

**FEATURE EXTRACTION MENGGUNAKAN LEXICON PADA
DATASET PENGENALAN EMOSI TEKS
BERBAHASA INDONESIA**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



diajukan oleh

APRILIA NURKASANA

18.11.1913

Kepada

PROGRAM SARJANA

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2021

**FEATURE EXTRACTION MENGGUNAKAN LEXICON PADA
DATASET PENGENALAN EMOSI TEKS
BERBAHASA INDONESIA**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



diajukan oleh

APRILIA NURKASANA

18.11.1913

Kepada

PROGRAM SARJANA

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

HALAMAN PENGESAHAN

SKRIPSI

**FEATURE EXTRACTION MENGGUNAKAN LEXICON PADA
DATASET PENGENALAN EMOSI TEKS
BERBAHASA INDONESIA**

yang disusun dan diajukan oleh

Aprilia Nurkasanah

18.11.1913

Telah dipertahankan di depan Dewan Penguji
pada tanggal 17 Juni 2022

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Arifiyanto Hadinegoro, S.Kom, MT.

NIK. 190302289

Arif Dwi Laksito, M.Kom.

NIK. 190302150

Mardhiya Hayaty, S.T., M.Kom.

NIK. 190302108

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 17 Juni 2022


DEKAN FAKULTAS ILMU KOMPUTER

Hanif Al Fatta, S.Kom., M.Kom.

NIK. 190302096

IDENTITAS JURNAL

Ultimatics : Jurnal Teknik Informatika
eISSN : 2581186X | pISSN : 20854552
Universitas Multimedia Nusantara





S3 Sinta Score		GARUDA Indexed by GARUDA	
2	H-Index	2	HS-Index
31	Citations	30	5 Year Citations





Nama Jurnal : Ultimatics: Jurnal Teknik Informatika
eISSN : 2581-186X
pISSN : 2085-4552
Publisher : Universitas Multimedia Nusantara
Akreditasi : <https://sinta.kemdikbud.go.id/journals/detail?id=4291> (Sinta 3)


LEMBAR REVIEW

Email Revisi:

[TI] Editor Decision, Revision Request: Feature Extraction using Lexicon on the Emotion Recognition Dataset of Indonesian Text  

Eksternal Kotak Masuk x

 **Rosa Reska Riskiana** <adminojs@umn.ac.id> Sel, 10 Mei 13.56   
kepada Mardhiya, saya ▾

 Inggris ▾ > Indonesia ▾ Terjemahkan pesan Nonaktifkan untuk: Inggris x

Dear Mardhiya Hayaty, Aprilia Nurkasanah,

After careful review, we have reached a decision regarding your submission to **Ultimatics : Jurnal Teknik Informatika**, titled "**Feature Extraction using Lexicon on the Emotion Recognition Dataset of Indonesian Text**".

Your article requires **REVISION** to be formally accepted and published in our Journal. You are asked to submit a revised full manuscript, according to the comments from reviewers. Our Editorial Team will check whether the revised version of the manuscript already addresses all points taken from the reviewers' comments.

Please upload the revised version on the Revision Section on your submission page within **2 weeks** using Word (.doc/.docx) format.

Editor Comments :

Thank you for your cooperation
--
Rosa Reska Riskiana
Telkom University
rosareskaa@telkomuniversity.ac.id
Editor, Ultimatics : Jurnal Teknik Informatika

Revisi Reviewer A:

Reviewer A:

Is the paper technically sound?

No

Does the paper contribute to the body of knowledge?

Neutral

Is the subject matter presented in comprehensive manner?

Neutral

Are the references provided applicable and sufficient?

Yes

Are there references that are not appropriate for the topic being discussed?

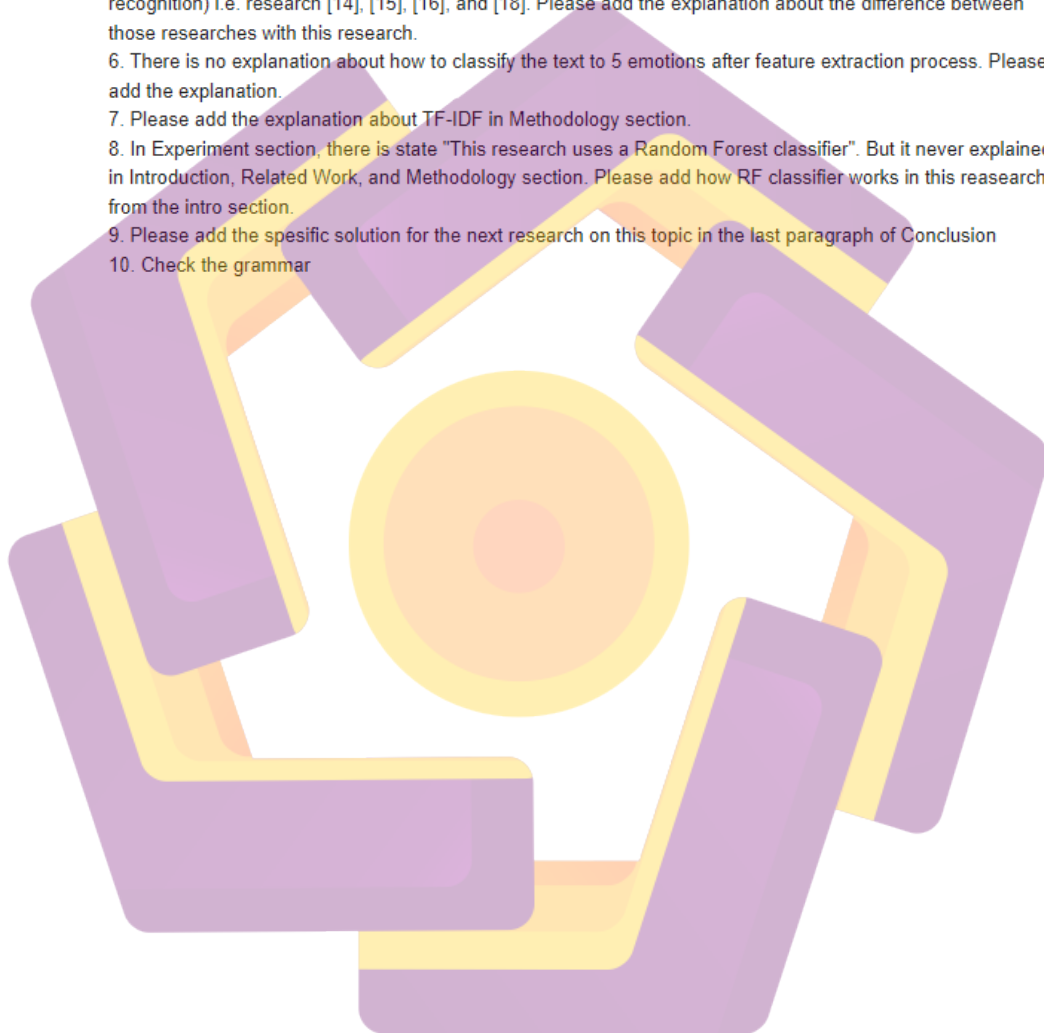
No

If yes, then please indicate which reference should be removed.

-

Comments to the Author :

1. There is Natural Language Processing word in the Abstract, but it did not explained in Introduction. Please add the explanation about NLP that is related to this paper.
2. The 1st paragraph in Introduction is not related to the 2nd paragraph. Please add the explanation so that communication (1st paragraph) and sentiment analysis (2nd paragraph) are related.
3. In section Related Work, please add the explanation about the difference between this reasearch and the previous researches.
4. The 2nd paragraph in Related Work has to be divided to two or three related paragraphs.
5. In the Related Work, there is some previous researches used the same method for the same topic (emotion recognition) i.e. research [14], [15], [16], and [18]. Please add the explanation about the difference between those researches with this research.
6. There is no explanation about how to classify the text to 5 emotions after feature extraction process. Please add the explanation.
7. Please add the explanation about TF-IDF in Methodology section.
8. In Experiment section, there is state "This research uses a Random Forest classifier". But it never explained in Introduction, Related Work, and Methodology section. Please add how RF classifier works in this reasearch from the intro section.
9. Please add the spesific solution for the next research on this topic in the last paragraph of Conclusion
10. Check the grammar



Revisi Reviewer B:

Reviewer B:

Is the paper technically sound?

Neutral

Does the paper contribute to the body of knowledge?

Yes

Is the subject matter presented in comprehensive manner?

Neutral

Are the references provided applicable and sufficient?

Should be improved

Are there references that are not appropriate for the topic being discussed?



No



If yes, then please indicate which reference should be removed.

Comments to the Author :

- Please add identity related to retrieval of tweets, including data taken on what date and trending topics at that time.
- Use the reference of the last 5 years.

LEMBAR PERSETUJUAN (LOA)

[TI] Editor Final Decision: Feature Extraction using Lexicon on the Emotion Recognition Dataset of Indonesian Text  


Eksternal  Kotak Masuk 




Rosa Reska Riskiana <adminojs@umn.ac.id>

Sel, 24 Mei 16.00



kepada Mardhiya, saya 



Inggris 



Indonesia 

[Terjemahkan pesan](#)

[Nonaktifkan untuk: Inggris](#) 

Dear Mardhiya Hayaty, Aprilia Nurkasanah,

We have reached a decision regarding your submission to *Ultimatics : Jurnal Teknik Informatika*, titled "Feature Extraction using Lexicon on the Emotion Recognition Dataset of Indonesian Text".

Our decision is to **ACCEPT** your article to be published in our Journal in the earliest edition. Congratulations on your article acceptance and publication in our Journal.

Thank you for your consideration in publishing your work in our Journals, we hope to hear from you and publishing your works again in the future.

Best regards,

--

Rosa Reska Riskiana

Telkom University

rosareskaa@telkomuniversity.ac.id

Editor, *Ultimatics : Jurnal Teknik Informatika*

This message is being delivered on behalf of the *Ultimatics: Jurnal Teknik Informatika*.

Please do not reply to this email.

Any inquiries related to our Journal can be addressed to fijurnal@umn.ac.id, ultimatics@umn.ac.id or directly to our editors.

BUKTI KARYA ILMIAH

Feature Extraction using Lexicon on the Emotion Recognition Dataset of Indonesian Text

Aprilia Nurkasanah¹, Mardhiya Hayati^{2*}

Faculty of Computer Science, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
¹aprilianurkasanah@students.amikom.ac.id, ²mardhiya_hayati@amikom.ac.id

*corresponding author

Accepted on mmmmm dd, yyyy
Approved on mmmmm dd, yyyy

Abstract— Text Mining is a part of Neural Language Processing (NLP), also known as text analytics. Text mining includes sentiment analysis and emotion analysis which are often used to analyse social media, news, or other media in written form. The emotional breakdown is a level of sentiment analysis that categorises text into negative, neutral, and positive sentiments. Emotion is organised into several classes. This study categorized emotion into anger, fear, happiness, love, and sadness. This study proposed feature extraction using Lexicon and TF-IDF on the emotion recognition dataset of Indonesian texts. InSet Lexicon Dictionary is used as the corpus in performing the feature extraction. Therefore, InSet Lexicon was chosen as the dictionary to perform feature extraction in this study. The results show that InSet Lexicon has poor performance in feature extraction by showing an accuracy of 30%, while TF-IDF is 62%.

Index Terms—Emotion Recognition Text, Lexicon, InSet Lexicon, Feature Extraction, Random Forest

I. INTRODUCTION

Communication between individuals is generally done verbally and non-verbally. Verbal communication begins and is obtained through words, sentences, or paragraphs. We often do not catch the emotion implied in the written words, sentences, and sections in verbal communication. As time goes by, the way of communication between individuals has changed. When technology is developing, the exchange of information flows spreads and develops rapidly. Social media is an oral communication tool that is often accessed and has become a trend. In 2021, Indonesia will have active social media users, as much as 61.8% of the existing population[1]. The number increased by 6.3% from the last year. Not only as a trend, but social media has also become a necessity. It can be seen from the statistical data of the social media users that it continuously grows every year. Various video, voice, and text data are added every minute. The flow of data information carried by social media is enormous. Thus, it is necessary to have an accurate processing model so that the information on social media can be processed correctly. Social media has become a part of life in modern society. Through social media, people often

express themselves regarding existing changes. Not only do they give objective opinions on an event, but social media users also often express their subjective views. Social media is considered a critical communication medium [2]. Not only opinions or criticisms about an event, but social media users also often criticize products, services, and even government decisions currently hotly discussed. Social media can be used as a medium for monitoring public opinion. This social media user shares their emotions in posts that many people from various platforms watch. Emotion influences humans to express hundreds of words. Therefore, it is considered a crucial thing in communication.

Natural Language Processing (NLP) in computational linguistics is a modern study of linguistics using computer science tools. NLP is also known as a subfield of Artificial Intelligence (AI). [3] NLP extract meaningful information from various texts, including text from social media. Social media has big data containing essential informations. These data give valuable information about a particular phenomenon, government policies, or reviews of using specific products. [4] NLP allows researchers to easily extract beneficial insights in textual datasets while avoiding burdensome computational work. These responses, comments, and reviews are then assessed based on sentiment class, commonly known as sentiment analysis.

Starting from the field of sentiment analysis, which analyzes opinions or messages into negative, positive, and neutral sentiment, emotion recognition is a sub-field of more detailed sentiment analysis that recognizes happy, sad, and angry emotions. Emotional interaction has been a part of a psychological phenomenon in daily life. It can be easily found in everyday interaction and encourages the existence of emotion recognition. Emotion recognition becomes the basis of successful human interaction, communication, and decision-making. The development of big data influences it to become a significant issue both academically and industrially. [5] Text data produced in emotional communication is often used to understand the human's emotional state. A person's emotional state is often associated with

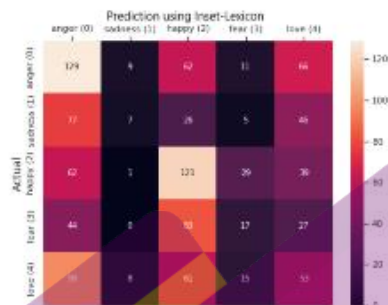


Fig. 7. Confusion Matrix of InSet Lexicon using Random Forest

The emotion class "Sadness" has a small proportion of false-positive values with moderate false negatives, which causes the "Sadness" class to have a small recall value even though it reaches a high precision value. Fig 8 shows the TF-IDF feature extraction confusion matrix in the Random Forest classifier.

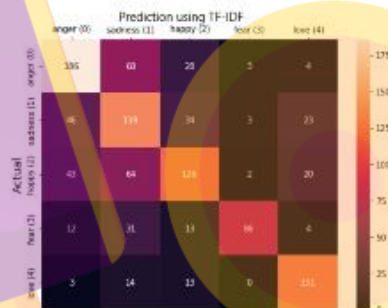


Fig. 8. Confusion Matrix of TF-IDF using Random Forest

Feature Extraction with TF-IDF provides better accuracy results on InSet Lexicon. Each emotion class can be predicted with a high F1-Score. The same with the Random Forest classification in InSet Lexicon feature extraction, the emotion class "Sadness", with a 49% score, has the lowest F1-Score value among other emotion classes.

V. CONCLUSION

This study examines the Lexicon based feature extraction on the emotion recognition dataset of Indonesian texts in the form of a corpus or dictionary called InSet Lexicon. The results in this study show an accuracy of 30%. The accuracy of this result is higher than the previous study that had the same dataset and classification. The difference in accuracy is influenced by the pre-processing stages carried out

in both studies. However, it has less accuracy than feature extraction using TF-IDF which has an accuracy of 62% on the Random Forest classifier. In this study, each emotion class (Anger, Sadness, Happy, Fear, and Love) can be detected with the used classifier. The low accuracy value in the InSet lexicon is caused by the polarity value that is not influenced by the emotion labels in the dataset. Lexicon feature extraction using the Lexicon InSet, which usually provides high accuracy when used to analyze Indonesian sentiment, produces low accuracy in the emotion recognition dataset of Indonesian texts.

We will examine and improve the current results for further research to achieve better performance. The researcher also suggests future studies to investigate the effect of class imbalance on the dataset for each text weighting scheme. The imbalance plays a significant role in creating a presupposition toward selecting the majority class in the emotion recognition dataset of Indonesian texts.

REFERENCES

- [1] S. Kemp, "DIGITAL 2021: THE LATEST INSIGHTS INTO THE 'STATE OF DIGITAL,'" *We Are Social*, 2021.
- [2] K. M. Carley, M. Malik, M. Kowalchuck, J. Pfeffer, and P. Landwehr, "Twitter Usage in Indonesia," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.2720332.
- [3] Y. Kang, Z. Cai, C. W. Tan, Q. Huang, and H. Liu, "Natural language processing (NLP) in management research: A literature review," *J. Manag. Anal.*, vol. 7, no. 2, pp. 139–172, 2020, doi: 10.1080/23270012.2020.1756939.
- [4] K. Li, C. Zhou, X. (Robert) Luo, J. Benitez, and Q. Liao, "Impact of information timeliness and richness on public engagement on social media during COVID-19 pandemic: An empirical investigation based on NLP and machine learning," *Decis. Support Syst.*, no. June 2021, p. 113752, 2022, doi: 10.1016/j.dss.2022.113752.
- [5] J. Deng and F. Ren, "A Survey of Textual Emotion Recognition and Its Challenges," *IEEE Trans. Affect. Comput.*, vol. 2045, no. c, pp. 1–20, 2021, doi: 10.1109/TAFFC.2021.3053275.
- [6] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Eng. Reports*, vol. 2, no. 7, pp. 1–24, 2020, doi: 10.1002/eng2.12189.
- [7] T. M. Li, H. C. Chao, and J. Zhang, "Emotion classification based on brain wave: a survey," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, 2019, doi: 10.1186/s13673-019-0201-x.
- [8] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems (2020)* 62, 2020.
- [9] Mulyono, A. S. Winarsih, and C. Supriyanto, "Evaluation of classification methods for Indonesian text emotion detection," *Proc. - 2016 Int. Semin. Appl. Technol. Inf. Commun. ISEMAN TIC 2016*, pp. 130–133, 2016, doi: 10.1109/ISEMAN TIC.2016.7873824.
- [10] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, vol. 36, no. 1, pp. 20–38, 2019, doi: 10.1016/j.ijresmar.2018.09.009.
- [11] R. C. W. Fera Fanesya Indriati, "Deteksi Emosi pada Twitter Menggunakan Metode Naive Bayes dan Kombinasi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 7, p. 3, 2019.
- [12] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon Based Sentiment Analysis in Indonesia Languages: A Systematic Literature Review," *RSF Conf. Ser. Eng. Technol.*, vol. 1, no. 1, pp. 363–367, 2021, doi: 10.31098/cset.v1i1.397.

- [13] J. Bata, "# AkuGalau: Korpus Bahasa Indonesia untuk Deteksi Emosi dari Teks," *J. Elektro*, vol. 12, no. 2, pp. 103–110, 2019.
- [14] M. A. Azim and M. H. Bhuiyan, "Text to emotion extraction using supervised machine learning techniques," *Telkomnika (Telecommunication Comput. Electron. Control)*, vol. 16, no. 3, pp. 1394–1401, 2018, doi: 10.12928/TELKOMNIKA.v16i3.8387.
- [15] D. William and D. Suhartono, "Text-based Depression Detection on Social Media Posts: A Systematic Literature Review," *Procedia Comput. Sci.*, vol. 179, no. 2019, pp. 582–589, 2021, doi: 10.1016/j.procs.2021.01.043.
- [16] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion Classification on Indonesian Twitter Dataset," *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, no. November, pp. 90–95, 2019, doi: 10.1109/IALP.2018.8629262.
- [17] P. Ekman, "An Argument for Basic Emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.
- [18] J. E. The, A. F. Wicaksono, and M. Adriani, "A two-stage emotion detection on Indonesian tweets," *JC1CSIS 2015 - 2015 Int. Conf. Adv. Comput. Sci. Inf. Syst. Proc.*, pp. 143–146, 2016, doi: 10.1109/ICACSIS.2015.7415174.
- [19] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *Eurasip J. Wirel. Commun. Netw.*, vol. 2017, no. 1, pp. 1–12, 2017, doi: 10.1186/s13638-017-0993-1.
- [20] V. Sundaram, S. Ahmed, S. A. Muqadeer, and R. Ravinder Reddy, "Emotion Analysis in Text using TF-IDF," *Proc. Conf. 2021 11th Int. Conf. Cloud Comput. Data Sci. Eng.*, pp. 292–297, 2021, doi: 10.1109/Confluence51648.2021.9377159.
- [21] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF feature weighting method and its analysis using unstructured dataset," *CEUR Workshop Proc.*, vol. 2870, pp. 98–107, 2021.
- [22] S. S. and P. K. V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, vol. 6, no. 4, pp. 300–305, 2020, doi: 10.1016/j.icte.2020.04.003.
- [23] N. Kholodna, V. Vysotska, and S. Albou, "A machine learning model for automatic emotion detection from speech," *CEUR Workshop Proc.*, vol. 2917, pp. 699–713, 2021.
- [24] A. Onan, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," *Comput. Appl. Eng. Educ.*, vol. 29, no. 3, pp. 572–589, 2021, doi: 10.1002/cae.22253.
- [25] J. L. Speiser, M. E. Miller, J. Toozre, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019, doi: 10.1016/j.eswa.2019.05.028.
- [26] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, pp. 1–15, 2019, doi: 10.1002/widm.1301.
- [27] H. Luzern, *Cyber Security Cyber Security*, no. March, 2017, doi: 10.1007/978-981-16-9229-1.
- [28] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern Recognit. Lett.*, vol. 93, pp. 133–142, 2017, doi: 10.1016/j.patrec.2016.12.009.
- [29] J. Akram and A. Tahir, "Lexicon and Heuristics Based Approach for Identification of Emotion in Text," in *2018 International Conference on Frontiers of Information Technology (FIT)*, Dec. 2018, pp. 293–297, doi: 10.1109/FIT.2018.00058.
- [30] D. D. Palmer, "Text pre-processing," *Handb. Nat. Lang. Process. Second Ed.*, pp. 9–30, 2010, doi: 10.4018/978-1-5225-4990-1.ch006.
- [31] Y. Nooryuda Prasetya and D. Winarsa, "Penerapan Lexicon Based Untuk Analisis Sentimen Pada Twitter Terhadap Isu Covid-19," *J. Fasilkom*, vol. 11, no. 2, pp. 97–103, 2021.
- [32] V. V. Nhlabano and P. E. N. Lutu, "Impact of Text Pre-Processing on the Performance of Sentiment Analysis Models for Social Media Data," *2018 Int. Conf. Adv. Big Data. Comput. Data Commun. Syst. ICABCD 2018*, pp. 1–6, 2018, doi: 10.1109/ICABCD.2018.8465135.
- [33] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, W. Alnumay, and A. P. Smith, "A lexicon-based approach to detecting suicide-related messages on Twitter," *Biomed. Signal Process. Control*, vol. 65, no. November 2020, 2021, doi: 10.1016/j.bspc.2020.102355.
- [34] M. Taboada, J. Brooke, M. Tofloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis DRAFT DRAFT!," *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, 2011.
- [35] F. Calefato, F. Lanubile, and N. Novielli, "EmoText: A toolkit for emotion recognition from text," *2017 7th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos. ACIIW 2017*, vol. 2018-Janua, no. July, pp. 79–80, 2018, doi: 10.1109/ACIIW.2017.8272591.
- [36] S. Liu and I. Lee, "Extracting features with medical sentiment lexicon and position encoding for drug reviews," *Health Inf. Syst. Syst.*, vol. 7, no. 1, 2019, doi: 10.1007/s13755-019-0072-6.
- [37] A. Bandhakavi, N. Wiratunga, P. Deepak, and S. Massie, "Generating a word-emotion lexicon from #emotional tweets," *Proc. 3rd Jt. Conf. Lex. Comput. Semant. *SEM 2014*, pp. 12–21, 2014, doi: 10.3115/v1/s14-1002.
- [38] A. Gholipour Shahraki and O. R. Zaiane, "Lexical and Learning-based Emotion Mining from Text," *Univ. Alberta, Canada*, pp. 24–55, 2017.
- [39] K. S. Nugroho, "Detecting Emotion in Indonesian Tweets: A Term-Weighting Scheme Study," vol. 8, no. 1, pp. 61–70, 2022, [Online]. Available: <https://www.e-journal.unair.ac.id/ISEBI/article/view/33017>
- [40] F. Koto and G. Y. Rahmanningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in *Proceedings of the 2017 International Conference on Asian Language Processing. IALP 2017*, 2018, vol. 2018-Janua, no. December, pp. 391–394, doi: 10.1109/IALP.2017.8300625.
- [41] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *CEUR Workshop Proceedings*, 2011, vol. 718, no. March, pp. 93–98.
- [42] R. N. Waykole and A. D. Thakare, "International Journal of Advance Engineering and Research A REVIEW OF FEATURE EXTRACTION METHODS FOR TEXT," pp. 351–354, 2018.
- [43] Q. Rajput, S. Haider, and S. Ghani, "Lexicon-Based Sentiment Analysis of Teachers' Evaluation," *Appl. Comput. Intell. Soft Comput.*, vol. 2016, pp. 1–12, 2016, doi: 10.1155/2016/2385429.
- [44] F. A. Bachtiar, W. Paulina, and A. N. Rusydi, "Text Mining for Aspect Based Sentiment Analysis on Customer Review: a Case Study in the Hotel Industry," *5th Int. Work. Innov. Inf. Commun. Sci. Technol.*, no. March, 2020.