International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Ina-SASet: Dataset for Developing Indonesian Sentiment Lexicon for Extracting Consumer Preference Based on Fine Grained Sentiment Analysis Technique

Bagus Setya Rintyarna [1*], Wiwik Suharso[2], Abadi Sanosra[3], Ika Safitri Windiarti[4]

Submitted: 17/05/2024 **Revised**: 28/06/2024 **Accepted**: 07/07/2024

Abstract*:* We collect and build a database called Ina-SASet, i.e.: dataset for developing Indonesian sentiment Lexicon for extracting consumer Preference based on fine grained sentiment analysis technique. The dataset was gathered from Twitter from January 1st 2022 to December 31st 2022, since Twitter is a popular microblogging platform for obtaining huge repository. We used Twitter API by employing several related keywords to crawl the dataset from Twitter. As many as 121.327 was successfully grabbed from Twitter platform. We then text pre-processed the dataset by 1) splitting the dataset into sentence, 2) applying tokenizing to the split sentences, 3) removing stop word by using online stop word collection, 3) stemming the token by matching them with Kamus Besar Bahasa Indonesia (KBBI) online repository. The stemmed token was then saved into MySQL database. The creation of an Indonesian sentiment lexicon resource pertaining to consumer preference extraction would be aided by this endeavour.

Keywords: *Sentiment analysis; Lexicon; Consumer preference, Twitter*.

1. Introduction

We can reveal one of the most important economic properties of consumer when choosing a bundle of goods based on their perception experiencing such product [1]. Common parameter that is used to identify consumer perception toward a product is called Consumer preference. Modelling consumer preference enable us to depict consumer acceptance to a certain commercial product [2]. Consumer preference represents subjective individual taste of consumer to various kind of product as measured by utility of product [3]. For the potential consumer, modelling consumer preference would help them finding the goods for their need [4] by lining their purchase decision more suitably based on a multiple attribute choice not only considering price as recommended by the consumer who has previously experienced such certain product [5]. Likewise, defining consumer preference is also beneficial for the companies [6].

Consumer preference is a prominent component for both product design and quality improvement since it represents purchase decision of consumer [7] value system based on multi attribute value theory i.e.: a multi criteria decision making problem. Main technique to be employed in such discrete choice modeling involves pencil survey by interviewing respondents to respond distinct description of products along with their level of attribute. It accounts the construct that any product can be represented in terms of its feature or characteristics that has any different level could be taken. However, discrete choice technique [8] based pencil survey is considered to be time consuming, human intensive and costly.

In that regard, this study attempts to model consumer preference automatically by using dataset from online platform using fine grained lexicon-based sentiment analysis technique. To assist the technique, we develop dataset to build an Indonesian sentiment lexicon platform for extracting consumer preference so called Ina-SA. For developing Ina-SA, we collect and build a database called Ina-SASet, i.e.: dataset for Developing Indonesian Sentiment Lexicon for Extracting Consumer Preference based on fine grained sentiment analysis [9].

In order to crawl the Twitter dataset, we employed the Twitter API and a number of relevant keywords. Using the Twitter platform, up to 121.327 were successfully taken. The dataset was then saved in a MySQL database following the use of a series of text preprocessing steps [10]. The following will be disclosed: 1) description of the data; 2) experimental design, materials and methodology; 3) conclusion; and 4) limitation.

[1]*Department of Electrical Engineering, Universitas Muhammadiyah Jember, Jember Regency, Indonesia.*
bagus.setya@unmuhjember.ac.id
[2]*Department of Informatics, Universitas Muhammadiyah Jember, Jember Regency, Indonesia*
[3]*Department of Management, Universitas Muhammadiyah Jember, Jember Regency, Indonesia*
[4]*Faculty of Business, Management and Information Technology, Universiti Muhammadiyah Malaysia, Perlis, Malaysia*

Specification Table

Subject	Computer Science
Specific Subject Area	Data set for developing Indonesian sentiment lexicon for extracting consumer preference based on fine grained sentiment analysis technique.
Type of the data	MySQL Database
How the data were acquired	Data was crawled from Twitter from January of 2022 to December of 2022 using Twitter Application Programming Interface (API) based on geolocation using several keywords i.e.: bisnis, consumer, consumer satisfaction, dagang, ekonomi, jual, kepuasan, kepuasan pelanggan, keuangan, layanan, manajemen, melayani, memuaskan, pangan, pasar, pelanggan, pemasaran, penjualan, puas, service, umkm.
Data format	Raw, Pre-Processed
Description of data collection	Data is firstly grabbed from Twitter using Twitter API by employing several keywords i.e.: bisnis, consumer, consumer satisfaction, dagang, ekonomi, jual, kepuasan, kepuasan pelanggan, keuangan, layanan, manajemen, melayani, memuaskan, pangan, pasar, pelanggan, pemasaran, penjualan, puas, service, umkm. Data was then pre-processed involving tokenizing, stop word removal and stemming. In compliance with Twitter Developer Policy, we only prensent Tweets (full text) and exclude user id, tweet link, and location [11] in the dataset.
Data source location	Universitas Muhammadiyah Jember
Data accessibility	The dataset is available at Mendeley: Setya Rintyarna, Bagus; Suharso, Wiwik; Sanosra, Abadi (2023), "Ina-SASet: dataset for developing Indonesian sentiment lexicon for extracting consumer reference based on fine grained sentiment analysis technique", Mendeley Data, V1, doi: 10.17632/52h3d44n85.1. Direct URL to data: https://data.mendeley.com/datasets/52h3d44n85/1

Value of the Data

- Consumer is the principal unit of economic. They define the suitable product to be obtained to satisfy their necessity [12]. Trough parameter so called consumer preference, we can reveal one of the most important economic properties of consumer when choosing a bundle of goods [13] based on their perception experiencing such product. By extracting consumer preference, we can portray the acceptance of consumer [14] toward a specific commercial product [15].
- Consumer preference, thus, describe subjective individual taste of consumer to various kind of product as measured by utility of product. In term of the potential consumer, modelling consumer preference would help them to decide the suitable product for their need as recommended by the previous consumer. Not only for consumer, defining consumer preference is also beneficial for the companies [16].
- Consumer preference is a prominent component for both product design and quality improvement [17] since it is represent purchase decision of consumer value system based on multi attribute value theory [18] i.e.: a multi criteria decision making problem. Since price is not the only determinant for purchasing the product, potential buyer likely savor consumer preference to line their purchase decision more suitably [15] by considering a multiple attribute choice.
- Common technique to calculate consumer preference employs technique called Discrete Choice based on a pencil survey [19]. The collection of the primary data in Discrete Choice based pencil survey involves survey of stated preference data by asking respondents to respond distinct description of products along with their level of attribute. Discrete choice [20] accounts the construct that any product can be represented in terms of its feature or characteristics that has any different level could be taken.
- Discrete choice by pencil survey [19], however, is considered to be time consuming, human intensive and costly [21]. In that regard, this study attempts to model consumer preference automatically by using dataset from online platform using fine grained lexicon-based sentiment analysis technique [7].
- To assist the technique, we develop dataset to build an Indonesian sentiment lexicon platform for extracting consumer preference so called Ina-SA. For developing Ina-SA, we collect and build a database called Ina-SASet, i.e.: Data set for Developing Indonesian Sentiment Lexicon (Ina-SA Ver 1.0) for Extracting Consumer Preference based on fine grained sentiment analysis technique [22].

2. Data Description

We have succesfully grabbed 121.327 tweets using Twitter API [23] by employing 20 keywords i.e.: bisnis, consumer, consumer satisfaction, dagang, ekonomi, jual, kepuasan, kepuasan pelanggan, keuangan, layanan, manajemen, melayani, memuaskan, pangan, pasar, pelanggan, pemasaran, penjualan, puas, service, umkm. The number of tweets that was succesfully grabbed for every keyword is presented in Fig. 1. The raw data was then saved on the MySql table dk_tbl_tweet.

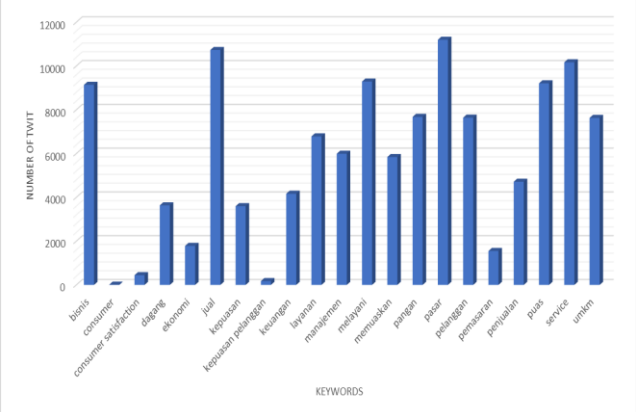

Fig 1. The number of Tweet

All of the dataset was saved in 5 MySql table i.e: 1) dk_tbl_tweet, 2) dk_tbl_katadasar, 3) dk_tbl_tokenizing, 4) dk_tbl_stopword and 5) dk_tbl_stemming as described in Table 1.

Table 1 My Sql table and the saved dataset

MySql Table	The saved dataset
dk_tbl_tweet	Raw data of the grabbed tweets
dk_tbl_katadasar	Collection from Kamus Besar Bahasa Indonesia (KBBI) grabbed from https://github.com/damzaky/kumpulan-kata-bahasa-indonesia-KBBI/blob/master/indonesian-words.txt
dk_tbl_tokenizing	The result of tokenizing
dk_tbl_stopword	The result of filtering using stop word removal
dk_tbl_stemming	The result of stemming

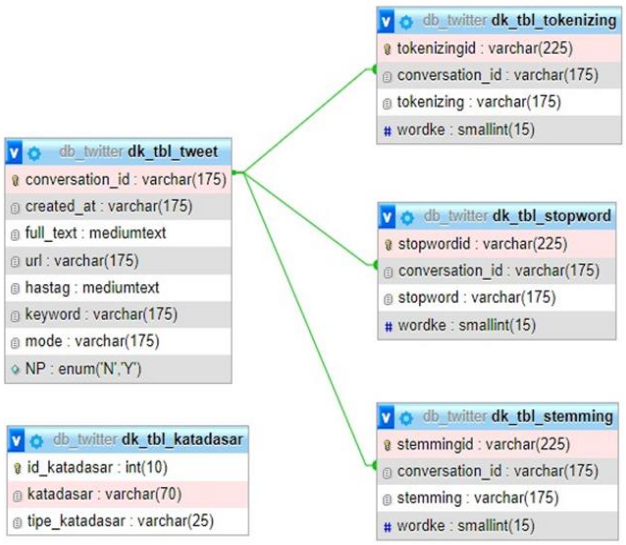

Fig. 2 MySQL table of the dataset

3. Experimental Design, Materials and Method

In developing Indonesian sentiment lexicon [24] for extracting consumer preference based on fine grained sentiment analysis (SA) technique [25], we need a collection of data from Twitter called Ina-SASet. Ina-SASet is the initial version of the lexicon for Ina-SA i.e.: platform for extracting consumer preference value based on SA technique [13]. The collecting process of Ina-SASet can be described in Fig 3. In the first step, data was grabbed from Twitter using Twitter API by employing several keywords i.e.: bisnis, consumer, consumer satisfaction, dagang, ekonomi, jual, kepuasan, kepuasan pelanggan, keuangan, layanan, manajemen, melayani, memuaskan, pangan, pasar, pelanggan, pemasaran, penjualan, puas, service, umkm. Example of the grabbed data is presented in Table 2. We also provide english translation of the tweet in Table 2.

After data was crawled from Twitter using Twitter API [26] and saved in dk_tbl_tweet, the data was then tokenized. The result of tokenizing was saved in dk_tbl_tokenizing. Before tokenizing, Tweets were split into sentence. To filter meaningless token, we filtered the sentence using stop word removal from https://github.com/masdevid/ID-Stopwords. The collection is then saved as MySQL database after stemmed using Nazief and Adriany stemming algorithm [10]. After tokenizing, data was stemmed. Beforehand, data was filtered using stop word removal. We use stop word data collection from https://github.com/masdevid/ID-Stopwords. The filtered data was saved on dk_tbl_stopword. After that, stemming was implemented to the data to get so called term (kata dasar) of every filtered token. Stemming can be done by using collection of Indonesian term as include in Kamus Besar Bahasa Indonesia (KBBI). We use the collection from https://github.com/damzaky/kumpulan-kata-bahasa-indonesia-KBBI/blob/master/indonesian-words.txt to apply stemming to the filtered token. We adopt algorithm from [27] to apply stemming. The term that was the result of stemming was saved on di_tbl_stemming. Stemming is of crucial important to deal with prefixes, suffixes, infixes, and confixes [28] of Indonesian language so the matching process of words [29] with KBBI collection could be carried out. The result of stemming is morphologinal root of the word [30]. Beforehand, the collection of KBBI was saved in dk_tbl_katadasar.

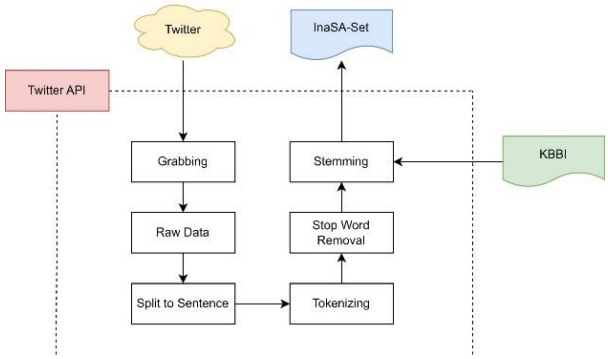

Fig 3. The collection process of Ina-SASet

Table 2 Example of the grabbed Tweet

Keywords	Tweets	English Translation of the Tweets
bisnis	Kelancaran dalam proses transaksi di bisnis retail dan F&B menjadi hal yang penting. Teknologi terdepat dari Epson untuk solusi point-of-sales (POS) akan membantu Anda meningkatkan efisiensi dan produktivitas yang akan berdampak pada kepuasan pelanggan. https://t.co/zKx2S4E8bL https://t.co/BOMqtJ61Xm	Smooth transaction processes in retail and F&B businesses are important. Epson's leading technology for point-of-sales (POS) solutions will help you increase efficiency and productivity which will impact customer satisfaction. https://t.co/zKx2S4E8bL https://t.co/BOMqtJ61Xm
dagang	Bisa bisanya dia makan yg lain sibuk dagang di live wkwkwk ???????????? Merangkap promosi semua bisnisnya ya mas.. Wajah jd cakep gini pake Elzagiarmi skincare ???????? https://t.co/3XOt93l4Jg	How could she eats while other people are busy trading on live ha ha ha ??????????? It's also a promotion for all your businesses, bro... Your face will look beautiful like this using Elzagiarmi skincare ??????? https://t.co/3XOt93l4Jg
kepuasan pelanggan	Kami menyadari bahwa kepuasan pelanggan dalam menikmati layanan internet dengan kualitas jaringan terbaik merupakan prioritas utama. Lewat senyuman tulus insan-insan terbaik kami berikan lewat pelayanan terbaik untuk pelanggan. Selamat Hari Konsumen Nasional https://t.co/xcgHFfwtvh	We realize that customer satisfaction in enjoying internet services with the best network quality is a top priority. Through the sincere smiles of the best people, we provide the best service to customers. Happy National Consumer Day https://t.co/xcgHFfwtvh

Keywords	Tweets	English Translation of the Tweets
keuangan	Di Wealth Wisdom 2022 #MindfullyRecover dari @PermataBank , kamu bakal dapetin banyak wawasan dari seminar dan workshopnya yang bakal membahas topik-topik seputar keuangan, ekonomi, bisnis, kesehatan, dan lifestyle #PermataBank	At Wealth Wisdom 2022 #MindfullyRecover from @PermataBank, you will get lots of insights from seminars and workshops which will discuss topics related to finance, economics, business, health and lifestyle #PermataBank
layanan	PLN juga semakin berorientasi terhadap kepuasan pelanggan dengan membuat proses pelanggan semakin terintegrasi. Salah satunya dengan terus mengembangkan aplikasi PLN Mobile untuk memudahkan pelanggan dalam menikmati layanan PLN. #TransformasiPLN #Power BeyondGenerations https://t.co/CCzhgxr3yV	PLN is also increasingly oriented towards customer satisfaction by making customer processes more integrated. One of them is by continuing to develop the PLN Mobile application to make it easier for customers to enjoy PLN services. #PLNTransformation #PowerBeyond Generations https://t.co/CCzhgxr3yV
melayani	Dibawah kepemimpinan Erick Thohir,BUMN benar2 hadir melayani dan membantu masyarakat.. Love you pak Erick.. God bless @erickthohir #BangkitBersamaET https://t.co/4UBbCd3jXB	Under the leadership of Erick Thohir, BUMN is truly here to serve and help the community.. Love you sir Erick.. God bless @erickthohir #BangkitBersamaET https://t.co/ 4UBbCd3jXB
pasar	Stray Kids memiliki 13 album di TOP400 (total penjualan bulanan tahun ini dari januari-oktober) artinya mereka telah mencatat 30.7% dari pangsa pasar (persentase dari total penjualan perusahaan). https://t.co/ED6O3NTdhp	Stray Kids has 13 albums in the TOP400 (total monthly sales of the year from January-October) meaning they have recorded 30.7% of the market share (percentage of the company's total sales). https://t.co/ED6O3NTdhp
pelanggan	Portofolio LaraDev kali ini berasal dari Siwi Pet Shop. Sistem Penjualan ini digunakan oleh Siwi Pet Shop untuk menangani penjualan kebutuhan hewan peliharaan secara online. Pelanggan dapat melihat dan melakukan pemesanan berbagai macam kebutuhan hewan peliharaan https://t.co/1DRdlW9K8T	LaraDev's portfolio this time comes from Siwi Pet Shop. This sales system is used by Siwi Pet Shop to handle online sales of pet needs. Customers can view and place orders for various pet needs https://t.co/1DRdlW9K8T
umkm	@GanjaranApp Semangat terus untuk UMKM lokal semoga kedepannya bisa makin sukses dan makin maju lagi, menurut saya ini salah satu cara untuk menarik pelanggan ya dengan inovasi yang baru kaya gini	@GanjaranApp Keep up the enthusiasm for local MSMEs, hopefully in the future they can be even more successful and more advanced, in my opinion this is one way to attract customers with new innovations like this
service	Promo Suzuki Ngegoal! All New Ertiga November Hujan Promo ✓ Bunga Ringan 0% ✓ DP Ringan 15% ✓ Smart Cash Bunga 0% [1 Tahun] ✓ Gratis Perawatan + Kaca Film Solargard Suzuki Pesta Akhir Tahun! ???? Info test drive, penjualan, service dan sparepart mobil Suzuki cek 082122214357 https://t.co/RGzGRMOgl8	Amazing Suzuki Promo! All New Ertiga November Rain Promo ✓ Low Interest 0% ✓ Low Down Payment 15% ✓ Smart Cash 0% Interest [1 Year] ✓ Free Maintenance + Solargard Suzuki Window Film End of Year Party! ???? Info on test drives, sales, service and spare parts for Suzuki cars, check 082122214357 https://t.co/RGzGRMO gl8

4. Conclusion

Consumer preference is a prevalent parameter that is employed to determine consumer perceptions of a product. It accounts for the concept that any product can be represented in terms of its features or characteristics, which can be interpreted at any level. The primary method used to extract customer preferences is called choice modelling, which involves conducting pencil surveys with interview subjects. Yet, discrete choice Technique-based pencil surveys are said to be expensive, time-consuming, and labour-intensive. In that sense, this study uses a fine-grained lexicon-based sentiment analysis technique to attempt automatically modelling consumer choice using a dataset from an online platform. This study is an initial attempt to create a platform for extracting customer preferences from sentiment lexicons in Indonesian called Ina-SA. A total of 121.327 were effectively taken from the Twitter network. The dataset was subsequently text pre-processed by 1) dividing it into sentences, 2) tokenising the split sentences, 3) eliminating stop words using an online stop word collection, and 4) stemming the tokens by comparing them to the Kamus Besar Bahasa Indonesia (KBBI) online repository. Next, the token with stems was stored in a MySQL database.

5. Limitation

Using a classification method will increase the efficacy of sentiment lexicon building. Using machine learning algorithms, such as deep learning, neural networks, or naïve bayes classifiers, is the most effective method. For the purpose of the training step, the machine learning algorithm's implementation will need that a portion of the dataset be pre-annotated by the annotator. Due to the sheer volume of Twits that were successfully collected throughout the crawling process, this dataset does not include any manually annotated data.

Ethics Statements:
This dataset complies with the Twitter developers' API terms of use and privacy policy [11].

Declaration of Competing Interest:

The authors declare that they have no competing interest that could influence the work reported in this paper.

Data Availability:

Direct URL to data:
https://data.mendeley.com/datasets/52h3d44n85/1

Credit Author Statement:

Bagus Setya Rintyarna: Conceptualization, Methodology, Software, Writing – original draft, Supervision, Funding acquisition; **Wiwik Suharso**: Conceptualiazation, Methodology; **Abadi Sanosra**: Conceptualization, Methodology.

Acknowledgement

We express our gratitude to the Indonesian Ministry of Education, Culture, Research, and Technology for financing and supporting this study as part of the Directorate General of Higher Education, Research, and Technology's ongoing research program.

References:

[1] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks," *J. Big Data*, vol. 6, no. 1, 2020, doi: 10.1186/s40537-019-0246-8.

[2] A. J. Najafabadi, A. Skryzhadlovska, and O. F. Valilai, "Agile Product Development by Prediction of Consumers' Behaviour; using Neurobehavioral and Social Media Sentiment Analysis Approaches," *Procedia Comput. Sci.*, vol. 232, no. 2023, pp. 1683–1693, 2024, doi: 10.1016/j.procs.2024.01.166.

[3] J. Chi, "Explaining US travel behavior with perceived threat of pandemic, consumer sentiment, and economic policy uncertainty," *Transp. Policy*, vol. 137, no. July 2022, pp. 90–99, 2023, doi: 10.1016/j.tranpol.2023.04.019.

[4] T. Anderson, S. Sarkar, and R. Kelley, "Analyzing public sentiment on sustainability : A comprehensive review and application of sentiment analysis techniques," *Nat. Lang. Process. J.*, vol. 8, no. August, p. 100097, 2024, doi: 10.1016/j.nlp.2024.100097.

[5] A. Daza, N. D. González Rueda, M. S. Aguilar Sánchez, W. F. Robles Espíritu, and M. E. Chauca Quiñones, "Sentiment Analysis on E-Commerce Product Reviews Using Machine Learning and Deep Learning Algorithms: A Bibliometric Analysisand Systematic Literature Review, Challenges and Future Works," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, 2024, doi: 10.1016/j.jjimei.2024.100267.

[6] H. Adiningtyas and A. S. Auliani, "Sentiment analysis for mobile banking service quality measurement," *Procedia Comput. Sci.*, vol. 234, pp. 40–50, 2024, doi: 10.1016/j.procs.2024.02.150.

[7] O. Alsemaree, A. S. Alam, S. S. Gill, and S. Uhlig, "Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions," *Heliyon*, vol. 10, no. 9, p. e27863, 2024, doi: 10.1016/j.heliyon.2024.e27863.

[8] N. Ortelli, M. de Lapparent, and M. Bierlaire, "Resampling estimation of discrete choice models," *J. Choice Model.*, vol. 50, no. January, p. 100467, 2024, doi: 10.1016/j.jocm.2023.100467.

[9] K. Lee, J. Kim, J. Kwon, and J. Yeo, "Maritime supply chain risk sentiment and the korea trade volume: A news big-data analysis perspective," *Asian J. Shipp. Logist.*, vol. 40, no. 1, pp. 42–51, 2024, doi: 10.1016/j.ajsl.2024.01.001.

[10] M. R. A. Rashid, K. F. Hasan, R. Hasan, A. Das, M. Sultana, and M. Hasan, "A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews," *Data Br.*, vol. 53, 2024, doi: 10.1016/j.dib.2024.110052.

[11] T. D. Platform, "Developer Agreement and Policy, (Accessed Aug 2023)".

[12] P. Krugman and R. Wells, "»Consumer Preferences and Consumer Choice," *Economics*, pp. 253–280, 2006, doi: 10.1007/978-1-349-91968-0_12.

[13] B. S. Rintyarna, H. Kuswanto, R. Sarno, and E. K. Rachmaningsih, "Modelling Service Quality of Internet Service Providers during COVID-19 : The Customer Perspective Based on Twitter Dataset," pp. 1–12, 2022.

[14] D. Bui, L. Dräger, B. Hayo, and G. Nghiem, "Macroeconomic expectations and consumer sentiment during the COVID-19 pandemic: The role of others' beliefs," *Eur. J. Polit. Econ.*, vol. 77, no. September 2022, p. 102295, 2023, doi: 10.1016/j.ejpoleco.2022.102295.

[15] P. Balart, "Semiorder preferences and price-oriented buyers in a hotelling model," *J. Econ. Behav. Organ.*, vol. 188, pp. 394–407, 2021, doi: 10.1016/j.jebo.2021.05.015.

[16] B. S. Rintyarna, R. Setyaningtyas, and I. S. Windiarti, "Assessment of Technology Acceptance Model of Water Quality Monitoring Technology Application on Rural Areas of Jember Regency," vol. 5, no. 2, pp. 1–7, 2023.

[17] J. Zhang, A. Zhang, D. Liu, and Y. Bian, "Customer preferences extraction for air purifiers based on fine-grained sentiment analysis of online reviews," *Knowledge-Based Syst.*, vol. 228, p. 107259, 2021, doi: 10.1016/j.knosys.2021.107259.

[18] Y. Chen, "ScienceDirect ScienceDirect support new energy automobile purchase decision Management Consumer preference disaggregation based on online reviews to Consumer preference disaggregation based online support new energy automobile purchase decision," *Procedia Comput. Sci.*, vol. 221, pp. 1485–1492, 2023, doi: 10.1016/j.procs.2023.08.013.

[19] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context," *Int. J. Inf. Decis. Sci.*, vol. 11, no. xxxx, 2018.

[20] M. Liao, "Identification of a rational inattention discrete choice model," *J. Econom.*, vol. 240, no. 1, p. 105670, 2024, doi: 10.1016/j.jeconom.2024.105670.

[21] B. S. Rintyarna, "Mapping acceptance of Indonesian organic food consumption under Covid-19 pandemic using Sentiment Analysis of Twitter dataset," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 5, pp. 1009–1019, 2021.

[22] W. Lin and L. C. Liao, "Lexicon-based prompt for financial dimensional sentiment analysis," *Expert Syst. Appl.*, vol. 244, no. September 2022, p. 122936, 2024, doi: 10.1016/j.eswa.2023.122936.

[23] S. Zervoudakis, E. Marakakis, H. Kondylakis, and S. Goumas, "OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data," *Mach. Learn. with Appl.*, vol. 3, no. September 2020, p. 100018, 2021, doi: 10.1016/j.mlwa.2020.100018.

[24] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Semantic Features for Optimizing Supervised Approach of Sentiment Analysis on Product Reviews," *MDPI Comput.*, vol. 8, no. 3, pp. 1–16, 2019.

[25] M. N. Hoque, U. Salma, M. J. Uddin, M. M. Ahamad, and S. Aktar, "Exploring transformer models in the sentiment analysis task for the under-resource Bengali language," *Nat. Lang. Process. J.*, vol. 8, no. June, p. 100091, 2024, doi: 10.1016/j.nlp.2024.100091.

[26] I. B. K. Manuaba, "A Sentiment Analysis Model for the COVID-19 Vaccine in Indonesia Using Twitter API v2, TextBlob, and Googletrans," *Procedia Comput. Sci.*, vol. 227, pp. 1101–1110, 2023, doi: 10.1016/j.procs.2023.10.621.

[27] J. Asian, B. Nazief, and H. Williams, "Stemming Indonesian : A confix-stripping approach . Stemming Indonesian : A Confi x-Stripping Approach," no. January, 2007, doi: 10.1145/1316457.1316459.

[28] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of

NLP-based sentiment analysis: A state-of-the-art review," *Nat. Lang. Process. J.*, vol. 6, no. February, p. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.

[29] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," *Nat. Lang. Process. J.*, vol. 7, no. October 2023, p. 100070, 2024, doi: 10.1016/j.nlp.2024.100070.

[30] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, no. March 2023, p. 102342, 2024, doi: 10.1016/j.is.2023.102342.