

Analysis Sentiment & Topic Modelling using Latent Semantic Indexing (A Case Study: Detik.com Application)

Glenzkovic Samuel Moniharapon^{1✉}, Dedi Iskandar Inan², Ratna Juita³, Victor Arie Lambadya Sirait⁴

^{1,2,3,4} Informatics Engineering, Engineering, Universitas Papua, Indonesia

Article Information

Article History

Received: September 13, 2024

Revised: September 22, 2024

Accepted: September 30, 2024

Keywords:

Application Quality, Latent Semantic Indexing, Sentiment Analysis, Topic Modelling, User Reviews

Kata Kunci:

Kualitas Aplikasi, Latent Semantic Indexing, Sentimen Analisis, Topik Modelling, Ulasan Pengguna

ABSTRACT

The proliferation of digital news platforms has transformed information consumption, making user experience (UX) a key success factor for Detik.com, a leading news application in Indonesia. This research aims to explore the relationship between UX and news through sentiment analysis and topic modelling. The methods include web scraping for data collection, descriptive analysis for sentiment, and Latent Semantic Indexing (LSI) for topic modelling. Out of 14,320 reviews analyzed, 45.6% were positive, focusing on the speed of updates and news quality. Conversely, 41.18% of reviews were negative, highlighting disruptive ads and login issues, while 13.76% were neutral. LSI identified eight main topics, including user experience and ad complaints. Visualizations through heatmaps, distance maps, and word clouds emphasized terms like "news" and "ads." This research recommends that Detik.com developers reduce ad disruptions, fix technical problems, and understand frequently discussed topics to enhance user satisfaction.

ABSTRAK

Proliferasi platform berita digital telah mengubah cara konsumsi informasi, menjadikan pengalaman pengguna (UX) sebagai faktor kunci kesuksesan, termasuk Detik.com, aplikasi berita terkemuka di Indonesia. Penelitian ini bertujuan mengeksplorasi hubungan antara UX dan berita melalui analisis sentimen dan pemodelan topik. Metode yang digunakan meliputi web scraping untuk pengumpulan data, analisis deskriptif untuk sentimen, dan Latent Semantic Indexing (LSI) untuk pemodelan topik. Dari 14.320 ulasan yang dianalisis, 45,6% bersifat positif, terutama terkait kecepatan pembaruan dan kualitas berita. Sebanyak 41,18% ulasan negatif menyoroti gangguan iklan dan masalah login, sementara 13,76% bersifat netral. LSI mengidentifikasi delapan topik utama, termasuk pengalaman pengguna dan keluhan iklan. Visualisasi melalui *heatmap*, *distance map*, dan *word cloud* menyoroti kata-kata seperti "berita" dan "iklan." Penelitian ini merekomendasikan agar pengembang Detik.com fokus mengurangi gangguan iklan dan memperbaiki masalah teknis, serta memahami topik-topik yang sering dibahas oleh pengguna untuk meningkatkan kepuasan pengguna.

Corresponding Author:

Glenzkovic Samuel Moniharapon

Informatics Engineering, Engineering, Universitas Papua, Indonesia

Gunung Salju, Amban, Manokwari, West Papua, Indonesia

E-mail: glenzkovic@gmail.com

INTRODUCTION

In this fast-paced digital era, mobile news applications like Detik.com play a crucial role in providing easily accessible information to the public (Denny Indah Suryawati et al., 2023). The Detik.com application is designed to meet the public's information needs by providing the latest news, informative articles, and various other features. Developed by Trans Digital Media, the application has a 4.3 rating on Google Play Store and was first released on September 6, 2010. Since its release, Detik.com has become one of the leading news sources in Indonesia (Alvendri et al., 2023). With the increasing use of smartphones, this application not only offers convenience in accessing news but also enables users to receive real-time updates. Through this application, the public can quickly obtain relevant and up-to-date information.

This study aims to understand user satisfaction levels by analyzing reviews of the Detik.com application obtained from the Google Play Store using sentiment analysis and topic modelling (Gifari et al., 2022). This allows us to comprehend users' attitudes, opinions, or emotions toward a particular topic. These factors can be either positive or negative and may originate from internal or external sources (Inan et al., 2022). Topic modelling is a natural language processing technique used to identify and group the main topics within large text documents (Naury et al., 2021). One of the methods in topic modelling, Latent Semantic Indexing (LSI), is employed to identify hidden relationships between words in a collection of text documents (Anggraini et al., 2021).

The study by (Alwasi'a, 2020) evaluated sentiment analysis of user reviews for the Detik.com application on the Google Play Store in 2019 using the Maximum Entropy method. The results demonstrated a high accuracy rate in sentiment classification, reaching 95.69%, with the system performing exceptionally well in classifying positive reviews at 97.45% and negative reviews at 86.17%. Furthermore, the study identified topics frequently mentioned by users, such as the words "good," "news," "okay," "update," and "excellent" for positive reviews, and "no," "news," "ads," "Detik," and "application" for negative reviews. These findings provide insights into the strengths and weaknesses of the Detik.com application based on user perceptions.

This study, on the other hand, focuses on the Detik.com news application, assessing application quality and user experience through descriptive statistical analysis of sentiment and employing Latent Semantic Indexing (LSI) for topic modelling. By applying this method to Detik.com application reviews on the Play Store, it is possible to identify dominant trends in user opinions. This can assist in guiding the development of the application and formulating more effective communication strategies for the Detik.com team to meet and exceed user expectations. User satisfaction is a key factor in the success of an application (Ericko Wicaksono et al., 2023). By understanding the dominant sentiments and topics in user reviews, the development team can more effectively direct application development and communication strategies. This study aims to apply topic modelling methods to extract valuable information from user opinion data provided for the Detik.com application on the Play Store. As news consumption in the digital information era has shifted from traditional platforms to online sources, Detik.com application users are also part of this trend. Therefore, understanding their perspectives is crucial in navigating strategic directions within the media industry (Steorts, 2023).

METHODOLOGY

This study in figure 1, begins with data on Detik.com application reviews, which were obtained through scraping from the Google Play Store. The results of the scraping consist of rating data, which is then used for Sentiment Analysis by evaluating the sentiment of the ratings received, categorized as positive, negative, or neutral. Subsequently, a research model design is prepared using Orange Data Mining, followed by preprocessing, which includes transformation, tokenization, filtering to enhance text quality and data cleansing. Afterward, topic modelling is performed using Latent Semantic Indexing (LSI) to identify semantic relationships between words by reducing data dimensions using Singular Value Decomposition (SVD).

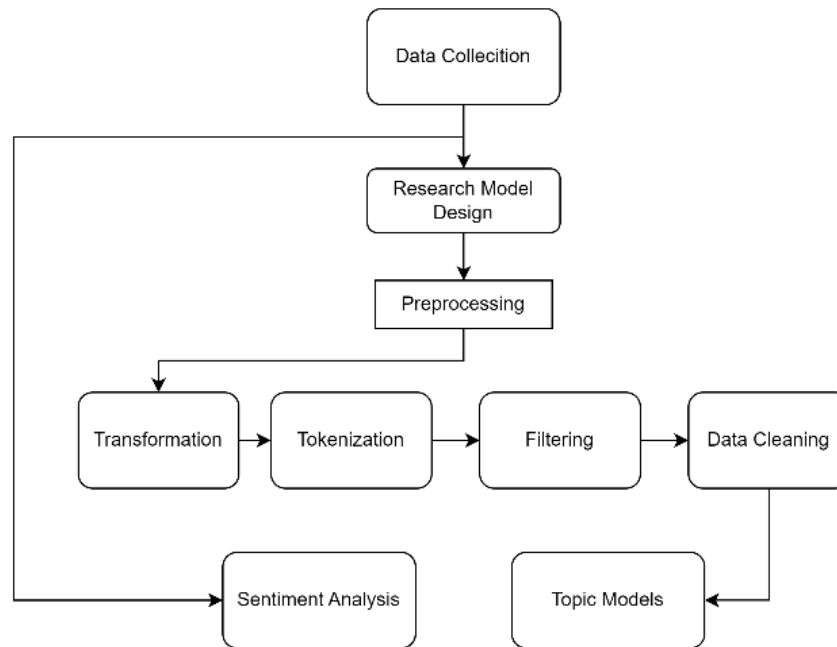


Figure 1. Research Workflow

Data Collection

Data is obtained from Detik.com application reviews on the Play Store in figure 2, through a process known as scraping. Scraping data refers to the automated extraction of information or data from a website (Flores et al., 2020). Scraping facilitates the collection of large datasets for trend analysis, user behavior, sentiment analysis, and market research (Matthew J. Denny & Arthur Spirling, 2018). This process is carried out using software programs that automatically navigate the Google Play Store web pages, extract Detik.com application review information, and store it. Data collection from the Google Play Store is performed using Google Collab with the Python programming language.

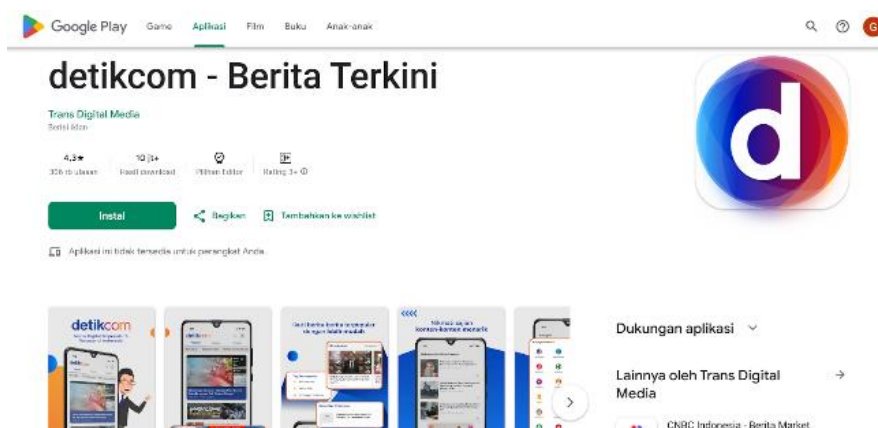


Figure 2. Detik.com Application on Google Play Store

Research Model Design

Orange Data Mining, commonly known as Orange, is an open-source component-based software for data mining, designed to manage data from cleaning through to knowledge extraction. This software is utilized for preprocessing and topic modelling. It offers numerous data cleaning features, such as handling missing values, data normalization, and data transformation. The design of the Orange widgets used in this study is illustrated in the figure 3.

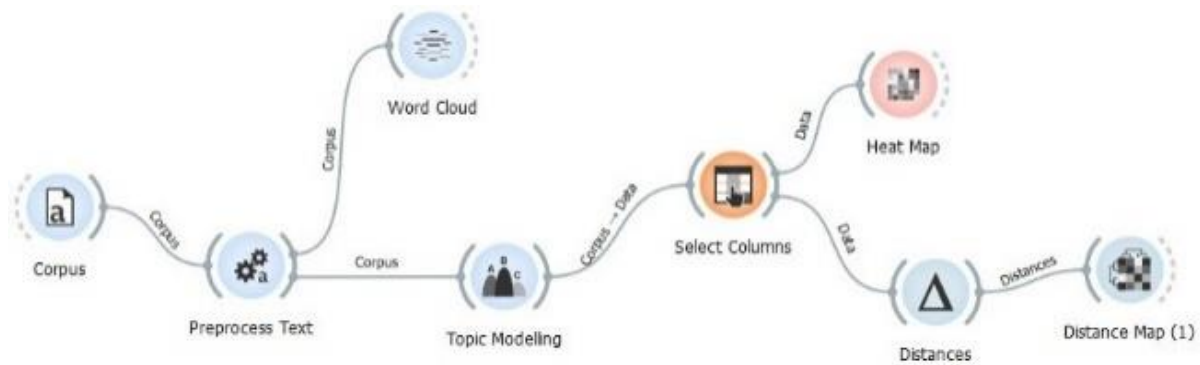


Figure 3. Research Model Design using Orange Data Mining

Preprocessing

In the context of text processing, preprocessing refers to a series of steps or techniques used to clean, transform, or adjust text data before further analysis. Preprocessing involves data cleaning and preparation processes aimed at enhancing the quality of text data before it enters the analysis phase (F. L. Hakim, 2023). Orange Data Mining software and Microsoft Excel offer a range of features that facilitate these preprocessing steps.

Transformation: The process of altering or modifying text in a specific manner. Transformation may involve changing formats, such as converting letters to lowercase or uppercase, removing accents from characters, removing punctuation, and so on. The primary goal of this transformation is to standardize the text to facilitate more efficient comparison and analysis, reduce ambiguity, and enhance the accuracy of the analytical tools or models employed.

Tokenization: The process of dividing text into smaller units known as tokens. Tokens can be words, phrases, or other meaningful segments of the text. Tokenization is typically carried out by separating the text into individual words, considering spaces and punctuation as delimiters, although it can be adjusted with specific rules depending on the requirements of the analysis.

Filtering: The process of removing or extracting specific parts of the text based on certain criteria. This may involve eliminating undesirable elements, such as URLs, punctuation, or specific words deemed irrelevant or disruptive to the analysis.

Data Cleaning: The results of preprocessing in Orange Data Mining are still suboptimal, necessitating manual cleaning using Microsoft Excel. This process involves removing words that lack meaning. It also includes replacing abbreviations with their full forms to make the text clearer and more comprehensible (Ismail et al., 2022).

Sentiment Analysis

Sentiment Analysis is the study of opinions, emotions, and attitudes towards entities such as individuals, events, and topics (Fernando & Toba, 2020). Sentiment analysis classifies the sentiment of text documents into three categories: positive, negative, and neutral. Its purpose is to evaluate human perceptions of these entities and enable businesses to make informed decisions. After data collection, the scraping results of the scores (ratings) are analyzed using descriptive statistics, examining scores on a scale of 1-5, where ratings of 1-2 are considered negative, 3 is considered neutral, and 4-5 are considered positive.

Topic Modelling

In this phase, the results of the sentiment analysis from the previous stage are further examined to determine the topics discussed by users of the application. This is represented by the number of samples used in this study. The goal is for the application's developers to understand the concerns of its users in relation to their level of satisfaction (user satisfaction) when using the application (user experience). In addition, evaluation is the process of determining the outcomes

of various efforts undertaken to support the achievement of specific objectives (Giyai et al., 2024). This feedback will be highly valuable for developers to improve the Detik.com application and subsequently enhance user satisfaction (user experience/UX) or even retain existing users. In this topic modelling phase, the algorithm employed will search for and group review results by identifying frequently co-occurring phrases. In this context, "topics" refer to recurring or prominent themes or concepts that reflect the main content of the review text.

In this study, the algorithm used for topic modelling is Latent Semantic Indexing (LSI). LSI is a method for topic modelling that enables the uncovering of semantic meaning or concepts contained within text using a mathematical technique known as Singular Value Decomposition (SVD) (Sallu Muharomah, 2023). The operation of LSI begins with creating a matrix where each row represents a document and each column represents a term (word). This matrix is then transformed using SVD to reduce data dimensions while preserving the important structural relationships between documents and terms. This process aids in identifying patterns or topics that characterize the dataset and helps group similar documents based on these topics. Additionally, LSI is effective for topic modelling.

RESULT AND DISCUSSION

Scraping

Table 1. Variabel from Scraping Result

No	Variable	Description
1	reviewId	Unique identifier for each review.
2	username	Username of the reviewer.
3	userImage	Image or avatar of the user.
4	Content	Text content of the review provided by the user.
5	Score	Score or rating given by the user to the application (scale 1-5)
6	thumbsUp-Count	Number of 'thumbs up' or 'likes' received by the review.
7	reviewCre-atedVersion	Version of the application reviewed at the time the review was given.
8	At	Timestamp or time when the review was created.
9	replayCon-tent	Response from the application owner or admin to the review.
10	repliedAt	Time when the application owner or admin responded to the review.
11	appVersion	Version of the application being reviewed

The data collection in this study involves secondary data obtained from Detik.com application reviews available on the Google Play Store. A total of 14,320 user reviews were collected out of 306,000 reviews, representing 4.64% of the total available data. The data was gathered using scraping techniques. The results of the scraping process yielded a set of variables related to Detik.com application reviews on the Play Store, which can be seen in Table1 and result in figure 4.

A	B	C	D	E	F	G	H	I	J	K	L	M
reviewId	username	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	appVersion	sortOrder	appId
32eb1599-8c56-4c3-5adyahani Elnr	https://play-ih.g	iklan	iklannya sangat mengganggu & menyebarkan. Aplikasi terbaru	1	12	06.05.11	07/03/2024 10.41			06.05.11	most_relevant	org.detikcom.rss
aad8f8ec-aac3-4ef-Rosandy Medi	https://play-ih.g	Parah, malizin, suka asal ketik, isi	berita suka salah, dan yg hebat'y lagi, .detikcom makin gila	1	1	06.01.13	06/07/2021 03.17			06.01.13	most_relevant	org.detikcom.rss
6a27297-9882-42a-Dermasido De	https://play-ih.g	Sekarang udah jadi aplikasi sampah, iklan	dimana mana	1	1	06.05.12	22/05/2024 21.04			06.05.12	most_relevant	org.detikcom.rss
889055c-6516-46fa-Aleg Aleq	https://play-ih.g	terlalu banyak iklan, sangat	mengganggu	1	1	06.05.12	11/05/2024 04.44			06.05.12	most_relevant	org.detikcom.rss
38c55c5-5d8b-4e5-Julia Thistyan	https://play-ih.g	bertar iklan, ganggu banget	iklan, tau2 langsung ke shopee / tokped	1	1	06.05.12	20/05/2024 09.07			06.05.12	most_relevant	org.detikcom.rss
3dbd0d9-70ff-4ad-Danang Perma	https://play-ih.g	Tolong untuk video otomatis nya	dinonaktifkan buang-buang kuota saat baca ada video yang b	1	1	06.05	20/07/2023 03.32			06.05	most_relevant	org.detikcom.rss
4511729v-3a5b-484-Pengguna Goo	https://play-ih.g	terlalu banyak iklan full screen yg	sangat mengganggu, setelah update aplikasinya jadi sering i	1	72	05.04.02	15/03/2019 02.11			05.04.02	most_relevant	org.detikcom.rss
40da06d-43ab-4ad-Bambang Utu	https://play-ih.g	Review/feedback button please?	Supaya kita semua menikmati berita yang berkualitas.	1	255	06.05.12	02/04/2024 00.08			06.05.12	most_relevant	org.detikcom.rss
4ceab894-85d7-41a-Rio Ardian	https://play-ih.g	uninstall aja, berita makin membosankan,	diulang-ulang terus, jebakan iklannya ngeselin	1	26	06.05.09	02/01/2024 03.02			06.05.09	most_relevant	org.detikcom.rss
45f0a4de-453b-453-Hadi Nur	https://play-ih.g	Untuk pojok pojok sepih sola	sekarang GA ada ksesmen dan hasil hasil lga kurang update,ko bukan	1	143	06.01.15	26/08/2021 07.45			06.01.15	most_relevant	org.detikcom.rss
46b4813b-322b-40b-Natan Simamo	https://play-ih.g	terlalu banyak iklan, jadi	malas baca beritanya	1	1	06.05.11	19/03/2024 16.27			06.05.11	most_relevant	org.detikcom.rss
4e9410b1-87a2-43b-Nirwana Chak	https://play-ih.g	Hmm sepertinya memang	detik sudah tidak bisa dijadikan media mainstream panutan lagi. Cu	1	23	04.03.00	06/06/2020 02.33			04.03.00	most_relevant	org.detikcom.rss
4e9522a-899a-424-SUTARMAN SU	https://play-ih.g	Saat akan membuka suatu	berita dihalaman kedua, bukannya lanjutkan, blank page dengan ada	1	25	06.00.05	12/07/2020 05.27			06.00.05	most_relevant	org.detikcom.rss
4520766f-1971-45b-Bagus D	https://play-ih.g	Setiap buka app, ikannya ga	ngutuin...full 1 layer tp dan kadang ga bisa dilewat, mesti rung	1	20	06.01.04	31/08/2020 14.23			06.01.04	most_relevant	org.detikcom.rss
8f7202b1-67b8-47e-Ahmad Sholah	https://play-ih.g	Sekarang terlalu banyak	iklan, kayak aplikasi game dan streaming aja ikalnya, terlalu ba	1	94	06.05	06/05/2023 11.37			06.05	most_relevant	org.detikcom.rss
33318b8-af4b-4bb-Muhammad Iq	https://play-ih.g	Sejak iklan membantir, aplikasi	sering error jaringan padahal berita muncul di notifi, dan aplikasi	1	557	06.01.13	29/05/2021 14.46			06.01.13	most_relevant	org.detikcom.rss
1b05a14b-4c79-437-DEANNOVA B	https://play-ih.g	Satu2nya berita yg ada	ikannya. ikannya bertapis lapis lagi. Iklan diawal dan iklan didalam, ser	1	13	01/08/2021	20.29			01/08/2021	most_relevant	org.detikcom.rss
4a827434-102b-48f-Hendi Hadury	https://play-ih.g	Dulu detik sangat bagus	setiap update berita pasti jadinya susah dengan is beritanya, sekari	1	3	05.07.04	02/06/2020 03.48			05.07.04	most_relevant	org.detikcom.rss
04d0eb7-3ab8-4e-Rudyoc	https://play-ih.g	Karna akhir akhir ini saya	jarang sekali baca detik, di tambah lagi berita begitu tentang deti, se	1	3	05.07.04	02/06/2020 03.48			05.07.04	most_relevant	org.detikcom.rss
47e7b9d6-b1a7-48b-Andrayah Ma	https://play-ih.g	Terlalu banyak iklan	sekarang dan video selalu putar otomatis, banyak salah dalam menyajik	1	9	06.01.04	22/08/2020 13.44			06.01.04	most_relevant	org.detikcom.rss
3b6c31f1-1a7b-407-Pengguna Goo	https://play-ih.g	main di update malah makin	lambat banget. Kurang kurangnya iklan di dalam berita nya atau.	1	3	03.04.00	09/02/2019 08.53			03.04.00	most_relevant	org.detikcom.rss
4f46897-70f7-4e3-Mly Yee	https://play-ih.g	iklan ini sg byk dan	mengganggu, susah ditutupi menurutku layer. Makin parah detik ke sekari.	1	4	03/05/2021	04.23			03/05/2021	most_relevant	org.detikcom.rss
42e2ef1d-e9fa-451-Hadi Marfuah	https://play-ih.g	Artikalnya kebanyakan	endorsement. Udah jauh dari asas jurnalistik. Beritanya ga berimbang	1	38	06.01.01	18/06/2020 17.62			06.01.01	most_relevant	org.detikcom.rss
16a2ac7-912b-451-kiky Kusuma	https://play-ih.g	detik sekarang gak	lengkap, detik yg dulu ada ksesmen liga dan ada kurs rupiah nya, sekarang i	1	23	06.01.05	09/10/2020 04.15			06.01.05	most_relevant	org.detikcom.rss
45a42c9-3a0b-411-Istari Istief	https://play-ih.g	Aplikasi yang jadi spam	iklan, cuma ngeliat keutangan bukan user experience dinonaktifkan	1	4	14/05/2020	03.18			14/05/2020	most_relevant	org.detikcom.rss
35a53ad-4a8d-4ad-Nelson Darwis	https://play-ih.g	Detik pemberitaannya	GAK LAGI Aktual, Objektif dan Independen... kecewa sekali Ketika Medi	1	3	07/06/2020	09.02			07/06/2020	most_relevant	org.detikcom.rss
27858ae-265b-4af-Karakito Nadek	https://play-ih.g	Komentar gak	langsung muncul, sering keluar sendiri... Harus sering login... Padahal yg di soro	1	4	06.01.13	11/06/2023 03.42			06.01.13	most_relevant	org.detikcom.rss
333210b-6971-40f-Marco Tampoe	https://play-ih.g	Berita dari atas	tempe bawah di scroll sama semua, cuma di bedain judul tapi isi sama aja, cum	1	4	07/06/2020	10.34			07/06/2020	most_relevant	org.detikcom.rss

Figure 4. Scraping Results

Preprocessing

Preprocessing involves several steps to prepare the data for analysis, including transformation, tokenization, and filtering, as shown in Figure 5 using features from Orange Data Mining. These processes help convert raw text into a structured format suitable for analysis by breaking it down into tokens and removing irrelevant data. Additionally, data cleansing was performed using Microsoft Excel to ensure accuracy and consistency by eliminating errors, duplicates, or incomplete entries. This combination of tools and techniques ensures that the dataset is clean, well-organized, and ready for further analysis.

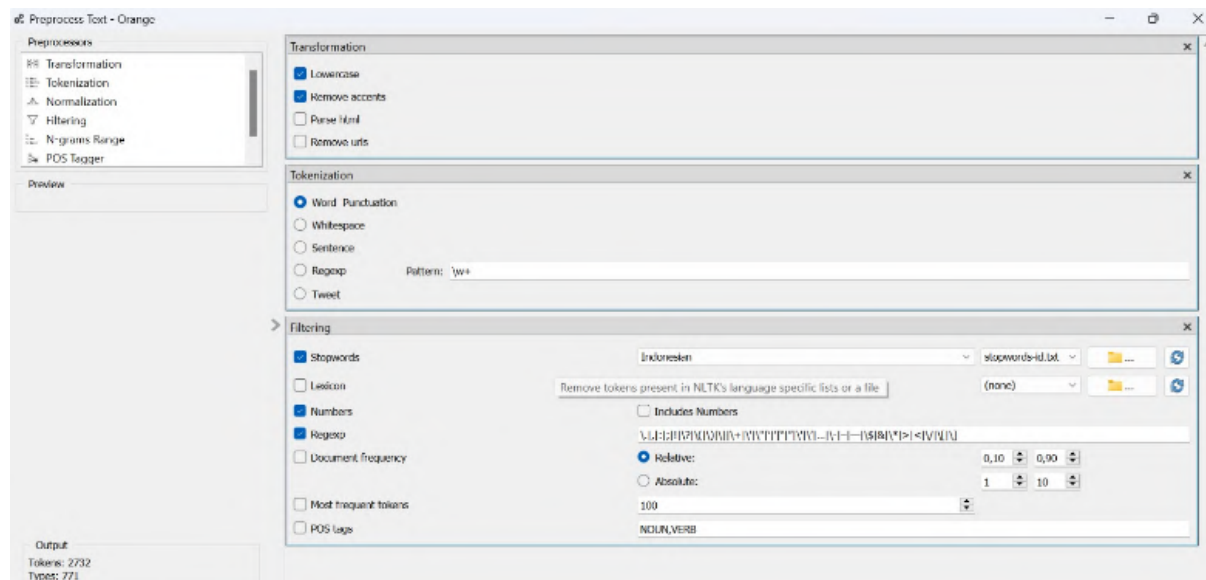


Figure 5. Data Preprocrssing in Orange Data Mining

Transformation involves converting all text to lowercase and removing accents or diacritics, such as changing "café" to "cafe." Tokenization divides the text into tokens based on spaces and punctuation, breaking down sentences like "Detik.com, berita-beritanya informatif." into individual components. Filtering includes removing stopwords that lack informational value, filtering out numbers, and using regular expressions to remove specific patterns such as URLs. Data cleaning eliminates inconsistencies and errors, resulting in more accurate and complete data, ultimately improving the quality and efficiency of data analysis.

Sentiment Analysis

Table 2. Rating result from Scraping

No	Rating	Quantity	Sentiment	Percentage (%)
1	★	3.226	Negative	22.53%
2	★ ★	2.671	Negative	18.65%
3	★ ★ ★	1.971	Neutral	13.76%
4	★ ★ ★ ★	1.990	Positive	13.90%
5	★ ★ ★ ★ ★	4.462	Positive	31.16%
Total		14.320		100%

Rating result from scraping in Table 2.

Overall, 45.06% of ratings have positive sentiment, while 41.18% are negative, and 13.76% are neutral. Topics related to positive sentiment include praise for the speed of news updates, the timeliness of information, and the quality of the news and information presented. Users also

appreciate the application's promptness in updating news, as well as the engaging quality of the headlines and content. Topics related to neutral sentiment include user experience and the quality provided by Detik.com.

Topics related to negative sentiment include the frequent appearance of advertisements, which not only appear once but repeatedly, becoming highly disruptive to readers. Advertisements often also appear in the form of auto-playing videos, causing users to feel disturbed and find it difficult to focus while reading the news. Additionally, there are inappropriate ads, such as those related to online gambling. Furthermore, the challenging login process, where users frequently get automatically logged out after exiting the application, forces them to log in repeatedly.

Topic Modelling

By identifying hidden patterns of relationships between words and documents, LSI enables the extraction of deeper semantic meaning, addressing issues of synonyms and polysemy in information retrieval, and improving accuracy in clustering documents based on more abstract conceptual similarities.

Topic	Topic keywords
1	berita, iklan, detik, aplikasi, update, klik, baca, detikcom, bagus, ga
2	iklan, berita, klik, update, mengganggu, kebanyakan, muncul, detik, detikcom, cepat
3	aplikasi, detik, berita, buka, bagus, klik, iklan, gak, video, hp
4	detik, aplikasi, bagus, detikcom, media, com, komentar, netral, berita, buka
5	bagus, update, aplikasi, detik, berita, buka, klik, informasi, bintang, gak
6	update, bagus, klik, detikcom, berita, informasi, cepat, judul, aplikasi, isi
7	klik, iklan, aplikasi, baca, mengganggu, update, muncul, kebanyakan, detikcom, detik
8	detikcom, klik, komentar, aplikasi, ya, berita, informasi, login, judi, media

Figure 6. Topic Model Results from Orange Data Mining

Based on the analysis of the 8 topics generated in table 3, it was found that 3 topics have positive reviews (Topics 5, 6, and 4), 2 topics have neutral reviews (Topics 1 and 3), and 3 topics have negative reviews (Topics 2, 7, and 8). Of the 8 topics, this distribution indicates that positive and negative reviews are relatively balanced, while a small proportion of reviews are neutral. However, the neutral reviews tend to lean towards a positive sentiment.

Table 3. Discussion of Topic Modelling Results

No	Topic	Description
1	News, Ads, Seconds, App, Up-dates, Click, Read, Detik.com, Good, No	User Experience: This topic covers various terms related to user interaction with the application, such as news, ads, and updates. It focuses on how users feel about and evaluate the application, including how well they can access news, how often they click, and their overall satisfaction with the Detik.com application.
2	Ads, News, Click, Updates, Disruptive, Excessive, Appears, Seconds, Detik.com, Fast	User Complaints: This topic addresses issues faced by users related to disruptive ads and news. Problems such as excessive ads or interruptions while reading news are primary concerns. It reflects user dissatisfaction with how the application displays ads and news.
3	App, Seconds, News, Open, Good, Click, Ads, No, Video, Phone	Application Quality: This topic evaluates various aspects of the Detik.com application, including app opening speed, news quality, and user experience with ads and videos. It reflects how well the application performs in providing content and features to its users.

No	Topic	Description
4	Seconds, App, Good, Detik.com, Media, Com, Comments, Neu-tral, News, Open	Information Assessment: The focus is on how users assess the quality of information provided by the application. Terms such as comments, neutral, and news indicate that this topic covers eval-uations of how well the application provides and presents news and information.
5	Good, Updates, App, Seconds, News, Open, Click, Infor-mation, Stars, No	User Reviews: This topic relates to how users review the applica-tion, including aspects such as updates, news quality, and infor-mation. Reviews often include assessments of how well the appli-cation manages content and features.
6	Updates, Good, Click, Detik.com, News, Information, Fast, Titles, App, Content	Speed & Information Update: This topic evaluates how quickly the application updates news and information and the quality of titles and content presented. It includes aspects of how efficiently the application provides the latest information to its users.
7	Click, Ads, App, Read, Disruptive, Updates, Appears, Excessive, Detik.com, Seconds	User Complaints: This topic covers issues related to disruptive ads and excessive updates, reflecting user dissatisfaction with disturbances caused by ads and the frequency of updates.
8	Detik.com, Click, Comments, App, Yes, News, Information, Login, Gambling, Media.	Content and Login Complaints: This topic includes issues related to the content presented and login problems that users may face. It also includes complaints related to gambling and media content, indicating concerns about content quality and accessibility as well as login features.

Figure 7, A heatmap is a visualization tool used to examine the relationships between variables or topics. The color on a heatmap indicates the strength or weakness of these relationships, and it can also identify patterns, trends, and outliers in the data. Brighter colors (yellow) or darker colors (blue) can indicate consistency or anomalies in the data, and it clusters data based on similarities or distances between data points. The heatmap is accompanied by a dendrogram that shows hierarchical clustering or grouping. The image above reveals 2 clusters:

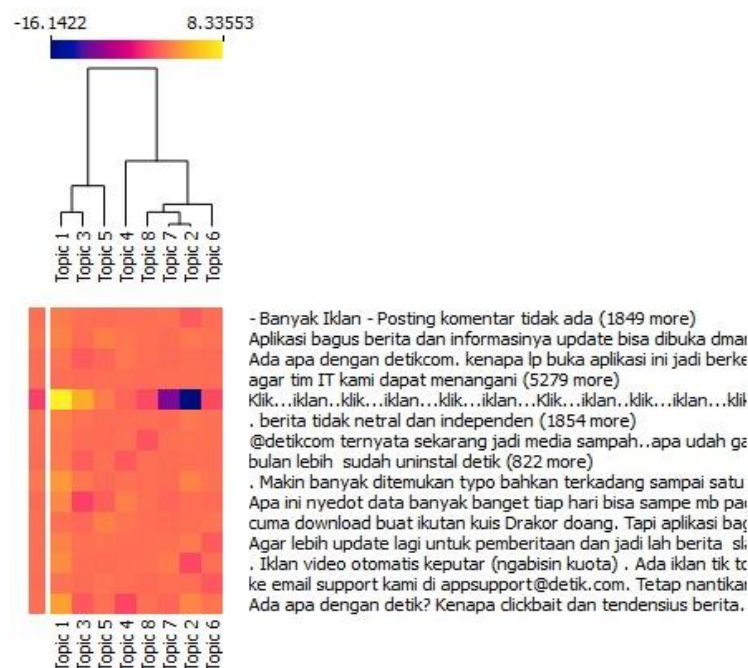


Figure 7. Topic Model Heat Map

The first cluster consists of Topic 1 (User Experience), Topic 3 (Application Quality), and Topic 5 (User Reviews), forming a single cluster. This cluster indicates a strong relationship between user experience, application quality, and the reviews provided. Users tend to leave reviews based on their experiences, which are influenced by the quality of the application.

The second cluster, in this thematic analysis, comprises several semantically related topics. Its internal sub-clusters are divided into two sections: Sub-cluster 2A includes Topic 2 (User Complaints), Topic 7 (User Complaints), and Topic 8 (Application Use & Complaints), while Sub-cluster 2B consists of Topic 6 (Speed & Information Updates) and Topic 4 (Information Assessment). The sub-cluster encompassing Topics 2, 7, and 8 merges with the sub-cluster including Topics 6 and 4, indicating that user complaints (Topics 2 and 7) and application usage (Topic 8) are often related to the speed of information updates (Topic 6) and the assessment of the information provided (Topic 4). This cluster encompasses various aspects of user complaints, frequently associated with application usage, the speed of information updates, and the evaluation of information provided by the application.

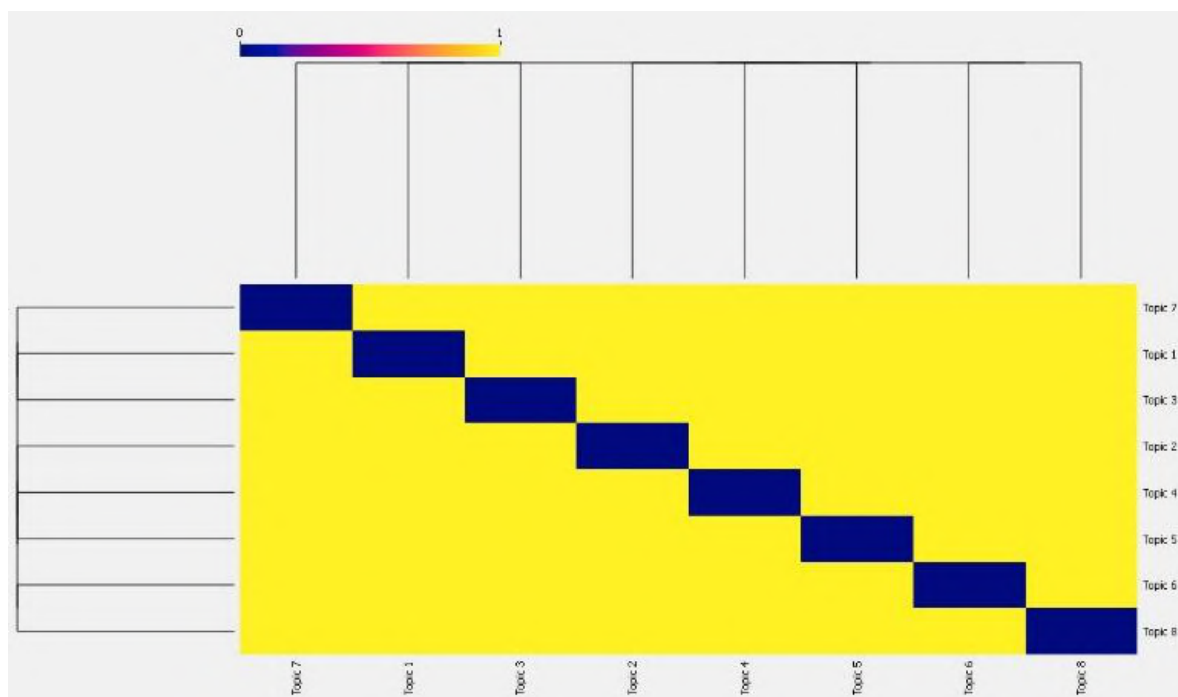


Figure 8. Heatmap of Topic Modelling Similarity

The distance map above illustrates the relationships between the identified topics. This map is used in cluster analysis to show how similar or different these topics are from one another. The X and Y axes represent the identified topics. In the distance map shown in Figure 8, there are eight topics labeled from "Topic 1" to "Topic 8."

In Figure 8, above, the matrix at the center of the graph represents the distance or similarity between the topics. Yellow indicates a greater distance (or less similarity), while blue indicates a shorter distance (or more similarity). The colors range from blue (nearest distance) to yellow (farthest distance), with the color scale above the graph showing the specific values of these distances. The blue diagonal from the top left to the bottom right indicates that each topic is identical to itself.

Topic 7 (User Complaints) is directly related to Topic 1 (User Experience) because complaints typically arise from negative experiences. Topic 1 (User Experience) is connected to Topic 3 (Application Quality) since the quality of the application affects user experience. Topic 3

(Application Quality) is related to Topic 2 (User Complaints) because a poor application often leads to complaints. Topic 2 (User Complaints) is also associated with Topic 4 (Information Assessment) as complaints frequently involve the evaluation of information.

Topic 4 (Information Assessment) is related to Topic 5 (User Reviews) because information assessment is a component of user reviews. Topic 5 (User Reviews) is connected to Topic 6 (Speed & Information Updates) as reviews often evaluate the speed and frequency of updates. Topic 6 (Speed & Information Updates) is associated with Topic 8 (Content and Login Complaints) because issues with speed and updates can lead to complaints about content and login. Topic 8 (Content and Login Complaints) is the final topic encompassing all the previous aspects, as these complaints can reflect various problems with the application.



Figure 9. Word Cloud of Frequently Mentioned Terms by Detik.com Users

Figure 9, word cloud is a visualization method used to display the most frequently occurring words in a text. In a word cloud, the more often a word appears in the text, the larger its size within the cloud. This provides a visual overview of the dominant words or concepts within a set of texts or specific documents. Below are the key terms that frequently appear: "news" (4928): This word appears most frequently, indicating that news content is a primary topic. "ads" (3259): The high frequency of this word suggests there are many advertisements or user complaints related to ads. "application" (1547): This indicates that the "Detik.com" application is a common topic of discussion. "second" (1409) and "Detik.com" (655): These terms indicate that the brand and platform are frequently mentioned. "good" (1346): This suggests that many users provide positive reviews.

CONCLUSION

This study reveals a strong correlation between sentiment analysis and topic modelling in user reviews of the Detik.com app. Out of the 14,206 reviews analyzed, 45.06% were positive, 41.18% negative, and 13.76% neutral. The applied topic modelling identified eight main topics, with three reflecting positive sentiment, three negative, and two neutral. This indicates that while positive reviews are slightly more prevalent, Latent Semantic Indexing (LSI) effectively captures a balance between positive and negative sentiments. The findings provide valuable insights into how users perceive various aspects of the app, such as the frequently praised speed of news updates and the major complaints about disruptive ads. This information is highly beneficial for Detik.com developers to enhance user experience by reducing ad interruptions and improving necessary features.

RECOMMENDATION

Based on the research that has been conducted, the following recommendations and suggestions are provided for further studies:

1. Increase Data Volume and Update Frequency: Expand the dataset to include a larger number of reviews and ensure that the data is recent to capture current trends and user sentiments more accurately.
2. Enhance Data Preprocessing: Improve data cleaning processes by incorporating techniques such as stemming and standardizing the Indonesian language dictionary to handle colloquial language more effectively.
3. Explore Alternative Topic Modelling Methods: Consider using different topic modelling methods such as Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), or Negative Matrix Factorization (NMF) for a more comprehensive analysis of topics.

REFERENCE

- Alvendri, D., Huda, Y., Darni, R., Negeri Padang, U., Hamka, J., Tawar Bar, A., Padang Utara, K., Padang, K., & Barat, S. (2023). Perancangan Media Pembelajaran Interaktif Konsep Dasar Seluler Menggunakan Aplikasi Unity Berbasis Android. *Journal on Education*, 05(04).
- Alwasi'a, A. (2020). Analisis Sentimen pada Review Aplikasi Berita Online Menggunakan Metode Maximum Entropy (Studi Kasus: Review Detik.com pada Google Play 2019). *Dspace.Aii.Ac.Id*.
- Anggraini, N., Surya Negara Harahap, E., Basuki Kurniawan, T., Bina Darma, U., Jenderal Yani No, J. A., & Sumatera Selatan, P. (2021). Text Mining-Text Analysis Related to COVID-19 Vaccination Issues. *Jurnal Ilmu Pengetahuan Dan Teknologi Komunikasi*, 23(2). <https://doi.org/10.33169/iptekkom.23.2.2021.141-153>
- Denny Indah Suryawati, Udi Rusadi, & Umaimah Wahid. (2023). Pergeseran Arah Transformasi Bisnis Media Cetak di Era Teknologi Informasi (Studi Kasus pada Harian Solopos). *AVANT GARDE: JURNAL ILMU KOMUNIKASI*, 11(02), 178–190.
- Ericko Wicaksono, Eristya Maya Safitri, M. Ailza Sifaul Anam, & Rizky Alamsyah Bimantara. (2023). Analisis Kepuasan Pengguna pada Aplikasi Gojek Menggunakan Pendekatan Metode DELONE-MCLEAN. *SITASI*.
- F. L. Hakim. (2023). Analisis Sentimen Berbasis Aspek Terhadap Produk Kecantikan Menggunakan Neighbor Weighted K-Nearest Neighbor (Nwknn). *Elibrary.Unikom.Ac.Id*.
- Fernando, E. H., & Toba, H. (2020). Pemanfaatan Latent Semantic Indexing untuk Mengukur Potensi Kerjasama Jurnal Ilmiah Lintas Universitas. *Jurnal Teknik Informatika Dan Sistem Informasi*, 6(3). <https://doi.org/10.28932/jutisi.v6i3.2894>
- Flores, V. A., Permatasari, P. A., & Jasa, L. (2020). Penerapan Web Scraping Sebagai Media Pencarian dan Menyimpan Artikel Ilmiah Secara Otomatis Berdasarkan Keyword. *Majalah Ilmiah Teknologi Elektro*, 19(2), 157. <https://doi.org/10.24843/mite.2020.v19i02.p06>
- Gifari, O. I., Adha, M., Rifky Hendrawan, I., Freddy, F., & Durrand, S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY)*, 2(1).
- Giyai, Y., Inan, D. I., & Baisa, L. Y. (2024). Evaluation of User Experience on Laporkitong Mobile Application Utilizing User Experience Questionnaire (UEQ). 4(3), 736–744. <https://doi.org/10.57152/malcom.v4i3.1292>

- Inan, D. I., Nizar Hidayanto, A., Juita, R., Andiyani, K., Hariyana, N., Tiffany, P., Prima Tangis Pertiwi, T., & Kurnia, S. (2022). Technology anxiety and social influence towards intention to use of ride-hailing service in Indonesia. *Case Studies on Transport Policy*, 10(3), 1591–1601. <https://doi.org/10.1016/j.cstp.2022.05.017>
- Ismail, Hanifah Nur Rahmah, & Rini Sulistiyowati. (2022). Penggunaan Software Orange Data Mining Pada Implementasi Text Mining Dalam Analisis Sentimen Netizen Di Twitter Terhadap Kelangkaan Minyak Goreng. *Jurnal.Polban.Ac.Id*.
- Matthew J. Denny, & Arthur Spirling. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Cambridge.Org*.
- Naury, C., Fudholi, D. H., & Hidayatullah, A. F. (2021). Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 5(1), 24. <https://doi.org/10.30865/mib.v5i1.2556>
- Sallu Muharomah. (2023). Topic Modelling Kasus Penipuan Di Twitter Menggunakan Metode Latent Dirichlet Allocation (LDA). *Dspace.Uii.Ac.Id*.
- Steorts, R. C. (2023). *A Primer on the Data Cleaning Pipeline*. <http://arxiv.org/abs/2307.13219>