Education on TikTok The Rise of Microlearning and Educational Content in Indonesia

Supervisor: Prof. Ingmar Weber

Alberto Martin Lionardi

Date: July 16, 2024

Contents

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken,_____   _____
             (Datum/Date)   (Unterschrift/Signature)

1 Abstract

The rising importance of social media in the current era of urbanization is undeniable, and one of the most prominent platforms with millions of active users is TikTok. Founded in 2016 and initially known as Douyin, TikTok has seen a steady increase in users since its launch, making it a fascinating subject for analysis. TikTok's user base is predominantly composed of individuals aged 18-24, a demographic that plays a significant role in shaping new generations. Indonesia, in particular, is heavily influenced by the platform, ranking second in the number of users.

While TikTok is primarily used for entertainment, this study explores the platform's educational content. To understand the relationship between demographic attributes across Indonesia, we propose a methodology for estimating TikTok's audience using the TikTok advertising platform. Our research reveals several key findings about the penetration and interest in educational content on TikTok across Indonesia.

Firstly, the penetration rate for educational content is relatively equal across rural, suburban, and urban areas. Interestingly, rural areas may exhibit a higher interest in educational content. Gender inequality in educational content interest exists on the platform, albeit with a small margin within each age group. However, age groups display a negative correlation between younger users and educational engagement.

Device-wise, audiences using iOS devices show greater engagement with educational content compared to those using Android devices. Additionally, better internet connectivity plays a significant role in fostering interest in educational content.

Moreover, we find that the number of universities or institutes in a province does not necessarily indicate a higher interest in educational content. In contrast, the number of students in a province positively correlates with engagement in educational content.

In summary, our study highlights the diverse engagement with educational content on TikTok across different demographic groups in Indonesia, emphasizing the platform's potential as a tool for educational outreach.

2 Introduction

As research on the impact of social media [Cristina M. Pulido, 2018, A. Verner Venegas-Vera1, 2020] is still ongoing, studies have shown mixed outcomes. Some studies have found that excessive use of social media can lead to distractions and decreased face-to-face interactions, negatively impacting knowledge sharing [Chen, 2022]. However, social media can also be a valuable tool for students to collaborate and share information, enhancing their knowledge development [Arunkumar Sivakumar, 2023].

Social media has played a crucial role in education Tess [2013], especially during crises like COVID-19 [Mr. Wesam Taiel Farhan AL-Bashabsheh, 2023]. Unlike Instagram, Snapchat, Facebook, or other platforms, TikTok emphasizes video and audio content, making educational material more engaging [Liu, 2022], particularly for younger generations [Ascencio, 2022]. TikTok also serves as a valuable resource for scientific studies [Miltsov, 2022].

According to TikTok, Indonesia ranks second in terms of user numbers, with over 100 million active users [**?**]. This paper will primarily focus on TikTok, given its significant and steady increase in user numbers over the past four years (see Figure **??**). Projections indicate that this growth will continue in the coming years, making TikTok an invaluable resource for research purposes.

(a) TikTok User Indonesia 2021-2024

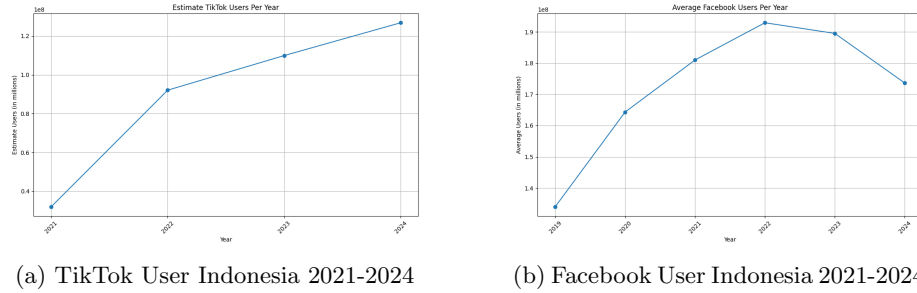(b) Facebook User Indonesia 2021-2024

Figure 1: TikTok Vs Facebook User 2021-2024

Since COVID-19, online learning has become one of the major teaching methods used by education systems around the world [Asaqli, 2020]. This shift to digital platforms has brought about numerous advantages [Rdouan Faizi and Chiheb, 2013a]. Online learning is often less costly, takes less time, and can be more engaging and entertaining for students. Additionally, it provides flexibility, allowing students to learn at their own pace and from any location.

However, these benefits are accompanied by several disadvantages. One significant drawback is the isolation that comes with the lack of social contact, which can impact students' mental health and development. The virtual environment also raises concerns about academic integrity, as it is easier for students to cheat. Furthermore, there are challenges related to the digital divide, where students without reliable internet access or appropriate devices are at a disadvantage.

The role of social media in education has also evolved significantly due to the pandemic. As discussed in the paper [Dr. Aarti Sharma, 2021], there has been a notable change in social media usage for learning before and after the lockdown. Data collected through surveys among students indicate that social media platforms have become crucial tools for communication, collaboration, and information sharing in the educational context. This shift highlights the increasing importance of digital literacy and the need for education systems to adapt to these new modes of learning and interaction.

This paper presents the first large-scale analysis of user demographics on TikTok in Indonesia, utilizing the platform's advertising data (see also [Miltsov, 2018, Wenqing Qian, 2023] for methodologies used on other platforms). We should also consider the impact of digital inequality on TikTok, particularly regarding the gender gap, as discussed by Laura Robinson [2015][David, 2018], which highlights the significance of digital inequalities in access, usage, skills, and self-perceptions, crucial for understanding broader social inequalities.

We contribute to the current state of the art in social media analytics. First, we provide a systematic data and in-depth analysis of TikTok advertising audience estimates for a broad range of user interests, behaviors, and demographic attributes across Indonesia. Our focus is on analyzing educational content on the platform. TikTok has an "Education" interest tag to indicate audience engagement in each area, but the targetable audience from this tag doesn't provide conclusive evidence of genuine interest due to the lack of supporting data.

To provide a more accurate estimate, we analyze educational-related hashtags. This method makes more sense and is validated by our secondary approach: scraping videos with educational hashtags and assessing how many are genuinely educational (More details in Chapter 3).

2.1 Research Questions

In this study, we examine TikTok Advertising as a resource for measuring the educational content in Indonesia, a country in which 30% of the population lives in rural areas. In particular, we address three major research questions:

- How does gender influence engagement with educational content across different age groups on TikTok?

- How do Indonesian demographic attributes relate to educational content?

- What are the penetration rates for educational content across rural, suburban, and urban areas?

- Do the number of universities and students in an Indonesian province impact educational content engagement on TikTok?

2.2 Related Work

Educational content on social media has gained significant attention in recent years due to its potential benefits. There are multiple factors supporting the usage of social media for educational purposes [Rinku Sanjeev, 2021]. Firstly, social media platforms provide a dynamic and interactive environment that fosters engagement and participation among students. Research by various scholars has shown that social media is an effective way to engage students, especially those with certain personality traits such as shyness, intimidation, or boredom. For instance, [Rdouan Faizi and Chiheb, 2013b] demonstrates that social media enables these students to share their ideas and participate in class discussions more freely, thereby enhancing their learning experience.

Furthermore, educators find social media to be a valuable strategy for information dissemination. According to [Risa Blair, 2014], social media platforms allow educators to share educational resources, updates, and important information efficiently and effectively. This method not only broadens the reach of educational content but also facilitates real-time communication and collaboration among students and educators.

The growing importance of research on the use of social media for educational purposes highlights its relevance in the current era. As technology continues to evolve, understanding the impact and potential of social media in education becomes increasingly critical. By leveraging social media, educators can create a more inclusive and engaging learning environment that caters to the diverse needs of students, thereby enhancing the overall educational experience.

There is substantial research analyzing countries using social media advertising platforms. [Wenqing Qian, 2023] explores demographic characteristics in China using the Weibo advertisement platform. We adopted the methodology from this study to analyze Indonesian digital inequality on TikTok, confirming the finding that young females (under 30 years old) are more likely to use social media.

The methodology from [Stalin Figueroa, 2021], which involves data scraping, video indexing, and multi-label categorization to identify educational content preferences on TikTok, inspired us to support our findings from the advertisement platform by targeting hashtags. However, our attempt to replicate the video indexer process was unsuccessful due to its limitation in detecting Indonesian spoken content. Nonetheless, data scraping provided additional support for our methodology.

A key aspect of our analysis focuses on categorizing Indonesian areas by municipalities into rural, suburban, and urban based on their "richness." We followed the methods outlined in [Daniele Rama, 2020] to investigate inequality at both the regency and province levels in Indonesia. A significant challenge we faced was the less "structured" and "organized" nature of available data in Indonesia compared to developed countries. While offline data was available at the province level, it was not readily accessible at the regency level, necessitating a two-tiered analysis approach.

While the report from [Fathul Qorib, 2020] on year 2020 finds that TikTok

usage as educational purpose is the lowest amongst other social media by using survey on students in Indonesia, our research methodology proves that on 2023 the number of students in province matter on the relation with educational engagement in province level. This might prove that on 2020, TikTok was still on early stages, and over the years the contents and the user has been growing significantly and starting to use TikTok as an educational content platform. However, we do not provide analytics on surveys with students, and so not able to prove the comparison between social media usage for educational purpose.

The rapid adoption of TikTok has sparked numerous studies exploring the factors influencing user engagement and adoption. A significant study by [Xinjie Deng, 2023] developed an adoption model for TikTok by incorporating perceived boredom and personal innovativeness into the Hedonic-Motivation System Adoption Model (HMSAM). Their research demonstrated that perceived innovativeness significantly and positively influences perceived ease of use, suggesting that users who view themselves as innovative are more likely to find TikTok easy to use. This ease of use, in turn, positively impacts behavioral intention and focused immersion through the mediating effects of curiosity and perceived boredom. Additionally, the study found that educational level moderates the relationship between joy and focused immersion, indicating that users with different educational backgrounds experience engagement differently on the platform.

Building on these findings, our study aims to further explore the demographic factors influencing TikTok usage, specifically focusing on age, gender, device OS, connection type, and regional differences. By analyzing data from various demographic groups, we seek to understand how these factors impact users' engagement with educational content on TikTok. Our approach extends the existing literature by examining the role of specific hashtags and content categories, providing a nuanced understanding of how educational material is consumed on the platform.

3 Methodology

3.1 Census Data

Indonesia is divided into 38 provinces and further into 516 regencies. We collect population data from each region to group the areas into rural, sub-urban, and urban categories. In order to work with our TikTok advertising platform, we need to cross-check data that are available on both platforms. We then send the available location IDs to the API and get the audience estimate. Below are the available offline data at the regency level and province level. Both data sets are gathered from the official government website `https://bpks.go.id/` and are from the year 2023.
Offline data that we could target in regency level are :

- Total Population in each Regency

- Poverty Rate in each Regency

Offline data that we could target in province level are :

- High School Completion Rate

- Poverty Rate in each Province

- Number of University / Institute

- Number of Students

- Total Teachers and Professors

As discussed in [Daniele Rama, 2020], we employ the same methodology for classifying both regencies and provinces into three categories based on their 'Poverty Rate.' The categories are defined as follows: 0 to quartile 1 (Q1) as urban, quartile 1 (Q1) to quartile 3 (Q3) as suburban, and above quartile 3 (Q3) as rural. A higher Poverty Rate indicates a greater inclination towards rural areas.

3.2 TikTok Advertisment Platform

TikTok Advertising audience estimates are available publicly see in Appendix 13, using this platform we could send a request to the API using Python package. For all queries, we request a number of user with certain demographic attributes (e.g Age Groups, Device OS, Location, etc). There are several advertising campaign available focusing on certain objective such as "Reach", "Traffic", "Community Interaction") in this case, we will choose "Reach" as we are interested in "maximum number of people". After performing the API call, we receive an audience estimate response based on the monthly active users on TikTok. The lowest value the API can return is 0-1000. There are regencies in Indonesia with specific demographic attributes (for example, Banda Aceh 3G

Users) that return this minimum audience estimate. To address this issue, we combine these regencies with other regencies within the same province to obtain a more accurate estimate.

3.3   Targeting

We target various attributes to understand about the relation with educational content across regencies in Indonesia. We list the attributes below :

- Gender (Male, Female)

- Cell Network (3G, 5G, WiFi)

- Cellphone OS (Android, iOS)

- Age Groups (18-24, 25-34, 35-44, 45-54, 55+)

- Language (English, Indonesia)

In order to analyse Educational content in Indonesia, ideally we target the "Education" tag provided by the advertisement platform in the section of "Interest", Figure 2 shows the proportion of user in Indonesia that we could target by selecting specific Interests.
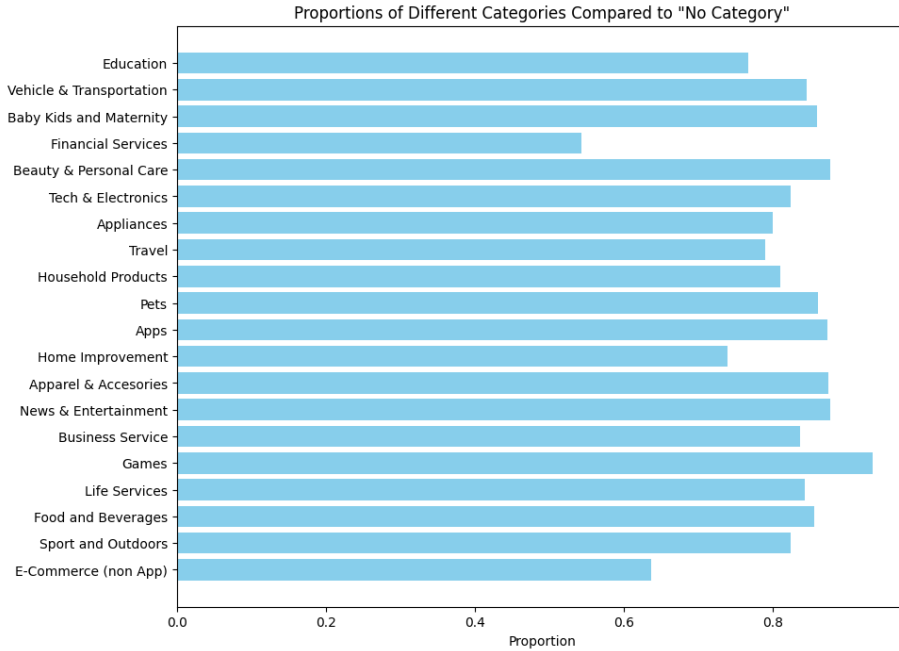

Figure 2: Audience Estimate gathered by selecting specific Interests

10

As we could see from selecting the Education tag we could reach almost 80% of the Indonesian audience. However, this tag doesn't provide much information (e.g what kind of education it is ?) and we should also consider the unknownity of the TikTok algorithm on gathering these number of audience. Therefore the estimate by using these Interests tag leaves us a lot of questions in there. Thus, as alternative we propose a solution by targeting hashtags. Targeting hashtag returns us fewer number of audience, however it benefits us to answer the previous question. As we are interested in Educational Content, we then target educational-related hashtag from the TikTok ads Platform using a set of hashtag in Indonesian language as additional variable as listed below :

- Art (Draw, Guitar, Dance)

- Language (Korean, Japanese, English ,Mandarin, Arabic)

- Life Skills / Hobby (Make up, parenting, public speaking, cooking)

- General Education Hashtag (Onlinelearning, Learntogether)

- Academic Subjects (Law, Psychology, History)

We then combine the data and model them for analysis using regression.

Once we have the demographic attributes and the interests identified, we then select the location. The TikTok Advertising Platform allows us to gather audience data at the regency level in Indonesia. However, not all regencies are available on the platform. Out of the total 516 regencies/cities in Indonesia, the platform provides data for only half of them. Figure 5 shows the regencies we could target, colored by the proportion of users we could reach.
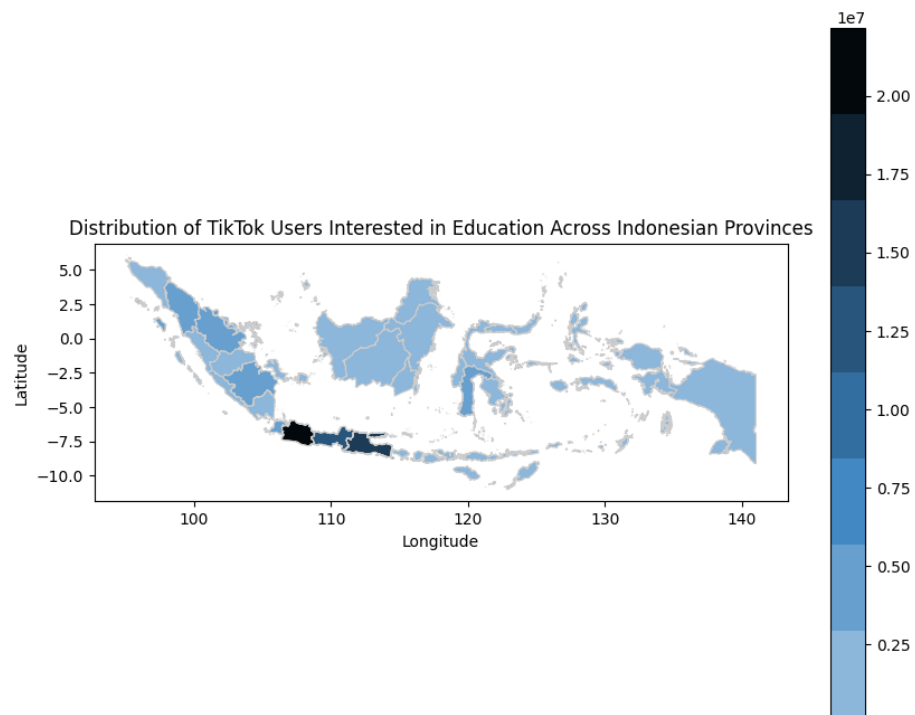
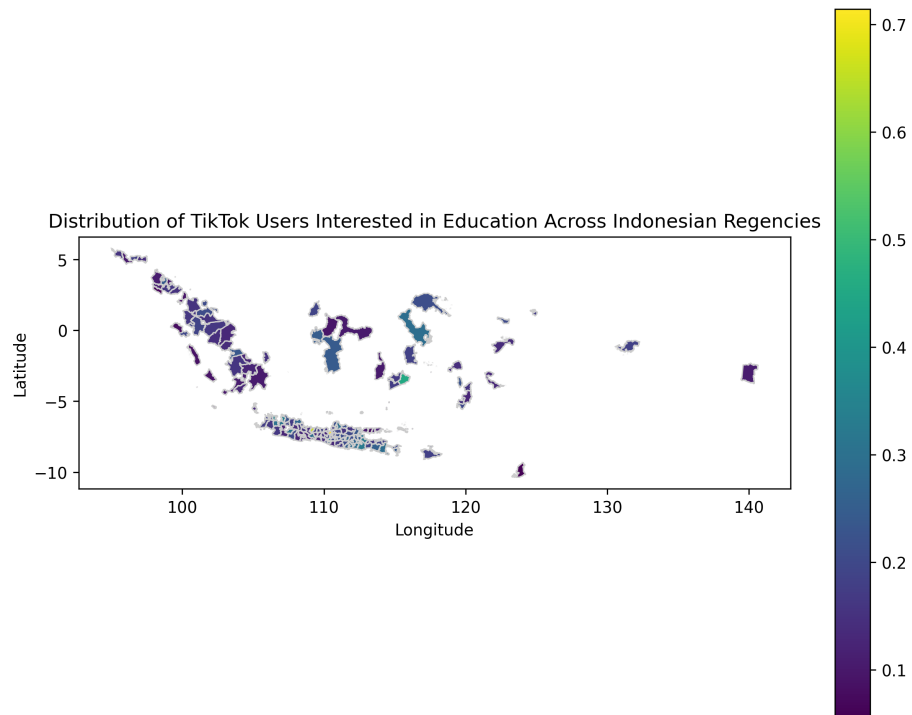Figure 3: TikTok User Across Indonesia by Provinces

Figure 4: Available Regions(Regencies) For Data Collection Colored by Penetration Rate in Educational Content (Interest in Education/All User)
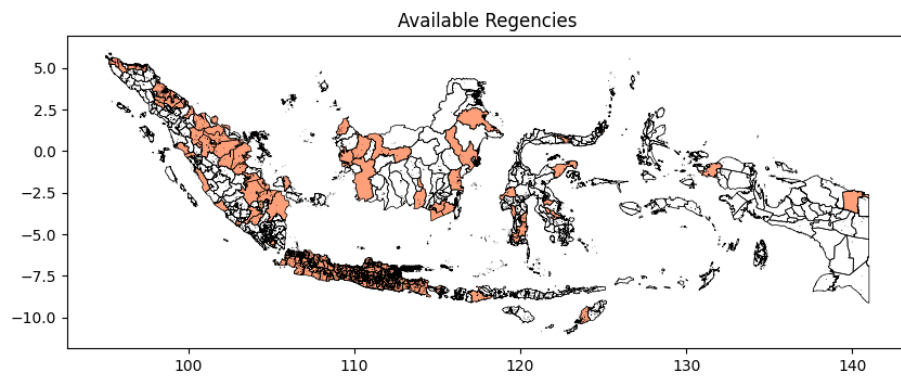

Figure 5: Clearer Map for Available Regencies

3.4 Scraping

To support our findings from hashtags, we conducted a comprehensive data collection from TikTok. Using Apify (Apify Actors Console), from this website, we put desired hashtags (educational content in this case) and started scraping. After 2-3 minutes (depending on how much video we would like to collect), we export the result in JSON for the attribute "webVideoUrl" in order to collect all the URLs. Example of collected webVideoUrl

webVideoUrl: "https://www.tiktok.com/@itechacademy/video/7251896828428307717"

Next, to automate the verification process, we utilized Selenium to check whether the content was indeed educational.

In this process, we collected a sample of 100-300 videos from each hashtag. For each video, we analyzed the comments to identify the presence of specific Indonesian words indicating that the video was perceived as helpful. We determine the "trueness" of educational content by examining user comments. For example, if comments such as "thank you for the mathematical course" are present, we increase the positive count and move to the next video.

However, the automation process faced a significant challenge with reCaptcha, which our program could not solve independently. This issue required manual intervention to complete the reCaptcha verification.

Ultimately, we calculate the percentage of videos that are truly educational. For instance, if 80% of videos with the hashtag #belajar are educational, we use this percentage to gauge the content's educational value

Table 1 presents the results of our analysis, showing the number of videos iterated for each hashtag and the corresponding findings. It is important to note that while the hashtags shown in the table are translated into English for clarity, the actual iteration used Indonesian hashtags to ensure that Apify collected videos specifically from Indonesia.

Table 1: Table of Hashtag Statistics

Hashtag	Positive_Percentage	Number of Videos
#education hashtag general (#belajar)	89.0	300
#language learning (#belajarbahasa)	79.0	100
#math(#belajarmatematika)	74.5	150
#draw(#belajargambar)	76.0	150
#guitar(#belajargitar)	58.0	150
#onlinelearning(#belajaronline)	88.0	100
#tipsandtrick(#caracepat)	70.0	200

The average percentage that a hashtag contains educational content is 76.21. This result will strengthen our next methodology specifically for collecting hashtags, and using the hashtag as target variable.

14

3.5 Quality Check

In addition to scraping, we also perform a quality check of the audience estimates from the advertisement platform. As illustrated in Figure 6, the number of male and female audiences interested in specific hashtags is analyzed. The results indicate that hashtags related to travel, shopping, and food are predominantly favored by the female audience. Conversely, male audiences exhibit a stronger interest in gaming content, reinforcing the common perception that males are more inclined towards gaming and female towards travel, shopping , and food. This quality check helps validate the data by confirming expected trends in audience interests.

(a) Interest in Gaming Hashtag

(b) Interest in Traveling Hashtag

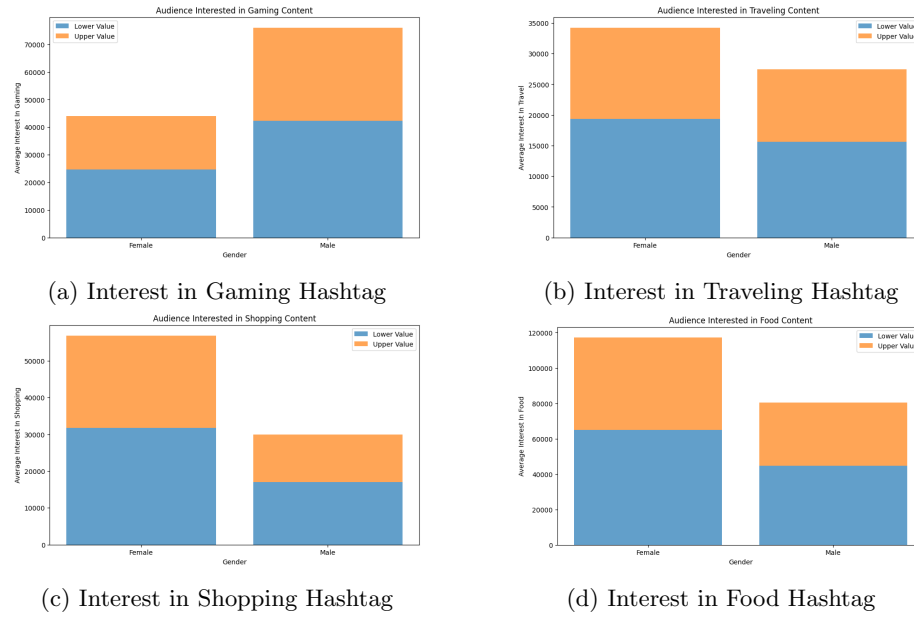(c) Interest in Shopping Hashtag

(d) Interest in Food Hashtag

Figure 6: Quality Check Performed in Placebo Hashtags

3.6 Gender Inequality

Graph 7 shows that in the younger age groups (18-24 and 25-34), there is a noticeable gender disparity, with a higher proportion of female users engaging with educational content. In the older age groups, the proportions of male and female users are more balanced, suggesting less gender disparity. The higher engagement of females in the younger age groups might indicate a greater interest in educational content among young women compared to young men. However, this trend does not persist in the older age groups.
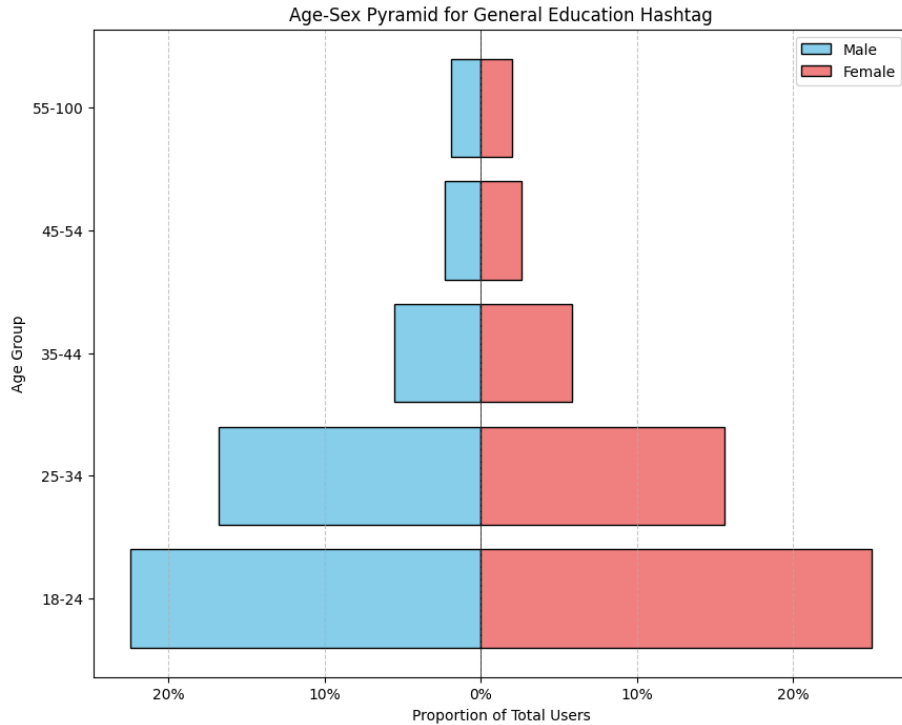

Figure 7: Age Gender Inequality in TikTok

3.7 Penetration Rate Educational content

To address our previous question regarding the penetration rate of educational content across regencies, we created scatter plots 8 to analyze the relationship between educational-related hashtags and the total number of users in each regency. The classification is based on the "Poorness" rate in each regency, where a high "Poorness" rate categorizes an area as rural.

Rural Areas Show Higher Penetration Rates

Contrary to common assumptions, urban areas do not always exhibit a higher penetration rate for educational content. In fact, rural areas show a higher penetration rate compared to suburban and urban areas. This finding challenges the notion that urban areas, with their larger student populations and better access to educational resources, would naturally have higher engagement with educational content.

Non-Metropolitan Areas Leading Engagement

Interestingly, the regencies with the highest penetration rates (user with interest in hashtag divided by total population in each areas) for each hashtag category are not among the major metropolitan areas in Indonesia. This suggests that smaller or less populated regions may have communities that are more actively engaging with specific types of educational content see 2.
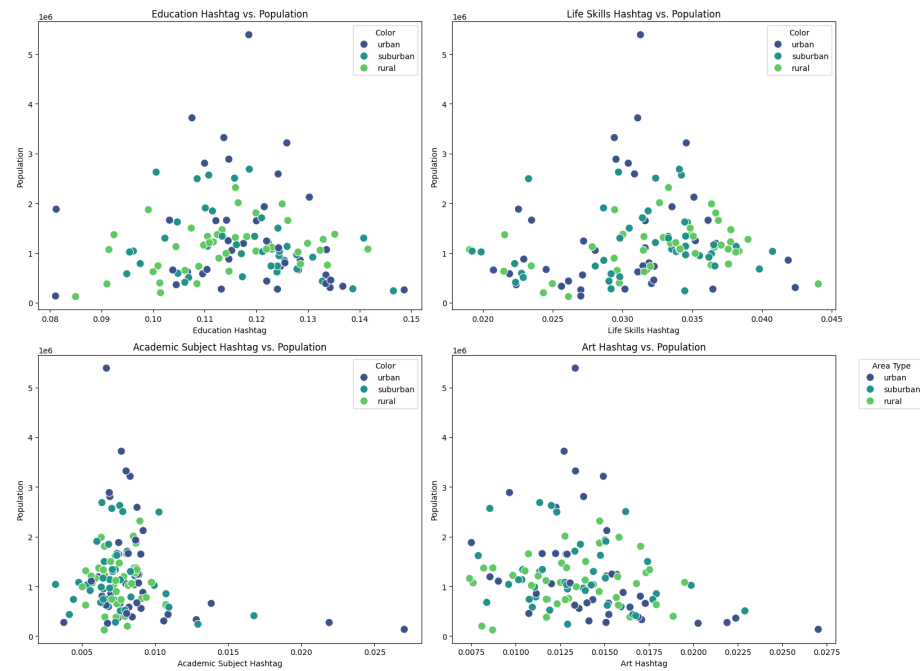

Figure 8: Penetration Rate for each Educational-Related Hashtag

Table 2: Highest Rates for Different Hashtags across Various Area Types

Hashtag Type	Area Type	Rate	Highest Location	Explanation
Education	Rural	0.1416	Bantul	Highest rate for Education Hashtag
	Urban	0.1486	Berau	
	Suburban	0.1466	Tanjung Pinang	
Life Skills	Rural	0.0441	Kupang	Highest rate for Life Skills Hashtag
	Urban	0.0424	Palangkaraya	
	Suburban	0.0419	Kudus	
Academic Subject	Rural	0.0108	Musi Banyuasin	Highest rate for Academic Subject Hashtag
	Urban	0.0270	Pulang Pisau	
	Suburban	0.0219	Pematangsiantar	
Art	Rural	0.0223	Ogan Komering Ilir, Ulu	Highest rate for Art Hashtag
	Urban	0.0270	Pulang Pisau	
	Suburban	0.0229	Labuhanbatu	

3.8 OLS Model Regency Level

We employed an OLS regression model (3) with final formula

$$
\begin{aligned}
\text{Education Hashtag} = 0.0778 \\
- 0.4684 \cdot (18\text{-}24) \\
- 0.4688 \cdot (25\text{-}34) \\
+ 0.6251 \cdot (2G) \\
- 0.2962 \cdot (35\text{-}44) \\
- 0.4802 \cdot (45\text{-}54) \\
+ 0.6289 \cdot (4G) \\
- 0.6679 \cdot (55\text{-}100) \\
- 0.0868 \cdot (\text{Academic Subject Hashtag}) \\
- 0.1197 \cdot (\text{Android}) \\
+ 0.8335 \cdot (\text{Art Hashtag}) \\
+ 0.5725 \cdot (\text{English}) \\
- 0.5939 \cdot (\text{English, Indonesia}) \\
+ 0.1292 \cdot (\text{Female}) \\
+ 0.1281 \cdot (\text{Food Hashtag}) \\
+ 0.2680 \cdot (\text{Gaming Hashtag}) \\
+ 0.6890 \cdot (\text{Indonesia}) \\
+ 0.1874 \cdot (\text{Language Hashtag}) \\
+ 0.4057 \cdot (\text{Life Skills Hashtag}) \\
+ 0.1443 \cdot (\text{Male}) \\
+ 0.0092 \cdot (\text{Poorness}) \\
+ 0.1194 \cdot (\text{Shopping Hashtag}) \\
- 0.0940 \cdot (\text{Sport Hashtag}) \\
+ 0.1302 \cdot (\text{Travel Hashtag}) \\
+ 0.6162 \cdot (\text{Wi-Fi}) \\
+ 0.0287 \cdot (\text{iOS})
\end{aligned}
$$

to examine the impact of multiple predictor variables on the dependent variable, "Education Hashtag" which is the percentage of user that engage in the hashtag per regions (User Interest in Education Hashtag / Total User). The predictors included age groups, connectivity types, device usage, and various content hashtags. The data consisted of 158 observations. The regression analysis yielded an R-squared value of 0.829, indicating that approximately 82.9% of the variance in "Education Hashtag" is explained by the model. The adjusted R-squared was 0.796, further confirming the model's robustness.

Significant Predictors

- **2G** ($\beta = 0.6251$, $p = 0.002$): Positive association with educational engagement.

- **4G** ($\beta = 0.5791$, $p = 0.007$): Positive association with educational engagement.

- **55-100** ($\beta = -0.6679$, $p = 0.033$): Negative association with educational engagement.

- **Art Hashtag** ($\beta = 0.8335$, $p < 0.001$): Positive association with educational engagement.

- **English, Indonesia** ($\beta = -0.9539$, $p = 0.001$): Negative association with educational engagement.

- **Food Hashtag** ($\beta = 0.1281$, $p = 0.012$): Positive association with educational engagement.

- **Gaming Hashtag** ($\beta = 0.2680$, $p = 0.001$): Positive association with educational engagement.

- **Life Skills Hashtag** ($\beta = 0.4057$, $p = 0.009$): Positive association with educational engagement.

- **Wi-Fi** ($\beta = 0.6162$, $p = 0.002$): Positive association with educational engagement.

Marginally Significant Predictors

- **18-24** ($\beta = -0.4684$, $p = 0.063$): Negative association with educational engagement.

- **25-34** ($\beta = -0.4688$, $p = 0.065$): Negative association with educational engagement.

- **Travel Hashtag** ($\beta = 0.1302$, $p = 0.059$): Positive association with educational engagement.

Non-Significant Predictors

Predictors such as **35-44**, **45-54**, **Academic Subject Hashtag**, and **Android** did not show statistically significant associations with the dependent variable.

Interpretation

Our analysis supports the claim made by [Xinjie Deng, 2023] that gender does not significantly influence educational engagement on social media platforms. This finding is consistent with our data, which shows no substantial gender-based differences in interactions with educational content.

Additionally, our study highlights the critical role of internet connectivity in enhancing engagement with educational content. Improved internet access facilitates greater usage of social media for educational purposes. According to [Kemp, 2023], internet accessibility in Indonesia reached 77% of the population by 2023 and has been steadily increasing over the years. This trend suggests that as internet penetration continues to grow, we can anticipate a corresponding increase in educational engagement on social media platforms. Ensuring widespread and reliable internet access remains a key factor in promoting educational content consumption and overall digital literacy.

Moreover, our findings indicate that English-language content exhibits notably low engagement, likely due to language barriers among users. Additionally, our analysis reveals that the type of device used, whether Android or iOS, does not significantly impact educational engagement. This supports the claims made by [Zulvia Trinova, 2022], which suggest that access to education via social media is not hindered by social status. These insights underscore the importance of considering language preferences and accessibility in the development of educational content for social media platforms.

3.9   OLS Model Regency Level Hashtag and Demographic Attribute

In order to gain a deeper understanding of the factors influencing engagement with educational content, we split the model into two separate analyses: Hashtag Attributes and Demographic Attributes. The results of these models are presented in Tables 4 and 5.

The Hashtag Model (Table 4) demonstrates a higher R-squared value compared to the Demographic Model, indicating that hashtags related to specific topics or content categories have a stronger predictive power for engagement with educational content. This suggests that users who interact with certain hashtags are more likely to engage with educational hashtags. Key findings include significant positive coefficients for hashtags related to language, life skills, food, travel, and gaming, indicating that these content areas are particularly engaging for educational content. Demographic factors in other hand, also play a role but are less influential compared to content. This suggests that while demographic factors are important, understanding what content users engage with provides deeper insights.

Table 3: OLS Regression Result

Dep. Variable:	Education Hashtag	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.796
Method:	Least Squares	F-statistic:	25.52
Date:	Sat, 08 Jun 2024	Prob (F-statistic):	5.79e-39
Time:	14:44:33	Log-Likelihood:	595.39
No. Observations:	158	AIC:	-1139.
Df Residuals:	132	BIC:	-1059.
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.0778	0.033	2.389	0.018	0.013	0.142
18-24	-0.4684	0.250	-1.873	0.063	-0.963	0.026
25-34	-0.4688	0.252	-1.860	0.065	-0.967	0.030
2G	0.6251	0.200	3.131	0.002	0.230	1.020
35-44	-0.2962	0.263	-1.127	0.262	-0.816	0.224
45-54	-0.4082	0.257	-1.591	0.114	-0.916	0.099
4G	0.6289	0.230	2.735	0.007	0.174	1.084
55-100	-0.6679	0.310	-2.157	0.033	-1.280	-0.055
Academic Subject Hashtag	-0.0868	0.282	-0.308	0.759	-0.645	0.472
Android	-0.1197	0.308	-0.389	0.698	-0.728	0.489
Art Hashtag	0.8335	0.231	3.601	0.000	0.376	1.291
English	0.5725	0.413	1.388	0.168	-0.244	1.389
English , Indonesia	-0.9539	0.295	-3.239	0.002	-1.536	-0.371
Female	0.1292	0.392	0.329	0.742	-0.647	0.905
Food Hashtag	0.1281	0.050	2.560	0.012	0.029	0.227
Gaming Hashtag	0.2680	0.078	3.430	0.001	0.113	0.423
Indonesia	0.6891	0.413	1.669	0.098	-0.128	1.506
Language Hashtag	0.1874	0.120	1.556	0.122	-0.051	0.426
Life Skills Hashtag	0.4057	0.154	2.632	0.009	0.101	0.711
Male	0.1443	0.390	0.370	0.712	-0.628	0.916
Poorness	0.0092	0.020	0.465	0.643	-0.030	0.048
Shopping Hashtag	0.1191	0.082	1.459	0.147	-0.042	0.281
Sport Hashtag	-0.0940	0.052	-1.796	0.075	-0.198	0.010
Travel Hashtag	0.1302	0.068	1.904	0.059	-0.005	0.265
Wi-Fi	0.6162	0.197	3.122	0.002	0.226	1.007
iOs	0.0287	0.309	0.093	0.926	-0.583	0.641

Omnibus:	11.842	Durbin-Watson:	1.872
Prob(Omnibus):	0.003	Jarque-Bera (JB):	13.470
Skew:	0.540	Prob(JB):	0.00119
Kurtosis:	3.939	Cond. No.	3.57e+03

Table 4: OLS Regression Result Hashtag Interaction

Dep. Variable:	Education Hashtag	R-squared:	0.632
Model:	OLS	Adj. R-squared:	0.602
Method:	Least Squares	F-statistic:	21.71
Date:	Mon, 01 Jul 2024	Prob (F-statistic):	5.49e-21
Time:	16:49:21	Log-Likelihood:	424.50
No. Observations:	124	AIC:	-829.0
Df Residuals:	114	BIC:	-800.8
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	-0.0090	0.011	-0.802	0.424	-0.031	0.013
Academic Subject Hashtag	-0.7177	0.303	-2.369	0.020	-1.318	-0.117
Art Hashtag	0.5341	0.298	1.795	0.075	-0.055	1.124
Language Hashtag	0.5423	0.148	3.665	0.000	0.249	0.835
Life Skills Hashtag	0.5000	0.181	2.760	0.007	0.141	0.859
Food Hashtag	0.1210	0.056	2.177	0.032	0.011	0.231
Travel Hashtag	0.1812	0.080	2.262	0.026	0.023	0.340
Sport Hashtag	-0.1853	0.064	-2.881	0.005	-0.313	-0.058
Gaming Hashtag	0.4458	0.093	4.802	0.000	0.262	0.630
Shopping Hashtag	0.0544	0.103	0.528	0.599	-0.150	0.258

Omnibus:	12.484	Durbin-Watson:	1.809
Prob(Omnibus):	0.002	Jarque-Bera (JB):	17.566
Skew:	0.528	Prob(JB):	0.000153
Kurtosis:	4.512	Cond. No.	521.

Table 5: OLS Regression Result Demographic Attribute

Dep. Variable:	Education Hashtag	R-squared:	0.486
Model:	OLS	Adj. R-squared:	0.415
Method:	Least Squares	F-statistic:	6.818
Date:	Mon, 01 Jul 2024	Prob (F-statistic):	3.74e-10
Time:	16:49:21	Log-Likelihood:	403.91
No. Observations:	124	AIC:	-775.8
Df Residuals:	108	BIC:	-730.7
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.1567	0.053	2.941	0.004	0.051	0.262
2G	0.2411	0.304	0.793	0.429	-0.361	0.843
4G	0.6490	0.354	1.835	0.069	-0.052	1.350
Wi-Fi	0.3446	0.296	1.163	0.247	-0.243	0.932
Female	1.1501	0.633	1.818	0.072	-0.104	2.404
Male	1.1858	0.629	1.885	0.062	-0.061	2.433
English	-0.4678	0.672	-0.696	0.488	-1.800	0.864
English , Indonesia	-0.4070	0.457	-0.891	0.375	-1.312	0.498
Indonesia	-0.2750	0.664	-0.414	0.680	-1.592	1.042
18-24	-0.3483	0.409	-0.851	0.397	-1.160	0.463
25-34	-0.6466	0.413	-1.566	0.120	-1.465	0.172
35-44	0.0937	0.422	0.222	0.825	-0.743	0.930
45-54	-0.1697	0.432	-0.393	0.695	-1.025	0.686
55-100	-0.9344	0.505	-1.849	0.067	-1.936	0.067
Android	-0.4603	0.487	-0.946	0.346	-1.425	0.505
iOs	-0.3444	0.496	-0.694	0.489	-1.328	0.639

Omnibus:	1.640	Durbin-Watson:	1.955
Prob(Omnibus):	0.440	Jarque-Bera (JB):	1.424
Skew:	0.262	Prob(JB):	0.491
Kurtosis:	3.012	Cond. No.	3.09e+03

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 3.09e+03. This might indicate that there are strong multicollinearity or other numerical problems.

3.10 Urban-Rural Disparities in Indonesian Regencies

We use the data to measure potential inequalities between the urban and rural regencies in Indonesia. In this experiment we exclude the suburban areas for more clarity, we substract the median attribute value of its rural from median of its urban municipalities. We then plot the distribution of these differences using a significance level of $p < 0.05$ (Note that the number of data points used here is small) The most significant difference from the difference in median, we find, are those associated with device operating system (there are more Android users in rural areas and iOS in urban areas)
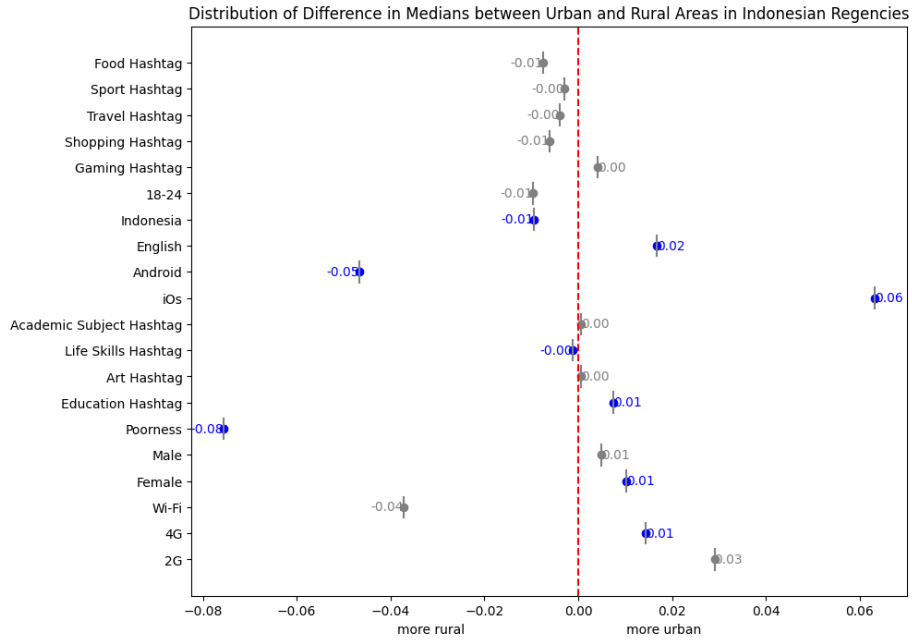
Figure 9: Distribution of difference in medians between urban and rural areas in Indonesian regencies for each attribute. Blue points show means, insignificant in grey

25

3.11 OLS Model Province Level

Due to some missing offline educational related data in regency level, we conduct analysis in province level to capture additional information. We use the #Educational_Hashtag as the target variable, similarly this hastag contains a set of Indonesian-Educational Hashtag which gathered from the advertisement platform. Note that this analysis 6 only contains 34 observations (provinces in Indonesia).
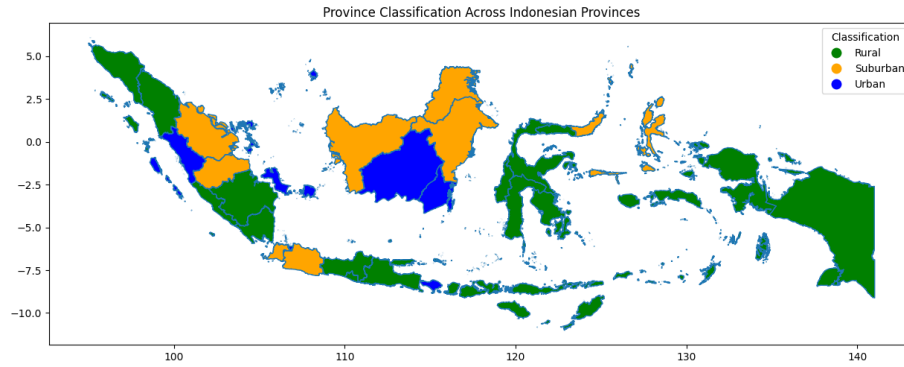
Figure 10: Provinces in Indonesia classified by Poorness Rate into Rural-Suburban-Urban

Key Findings

Overall Model Significance

The p-value for the F-statistic is 0.00367, which is less than the common significance level of 0.05. This suggests that the model explains a significant portion of the variance in #Educational_Hashtag.

Number of Universities

The coefficient for Number Of University is negative (-0.2581) and has a p-value of 0.246. Although it is not statistically significant $p > 0.05$, the negative coefficient suggests that an increase in the number of universities is associated with a decrease in #Educational_Hashtag. This could imply that more universities do not necessarily indicate higher interest in educational hashtags.

Number of Students

The coefficient for Number Of Students is positive (0.2413) and has a p-value of 0.430, which is not statistically significant. However, the positive coefficient suggests that an increase in the number of students is associated with an increase

in #Educational_Hashtag. This indicates that areas with more students might have higher engagement with educational hashtags.

Device Type

The coefficient for iOS is positive (0.2226) while the coefficient for Android is negative (-0.2157). Neither is statistically significant (p-values of 0.732 and 0.735, respectively). Nonetheless, these coefficients suggest that iOS users may have a higher interest in educational content compared to Android users. This finding also supports the assertion by [Zulvia Trinova, 2022] that access to online learning is affordable for all people in Indonesia, regardless of their financial situation.

Table 6: OLS Regression Result

Dep. Variable:	#Educational_Hashtag	R-squared:	0.525
Model:	OLS	Adj. R-squared:	0.397
Method:	Least Squares	F-statistic:	4.109
Date:	Sun, 09 Jun 2024	Prob (F-statistic):	0.00367
Time:	20:07:01	Log-Likelihood:	99.383
No. Observations:	34	AIC:	-182.8
Df Residuals:	26	BIC:	-170.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3415	0.627	0.545	0.590	-0.947	1.630
High School Completion Rate	-0.0270	0.037	-0.738	0.467	-0.102	0.048
Poor Population Percentage	-0.0832	0.063	-1.330	0.195	-0.212	0.045
Number Of University	-0.2581	0.218	-1.186	0.246	-0.706	0.189
Number Of Teacher/Professor	0.1654	0.420	0.394	0.697	-0.698	1.028
Number Of Students	0.2413	0.301	0.802	0.430	-0.378	0.860
Android	-0.2157	0.630	-0.342	0.735	-1.511	1.080
iOs	0.2226	0.642	0.347	0.732	-1.098	1.543

Omnibus:	1.248	Durbin-Watson:	2.140
Prob(Omnibus):	0.536	Jarque-Bera (JB):	0.413
Skew:	-0.170	Prob(JB):	0.813
Kurtosis:	3.420	Cond. No.	651.

3.12 ANOVA

The plot 11 describes the educational distribution across municipalities. There are significant variations in educational engagement in rural areas, however the median in urban and sub urban areas are still higher in compare to rural areas, the reason for this could be potentially due to more students in urban and suburban areas in compare to rural areas and also in general better connection type. Keep in mind however the p-value in this experiment is bigger than 0.05, indicating that the the difference between the municipalities could be not statistically significant.
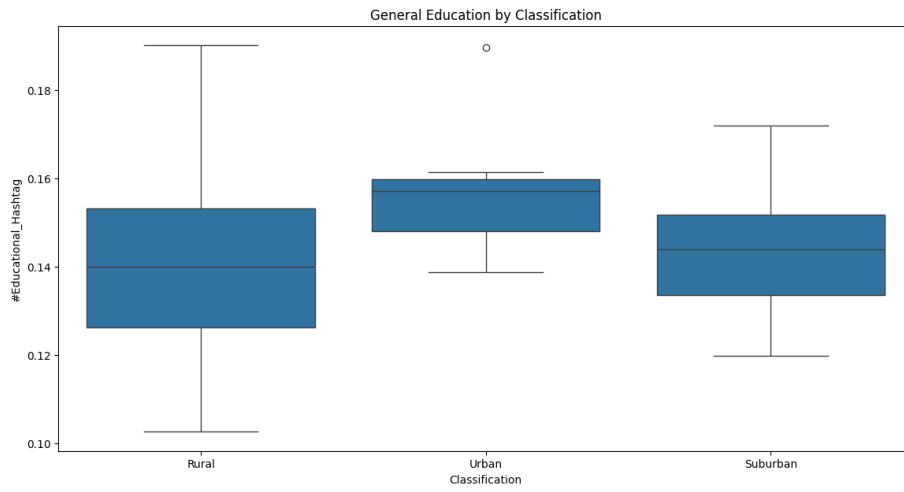
Figure 11: Boxplot for distribution of "Educational Content" engagement across Municipalities

3.13 Interest over Hashtags

Table 12 presents our findings on the interest in various hashtags in Indonesia. According to the report by [Muhammad, 2023], Indonesian users primarily seek entertainment, products, food, and travel content on social media. Our findings, however, based on targetable hashtags and interests that we could analyze, show a slightly different trend. While food-related content remains the most popular, educational content surpasses travel, shopping, and gaming hashtags in interest and is only slightly less popular than sport-related hashtags. It is important to note that our results are derived solely from targetable interests on the advertising platform, which may limit the overall validity of the findings.
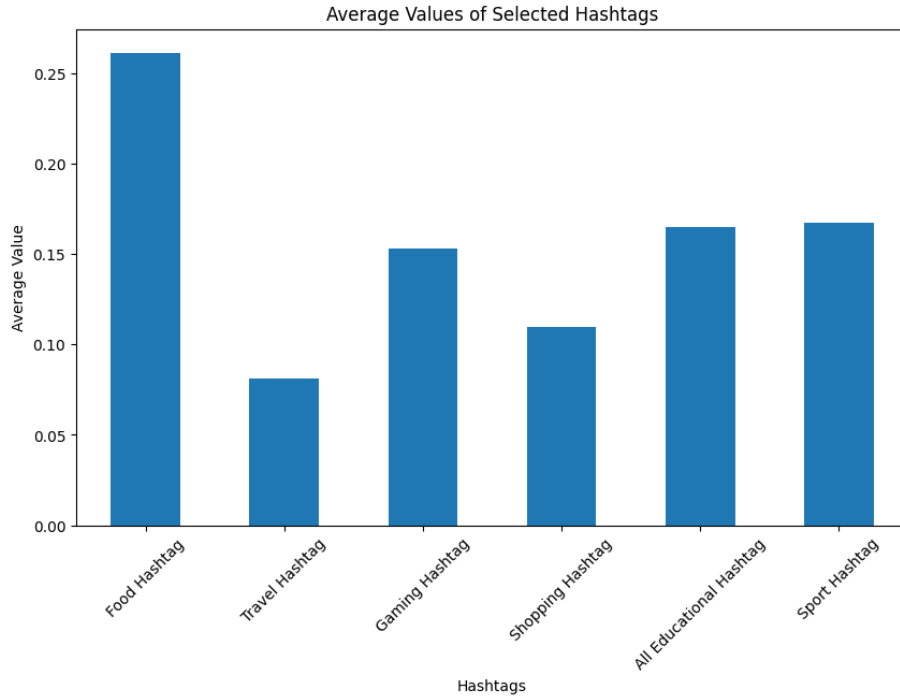
Figure 12: Hashtags Interest

3.14 Validity

From all the methodology and the results we have for analysis, there are also factors that need to be considered. One of it is the small number of observations in our models, specifically in province level. The resulting model from OLS do provides us an "image" of the relation between offline data and Educational Interest in TikTok, however we still need to keep in mind that these data are not statistically significant due to the small number of observations. Additionally we are still holding the assumption of the TikTok Advertisement Platform's Algorithm are targeting audiences correctly, we collected 5 data from friends and family which use TikTok daily in order to check the "Ad Interest" signed to the account, and 5 of them result in "N/A". Additionally inactive users and bots are still a factor to consider in the audience estimate. Moreover we are only targeting audience using the "registered" hashtag that are available in the advertisement platform, resulting in pretty low audience with interest in educational content across Indonesia.

4 Discussion

While our research highlights the correlation between various demographic variables with educational content in Indonesia through user estimate from the TikTok Ads Platform, it is essential to acknowledge the factors that were excluded from the methodology, such as TikTok audience estimate algorithms, existence of bots [Mariam Orabi, 2020][Richard J. Oentaryo, 2016][Ho-Chun Herbert Chang, 2020], number of inactive users. It is important for future papers to take these factors into account for more reliability. Other than that, comparing the findings with data from other social media platforms that is widely used in Indonesia such as Instagram and Facebook could validate our findings and finds platform-specific educational trends. Additionally, employing a video indexer [Stalin Figueroa, 2021] as a method to verify educational content could enhance the reliability of the confirmation process. Previous work by [Fathul Qorib, 2020] found that in 2020, TikTok had the lowest percentage of students seeking educational content. However, based on our findings, this trend may shift in the coming years. Given the rapid evolution of social media platforms and their increasing integration into educational contexts, it is plausible that the use of TikTok for educational purposes will grow. Conducting another iteration of surveys in future research could provide valuable insights into these changing patterns and the evolving role of social media in education. This ongoing research is essential to understand and adapt to the dynamic landscape of educational content consumption on social media platforms. Our research, focusing on Indonesia—a developing country with a higher proportion of rural areas than urban areas—demonstrates that engagement with educational content on TikTok is predominantly seen in rural and suburban regions. This highlights that students from various economic backgrounds can and do access educational materials via social media platforms, indicating a broad and inclusive reach of educational content beyond urban centers, which supports the findings of [A. Verner Venegas-Vera1, 2020] that income does not prevent students from engaging with educational content. Furthermore, this finding supports the conclusions of [Mücahit Aydoğmuş, 2023] and [Risa Blair, 2014] that social media is a proven and useful strategy for educators to effectively disseminate information and engage students.

5 Conclusion

To answer our research questions, we have discovered several key insights about user interest in educational content on TikTok. Our analysis indicates that the majority of users interested in educational content fall within the younger age group of 18-35 years, with a slight predominance of female users. This trend is likely due to the overall higher number of female users on the platform.

Demographically better internet connectivity has positive correlation with education interest, and educational related hashtag such as Art, Food, and Life Skills also has positive correlation with education, this might due to relation with learning, for example #howtodraw, or #howtocook. We find that however, younger age does not seem to have positive correlation with the educational engagement.

When examining user demographics across municipalities, we found that, aside from the prevalence of more Android users in rural areas and more iOS users in urban areas, the engagement with educational hashtags is nearly even, with only a slight difference in the median. This could explain the even distribution of penetration rates across Indonesia's regencies. In fact, there seems to be more engagement in rural and suburban areas compared to metropolitan areas.

Moreover, by incorporating offline data (number of universities, teachers/lecturers, and students) into our model, we found that the number of universities does not indicate a higher interest in educational hashtags, whereas the number of students and lecturers does. Additionally, iOS users potentially show more interest in educational content compared to Android users.

Despite these findings, it is important to consider them as a foundation for further research on educational content in Indonesia, rather than definitive conclusions. As highlighted in our methodology sections, there are still gaps and weaknesses in our research that need to be addressed.

6 ChatGPT Disclosure

In this section, we document the usage of ChatGPT, detailing how it was integrated into our research methodology, the ChatGPT usage in this paper was mainly for enhancing texts, and as a method for learning the machine learning algorithms in python. As an example we use ChatGPT to generate a sample code for performing OLS given a dataframe.

7 Appendix

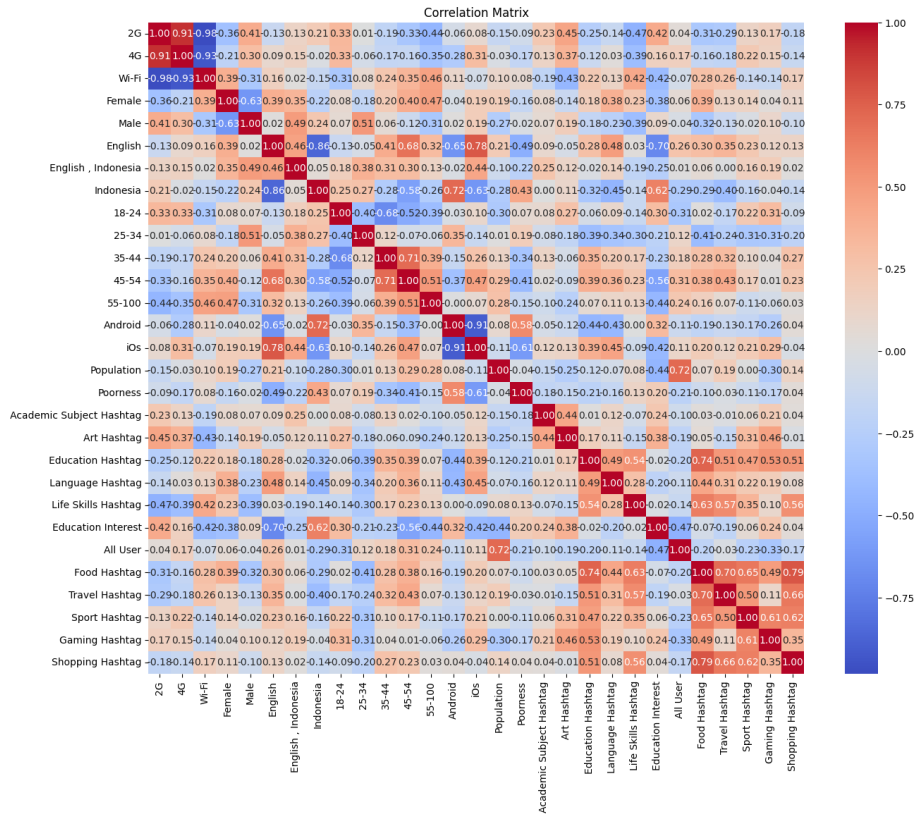Figure 13: Audience Estimate gathered by selecting specific Interests
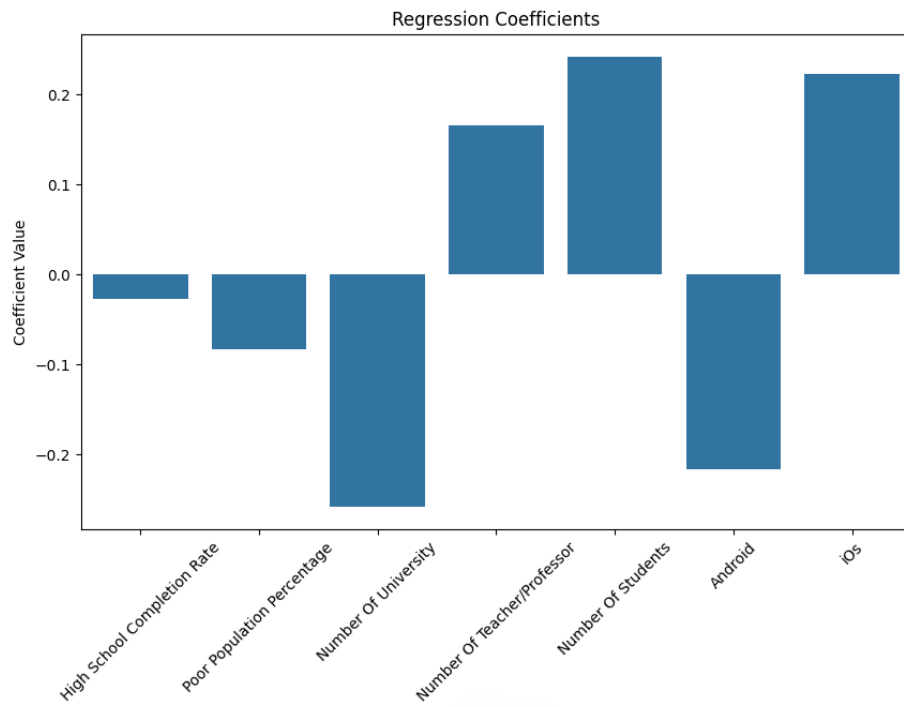
Figure 14: Correlation Matrix in Regency Level

Figure 15: Coefficients Value from OLS Province Level

References

Edgar V. Lerma A. Verner Venegas-Vera1, Gates B Colbert2. Positive and negative impact of social media in the covid-19 era. 2020.

Sudarsan Jayasingh Arunkumar Sivakumar. Social media influence on students' knowledge sharing and learning: An empirical study. 2023.

Eisam Asaqli. Online education: A change or an alternative? 2020.

Monica Bazan Ascencio. Discussion and guidelines about using tiktok to stimulate creativity in the classroom. 2022.

X Chen, M.; Xiao. The effect of social media on the development of students' affective variables. 2022.

Teresa Sordé-Martí Ramon Flecha Cristina M. Pulido, Gisela Redondo-Sama. Social impact in social media: A new method to evaluate the social impact of research. 2018.

Michele Tizzoni Kyriaki Kalimeri-Ingmar Weber Daniele Rama, Yelena Mejova. Facebook ads as a demographic tool to measure the urban-rural divide. 2020.

Angel Cuevas Manuel Cebrian-Esteban Moro Iyad Rahwan David, Yonas Mitike Kassa. Analyzing gender inequality through large-scale facebook advertising data. 2018.

Dr. Renu Jain Dr. Neena Mittal Dr. Aarti Sharma, Dr. Amrita Kaur. Education reform: Role of social media in education. 2021.

Jumrana La Tarifu Fathul Qorib, Sitti Utami Rezkiawaty Kamil. Reshaping today's education with social. 2020.

Meiqing Zhang Goran Muric Emilio Ferrara Ho-Chun Herbert Chang, Emily Chen. Social bots and social media manipulation in 2020. 2020.

Simon Kemp. Digital 2023: Indonesia. 2023.

Hiroshi Ono Anabel Quan-Haase Laura Robinson, Shelia R. Cotten. Digital inequalities and why they matter. 2015.

Yunqing Liu. Tiktok's influence on education. 2022.

Zaher Al Aghbari Ibrahim Kamel Mariam Orabi, Djedjiga Mouheb. Detection of bots in social media: A systematic review. 2020.

Alex Miltsov. Online health monitoring using facebook advertisement audience estimates in the united states: Evaluation study. 2018.

Alex Miltsov. Researching tiktok: Themes, methods, and future directions. 2022.

Dr. Pitabas Pradhan Mr. Wesam Taiel Farhan AL-Bashabsheh. The importance of social media in education during covid-19. 2023.

Nabillah Muhammad. Jenis konten media sosial yang sering diakses responden. 2023.

Yıldıray Karadağ Mücahit Aydoğmuş, Edip Tut. Teachers' experiences regarding the use of social media for educational purposes. 2023.

Abdellatif El Afia Rdouan Faizi and Raddouane Chiheb. Exploring the potential benefits of using social media in education. 2013a.

Abdellatif El Afia Rdouan Faizi and Raddouane Chiheb. Exploring the potential benefits of using social media in education. 2013b.

Philips K. Prasetyo Ee-Peng Lim Richard J. Oentaryo, Arinto Murdopo. On profiling bots in social media. 2016.

Nidhi Shridhar Natrajan Rinku Sanjeev. A systematic review on education 4.0 using social media platform. 2021.

Tina M. Serafini Risa Blair. Integration of education: Using social media networks to engage students. 2014.

Angel Fiallos Ordoñez Stalin Figueroa, Carlos Fiallos. Tiktok and education: Discovering knowledge through learning videos. 2021.

Paul A. Tess. The role of social media in higher education classes (real and virtual) – a literature review. 2013.

Emilio Zagheni Ridhi Kashyap Ingmar Weber Wenqing Qian, Ole Hexel. Demographic inequalities in digital spaces in china: The case of weibo. 2023.

Zhonggen Yu Xinjie Deng. An extended hedonic motivation adoption model of tiktok in higher education. 2023.

Khasanah Lela Susanty Ulfa Maulani Zulvia Trinova, Ali Mu'ammar Zainal Abidin. Online school future: Challenges and expectations of modern education in indonesia. 2022.