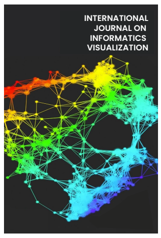
JOIV : Int. J. Inform. Visualization, 8(1) - March 2024 88-95

INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv



Text-Based Content Analysis on Social Media Using Topic Modelling to Support Digital Marketing

Gandhi Surya Buana [a], Raras Tyasnurita [a,*], Nindita Cahya Puspita [a], Retno Aulia Vinarti [a], Faizal Mahananto [a]

[a] *Department of Information Systems, Institut Teknologi Sepuluh Nopember, Jl. Teknik Kimia, Surabaya, 60111, Indonesia*

*Corresponding author: *raras@is.its.ac.id*

***Abstract*— This study aims to create Social Media Analytics (SMA) tools to help Digital Marketers or Content Creators create content topics for creating text-based Instagram content and support digital marketing strategy. Since no SMA tools can provide topic discovery for text-based Instagram content, this research aims to make an SMA tool. The data requirements to make an SMA tool include content text, content caption text, likes, comments, upload time, and content category obtained through the Instascrapper. The method used in this study is the Topic Modelling method using the Latent Dirichlet Allocation (LDA) approach to find the most dominant topic in the content. Optical Character Recognition (OCR) performs an image transformation process to extract text from text-based Instagram content images. The results of SMA tool creation are tested on three expert users, which shows that 93% of test participants could use the SMA to find topic references, and 85% can still be used by users even though they find it difficult. Since the test result shows that SMA tools still need development, for further research, SMA tools can focus on developing the user experience to increase the value of user acceptance by paying attention to the ease of the SMA tools. Also, SMA tools can focus on target users such as Data Analysts, Business Intelligence Analysts, or others within a company to support decision-making for the marketing department.**

***Keywords*— Digital marketing; Instagram; latent dirichlet allocation; optical character recognition; social media analytics; topic modelling.**

Manuscript received 10 Feb. 2022; revised 5 Aug. 2023; accepted 24 Sep. 2023. Date of publication 31 Mar. 2024.

I. Introduction

As a social media platform designed to share photos and video-based content, Instagram has grown in Indonesia, with the number of Instagram users reaching 98.060,000 in August 2021 [1]. Of the many types of Instagram content, text-based content has become a trend for users to spread social information [2], [3]. Text-based content on Instagram appears because users are growing skeptical of traditional marketing methods. Hence, they prefer the text-based content type that can give more value to their lives [4], [5]. It encourages content creators, from businesspeople to influencers, to create text-based content that can spread information and education to its target users [6], [7].

Like the other content, content creators require an ideation process [8] in creating text-based content because the ideation process is trying to choose a crucial relevant topic [9], [10]. In the topic selection process, content creators are trying to search, collect, and create insights by analyzing social media, which is accommodated by a tool or system called social media analytics (SMA) [11], [12].

So far, SMA uses content metadata such as the number of interactions visualized as a dashboard; some SMA tools using website-based are Instagram Insight, SocialSprout, or Brandwatch Benchmark. Besides, several studies on SMA used metadata, captions, and comments for social listening by analyzing the sentiment [13], [14]. However, no SMA tool uses text on the image of the content. The absence of an SMA tool that uses text on the image of content is because the existing SMA is still being developed for content in general; therefore, using text on the image of content is not a top priority [15]. Moreover, before analyzing the text on the image of the content, the text needs to be extracted, which often misrecognizes other forms in the text [16], [17], [18].

In several studies, the extraction process on images containing text has been done to extract memes [19] or text on promotional images [20]. However, the topic of the text on the image has yet to be found. Therefore, it is necessary to do topic modeling to process information to find hidden topics from images that have become text documents based on what has been done in several previous studies on product development planning [21], [22] or on finding a research topic [23].

Optical Character Recognition (OCR) is used to convert documents into ASCII characters (machine-readable characters) [24]. In its application, OCR needs to perform an image transformation process first to allow machines to recognize characters just like humans use their eyes to see an image [25]. The process consists of image input, pre-processing, segmentation [26], normalization, extraction, and recognition [27].

Furthermore, topic modeling using Latent Dirichlet Allocation (LDA) will work as a form of textual data modeling process to identify topics [28] attached to a document or data. Text documents can be in the form of books, journal articles, or various kinds of unstructured text. In the identification process, topics from documents that have been previously extracted using OCR will be found.

The results of this study are expected to make an SMA tool using OCR and Topic Modeling, which can provide relevant topic references to support topic selection in the text-based content ideation process on digital marketing content.

II. Materials and Method

This study applies Optical Character Recognition (OCR) and Topic Modelling to process the discovery of text-based Instagram content on developed Social Media Analytics. It provides insights into the Digital Marketer or Content Creators for supporting digital marketing strategy. Social Media Analytics development will be done with the analysis, design, and integration stages [29]. After SMA is developed, the next step is to implement and integrate SMA. The developed SMA tool is delivered to the users to get opinions and input from users who are digital marketers, social media specialists, and content creators from different companies.

A. Social Media Analytics Basic Needs Analysis

Social Media Analytics has developed into one of the business features needed in various stages of social media [30], [31], one of which is the creation of content ideas [32]. Therefore, it is necessary to analyze the basic needs of various SMA tools to ensure that the SMA developed can meet the unique needs of supporting idea reference discovery and the basic needs of Social Media Analytics. The basic needs analysis is carried out by considering the most popular Social Media Analytics such as Instagram Insight, Sprout Social Instagram, and Brandwatch Benchmark.

1) Instagram Insights: Instagram Insight is an SMA by Instagram to support users in observing the progress of their Instagram accounts [33]. Some of the analysis supported by Instagram can be seen in Table I.

2) SproutSocial Analytics Tools: SproutSocial Analytics Tools is one of the SMA that provides support in marketing analysis across various channels and has been trusted by more than 30,000 world-class companies [34]. Some of the analysis supported by SproutSocial Instagram Analytics Tools can be seen in Table II.

3) Brandwatch Benchmarks: Brandwatch Benchmark is one of the features used to support finding information by comparing one account to another [35]. Some of the analyzes supported by Brandwatch Benchmark can be seen in Table III.

TABLE I
INSTAGRAM INSIGHT ANALYTICS MAPPING

Analysis	Information provided
Reach analysis	• Number of reached accounts. • Reach accounts status (Follower/Non-Follower). • Distribution of the amount of reach for each type of content. • Best reach content for each type of content. • Number of impressions. • Number of profile visits.
Engagement Analysis	• The number of interesting accounts. • Interested account status (Follower/Non-Follower). • Distribution of the amount interactions for each type of interaction and content type.
Followers Analysis	• Number of followers. • Follower growth in a certain period. • Follower distribution for each of the 5 best locations. • Distribution of Followers for each age range. • Distribution of Followers for each gender. • Distribution of Time and Active Days of Followers and Content.

TABLE II
SPROUT SOCIAL INSTAGRAM ANALYSIS MAPPING

Analysis	Information provided
Post analysis	• Best Content Photos. • Engagement of each content. • Captions for each content. • Distribution of time and days of uploaded content.
Audience analysis	• Distribution of the amount of content, likes, and comments uploaded in hours or days.
Competitor analysis	• The distribution of the amount of content, likes and comments within a certain period of time and the development of followers from time to time.

TABLE III
BRAND WATCH BENCHMARK ANALYSIS MAPPING

Analysis	Information provided
Marketing strategy analysis	• Information content. • Engagements for each content. • Spread time upload. • Use of captions.
Performance analysis	• Best content performance. • Use of captions. • Content performance in certain time.

The SMA developed in this research can be used as a Social Media Information Discovery to find topics using Natural

Language Processing [36], [37]. The results of the research mentioned that Instagram Insight could support an in-depth analysis of the target market profile, the SMA analyzed have similarities in some of the information provided but have different contexts and analysis objectives, and the similarity in the information given to each SMA can signify that the information has become the basic information needed by SMA users in making decisions.

Some of the commonalities of information that can be essential and required by SMA users are:

- The best types of content and information are in the form of captions, likes, and comments.
- Content distribution based on a particular period.
- Distribution of content by the hour and day

B. Data Collection Process Design

At this stage, it is necessary to identify related to the data needed so that further identification can also be related to the source and the retrieval and storage process needed to retrieve the required data [38]. Data needs are identified by mapping information needs with the data needed for that information, as shown in Table IV.

TABLE IV
MAPPING OF INFORMATION AND DATA NEEDS

Information Needs	Data Needs
Topics of analyzed content	Content text and content caption
The best type of content	Content categories, likes, comments
Distribution of content within a certain period, Content distribution by the hour and day	Content upload time

Furthermore, identifying data sources is carried out by mapping the data needs with the data source of the data provider, and the mapping results are shown in Table V.

TABLE V
MAPPING OF INFORMATION AND DATA NEEDS

Data Needs	Data source
Content Text	Not available directly because the content needs to be extracted first from the content image in Instagram metadata using the Optical Character Recognition method
Content Caption Text, Likes, Comments, Upload Time	Available directly using Instagram metadata.
Content Category	Not available directly because the content categorization is required using the Topic Modeling method

Finally, the data retrieval and storage process design are carried out by considering the data source, which results in a process seen in Figure 1 that starts from scrapping targets, extracting text, converting data, and storing data in the provided data frame.

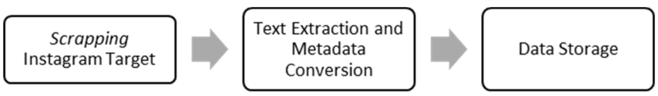


Fig. 1 Process of data retrieval and storage

1. Scrapping Instagram Metadata: Scrapping can be done using the Instaloader Python library [39]. The metadata retrieved focuses on likes, comments, captions, upload times, and content images. The results are temporarily stored in the form of files in JSON format for metadata likes, comments, captions, and upload times, while in jpg format for any existing content.

2. Extracting image to text: After the image has been successfully retrieved, the text can be extracted using OCR for each available image in each content. The available JSON file and the extraction process will then be converted to CSV. The results of this stage will be stored temporarily as variables in the data frame format with the names in Table VI.

TABLE VI
TEXT EXTRACTION RESULT DATA

Data	Variable name
Text from image 1 to 10	text1, text2, text3, text4, … text10
Caption	Caption
Likes	Likes
Comments	Comments
Upload date	date

3. Final data storage: The text from the caption and images will be combined because they have the same context. This process aims to reduce storage memory and retrieve data more effectively. Table VII shows the variable name of the final data storage.

TABLE VII
FINAL DATA COLLECTION RESULT

Data	Variable name
Content text	Finals
Likes	Likes
Comments	Comments
Upload date	Date

C. Data Pre-Process Process Design

The results of the data that have been collected then need to be prepared in advance in the pre-processing data stage. Therefore, at this stage, the pre-processing of the data will go through two stages as follows.

1. Data Cleaning: Data cleaning is done first by using the Tokenizing method to separate one word from another, Filtering to select tokens that are not important, and stemming from converting the word into its basic word until finally, the data eraser that has no text in it is done [40]. This process will be carried out with the support of English and Indonesian libraries and the addition of non-standard words often used in some instances.

2. Data Preparation: Data preparation was done to create a corpus and library/dictionary list for later use in the topic modeling process [41].

D. Data Processing Design

The results of the pre-processed data will then be processed using topic modeling, which needs to determine the ideal topic first. Determination of the perfect issue is done by comparing the number of coherence values [42] for each selected topic [43]. From the results of the coherence value, the user can determine the number of topics he wants by considering the coherence value or the number of issues according to his preferences [42], [44]. Then, each piece of content will be labeled according to the most dominant topic.

E. Type Design and Visualization Process

The visualization used in SMA is designed by considering the basic needs of SMA and the results of the topic modeling to suit user needs. The results of the visualization mapping table with user needs are shown in Table VIII.

TABLE VIII
TYPES OF VISUALIZATION

Information Needs	Visualization
The best type of content	Bar Chart
Distribution of content within a certain period	Scatter Plot
Content distribution by the hour and day	Bar Chart
Topics of analyzed content	Wordcloud and Barchart

F. Integration Process

The SMA process is integrated by combining several existing main processes into a class containing functions to be used separately but still in a good unit. In addition, the integration process is carried out by understanding the interactions that users can carry out. This integration process is carried out to produce optimal topics using the InstaScrapper, Cleaning, and TopicModelling class. InstaScrapper class is used to collect text data from Instagram so that it can be used at the Topic Modeling stage or different stages, the cleaning class is used to carry out the pre-processing data stage, especially at the cleaning stage, and the TopicModelling class is used to carry out the topic modeling process from preparing the corpus to the dictionary used. In addition, this class is also used to carry out the visualization process of a predetermined topic so that, in the end, it produces knowledge that users can use. The design of the integration process for each class can be seen in Figure 2.

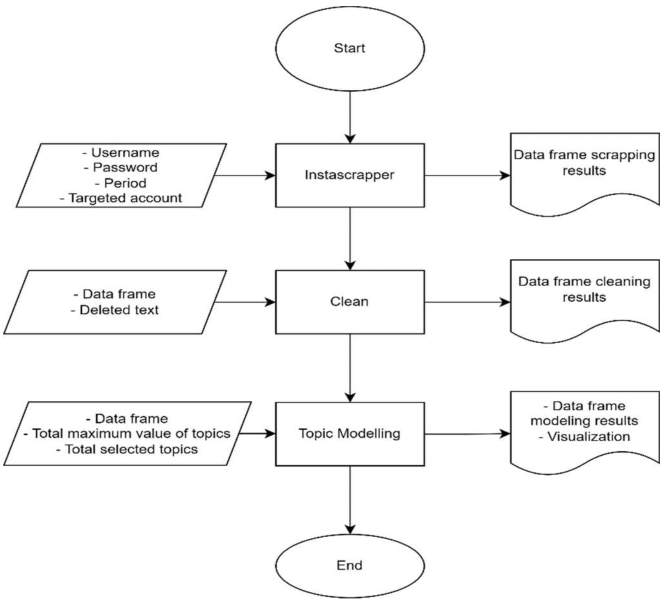


Fig. 2 Stages of integration process

G. Expert Evaluation

Expert evaluation is conducted through testing to assess the understanding and acceptance of the developed SMA tool. This process is not intended as a sample test. Instead, it serves as supporting data to determine whether the developed model of the SMA tool can be comprehended and accepted by expert users. In this context, expert users are the participants in the expert evaluation. The expert user's profile is a digital marketer, social media specialist, and content creator with a stake in developing digital marketing strategies.

III. RESULTS AND DISCUSSION

The results from developing and implementing the Social Media Analytics tool are presented in this subsection.

A. OCR Implementation

In the data collection process, the InstaScrapper class is defined as scrapper = InstaScrapper(), so it can call functions inside InstaScrapper to collect Instagram content data. Data retrieval was successfully carried out by targeting the dibimbing.id [45] account from 2023-01-05 to 2023-04-06 is stored in a data frame. The data downloaded in data collection is in the form of .jpg format containing Instagram content images and. json.xz format containing content metadata. At the same time, the program runs to perform text extraction for each downloaded image.

Text extraction for each image is done using OCR. This extraction is essential because the results of this extraction will then be processed again in the cleaning process and data processing using topic modelling. Therefore, to ensure the extraction of text produced by OCR implemented in the SMA tool, it is necessary to pay attention to the performance of OCR extraction. The performance of OCR extraction can be assessed by comparing the precision of the word results from extracting content images into text against the actual words in text-based Instagram content.

Fig. 3 Piece of text based Instagram content

Figure 3 shows a piece of Instagram text-based content in the Instagram account of dibimbing.id. From the original content in Figure 4, there are ten words in it, and those are “dibimbing; #YourCareerPartner; Arief; Muhammad; Resmi; Ditunjuk; Sebagai; Duta; Nasi; Padang”. Meanwhile, the OCR extraction results only managed to get eight words that matched the original content, and those are “Aibimbing; gp Yourcareer; Arief; Muhammad; Resmi; Ditunjuk; Sebagai; Duta; Nasi; Padang”. Therefore, the performance value obtained in this piece of content is 80%. Therefore, the extraction results will be compared to all images in the Instagram content to assess the text extraction performance. OCR extraction performance results are described in Table IX.

TABLE IX
OCR EXTRACTION PERFORMANCE

Content Image	Total Correct Words Original Content	Total Correct Words Extraction Result	Performance
Content Image 1	10	8	80%
Content Image 2	96	93	96.8%
Content Image 3	40	38	95%
Content Image 4	31	28	90.32%
Content Image 5	33	23	69.70%
Content Image 6	45	20	44.44%
Total Correct Word	255	210	82.3%

Table IX shows the OCR performance result is quite good, with a value of 82.3%, which successfully extracted 210 words from a total of 255 words. Some words that cannot be identified well are brand logos like "dibimbing.id", graphic writing from photos, and contrast differences in the content images.

B. *Topic Modelling Implementation*

The implementation of Topic Modelling involves the process of data collection, data cleaning, and topic modeling. It provides topic content recommendations, insights on content performance, and content upload strategies from the analyzed Instagram account. The data collection process is performed using the InstaScrapper() function, which retrieves data such as images, metadata, and text extraction from the content. The extracted text is cleaned using the Clean () function.

Fig. 4 Word cloud result for automatic cleaning

Data cleaning in the SMA tool is carried out twice: first automatically and then manually by the end-user to remove irrelevant words. The extracted and cleaned text is visualized in a word cloud. Figure 4 shows the results of automatic data cleaning by the SMA tool. However, manual cleaning is conducted to generate essential words, as shown in Figure 5, since there are still some irrelevant words like dan, none, di, gp, swip, yang, etc.

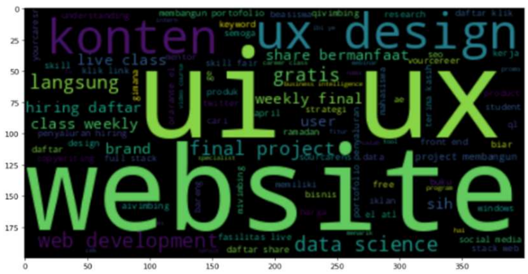


Fig. 5 Word cloud result for manual cleaning

The cleaned data is used to process the Topic Modeling by invoking the TopicModelling() function. The Topic Modeling process provides content recommendations based on the highest coherence score from the topic.viz() function. Figure 6 shows the results of the coherence score, and the highest coherence score is 4; therefore, the optimum number for generating topic recommendations is 4 topics.

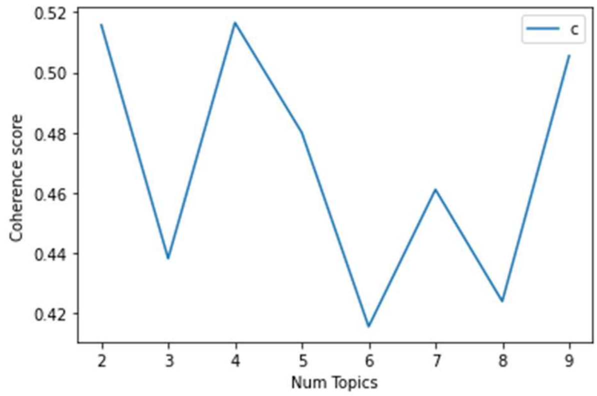


Fig. 6 Coherence Score Visualization

The content recommendations display word clouds and related keywords for each content, as shown in Figure 7.

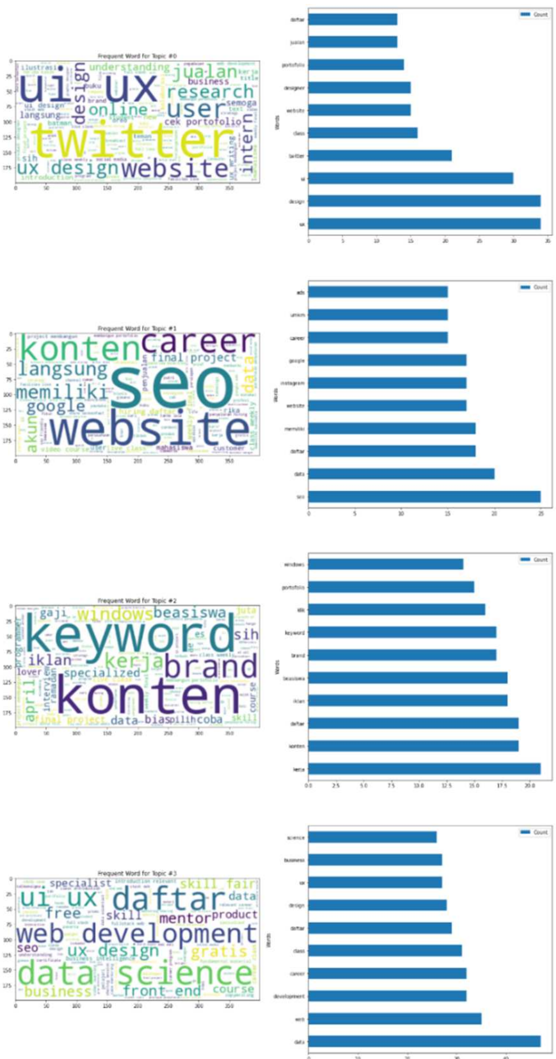


Fig. 7 Topic recommendations and its keywords

Figure 7 shows the four topic recommendations generated from the topic modelling exhibit varying dominance of keywords. The dominant keywords for Topic 0 revolve around UI/UX design and website-related matters. Topic 1 focuses on SEO, website data, and Google-related topics. Topic 2 centers on job listings and portfolio-related discussions. Lastly, Topic 3 encompasses career-related matters, classes, and skills such as web development, data science, business, etc.

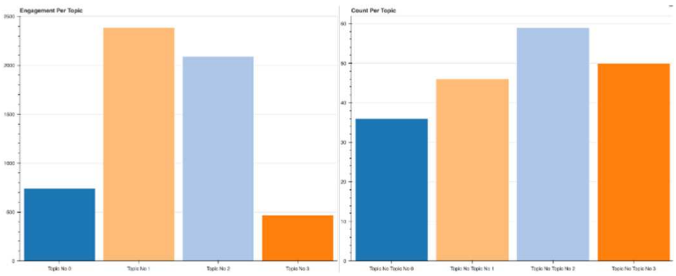


Fig. 8 Content Performance: engagement and total content

Besides, the analysis yields valuable insights into content performance, offering a comprehensive understanding of engagement levels, as shown in Figure 8. Content performance shows the results of topic engagement on the left-hand graph. The number of comments and likes influences the topic engagement on Instagram content, indicating the virality of the topic. On the other hand, the number of content pieces per topic merely shows the quantity of uploaded content.

The engagement and total content do not directly impact each other. It is important to note that despite having the highest engagement, it may have less uploaded content. However, both insights are equally crucial in considering the urgency of creating Instagram content, whether it aims for virality or aims to address commonly discussed topics.

Therefore, it's crucial to consider the content trends, as illustrated in the graph in Figure 9. The graph represents the distribution of content over time. Each dot on the graph represents a piece of content, where the x-axis indicates the specific time range during which various topics are extensively discussed, and the y-axis represents the corresponding engagement levels of each piece of content.

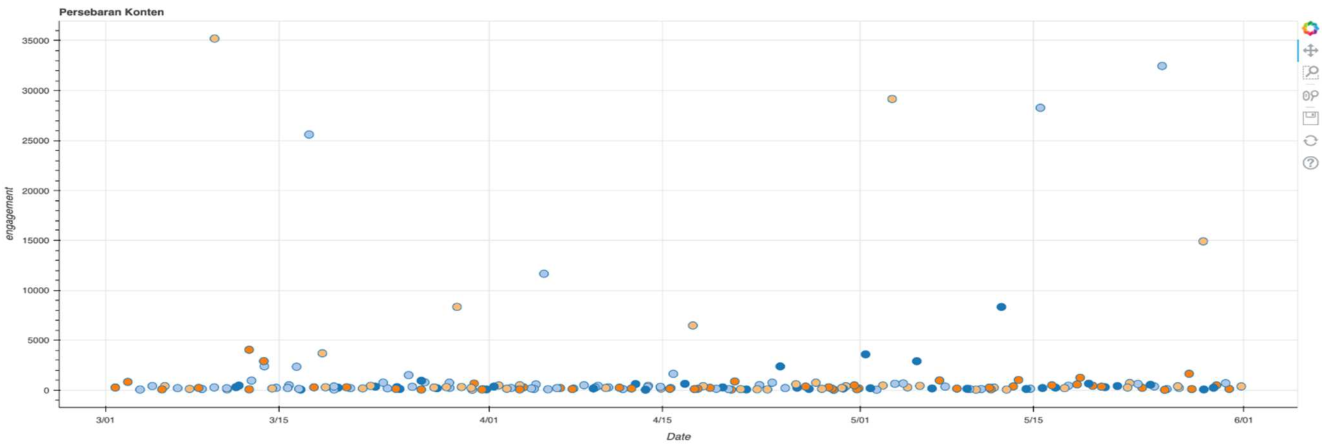


Fig. 9 Content Performance: Content trend by its distribution

The last result for Topic Modelling is visualizing insight for the Upload Content Strategy. This result visualizes the amount of content uploaded based on the day and time of upload, which can be seen in Figure 6. The graph on the left side indicates the best days for content upload, while the chart on the right shows the optimal hours for content upload. Based on the results, topics with the highest engagement are predominantly uploaded on Wednesdays, Thursdays, and Saturdays, specifically from 17:00 WIB to 18:00 WIB. This result suggests that uploading content during those time frames holds the potential for achieving high reach and engagement.

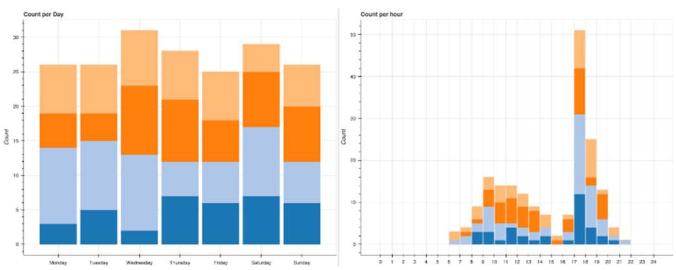


Fig. 10 Visualization of the amount of content by day and time

However, it is essential to note that the interpretation of data from each visualization in Topic Modeling may vary for different users. Digital marketers, social media specialists, and content creators have their way of interpreting data and determining strategies to implement in their digital marketing plans. Due to the dynamic nature of social media, the trends are rapidly changing, and the selection of topics is tailored individually to user needs and requirements.

C. *Expert Evaluation Result*

Expert evaluation is conducted by testing expert users through two tests to measure their understanding and acceptance. Also, the purpose of this evaluation is to gather user opinions regarding the SMA tool that has been developed. User feedback is valuable as it serves as input for future improvements. The first test is for measuring user understanding and usage to determine the performance of SMA in providing topic references. The results can be seen in Table X, which shows that of the 15 commands and questions given, the average participant managed to answer 14 (93%). This result indicated that the test participants could use the SMA to find topic references.

TABLE X
TEST RESULTS FOR USER UNDERSTANDING AND USAGE

No	Questions/Assignments		Correct
1.	What is the purpose of each currently visible visualization?	Bar Chart Engagement Per Topic	3 testers
2.		Bar Chart Total-Content Per Topic	3 testers
3.		Scatter Plot of Content Performance Distribution in one period	3 testers
4.		Stacked Bar chart Per Topic Content/Day	2 testers
5.		Stacked Bar chart Per Topic Content/Hour	3 testers
6.		Word cloud and Bar chart for each Topic	3 testers
7.	What information can be inferred from this visualization?	Topics with the best engagements	3 testers
8.		Topics with the highest number of uploads	3 testers
9.		Content performance trends	3 testers
10.		The best upload days for the best topics	2 testers
11.		Best upload day for the best time	3 testers
12.		Summarizing the best topic content	3 testers
13.		Summing up the contents of the worst topics	3 testers
14.	From the information in each visualization, try to explain the overall conclusion you get!		3 testers
15.	From this conclusion, try to explain the steps you might take after!		3 testers

The second test was carried out using the Systems Usability Scare (SUS) method with a scale of 1-5 for users with negative questions having a formula (5-n) and positive questions having a formula (n-1) to get a maximum score of 4 for each question and 40 for the maximum total score. The results of the test for each statement can be seen in Table XI.

TABLE XI
TEST RESULTS FOR USER ACCEPTANCE USING SUS

#	Statement	Score
1.	Users think to use this SMA as often as possible	3.00
2.	Users think SMA is very complicated	3.60
3.	The user thinks it will need someone's help to use this SMA	3.00
4.	Users think that SMA is easy to use	4.00
5.	Users find the functions well integrated	3.60
6.	Users think too much information is inconsistent in SMA	4.00
7.	Users think many people will be able to learn to use this SMA quickly	3.33
8.	Users think SMA is hard to use	3.33
9.	Users feel confident using SMA	3.00
10.	Users need to learn many things before being able to use this system	3.00

From the average results of each statement, the lowest value (3.00) comes from statements 1, 3, 9, and 10. Users explain that they feel insecure because the developed SMA is in the form of a Python program. Since the expert users are not IT people, they need help to operate the SMA especially if they have some problems. Meanwhile, the total average value for the User Acceptance test is 85%, which signifies that users can still use SMA even though they find it difficult. The results from the User Understanding and Usage and User Acceptance give insights into the performance of the SMA tool that has been developed.

IV. CONCLUSIONS

The SMA tool has been successfully developed using Python, using three essential classes: InstaScrapper, Clean, and TopicModelling. The tool's success is measured by the OCR performance value of 82.3% and the implementation of Topic Modelling, which generates topic recommendations. Additionally, it provides valuable insights into content performance analysis and strategies for content uploading. The model of the SMA tool was tested on the expert users to get their opinions and feedback regarding the SMA through Expert Evaluation. The result of the Expert Evaluation shows that 93% of participants were able to accurately identify relevant topics and leverage the SMA tool's capabilities in generating topic references and facilitating ideation for text-based Instagram content. Moreover, the acceptance test score from the participants is 85%. Notably, the lowest score was attributed to participants who needed a programming background, necessitating occasional assistance from others. The results indicated that the SMA tool needs user proficiency in programming, and the users need user-friendly interfaces and guidelines to maximize the SMA tool's effectiveness. Some suggestions for further development are that future research can focus on developing the user experience on a web application to increase the value of user acceptance by paying attention to the ease of the SMA process. The development can focus on targeted users such as Data Analysts, Business Intelligence Analysts, or other roles within a company to support decision-making from the marketing department.

ACKNOWLEDGMENT

This research was funded by Department of Information Systems, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia (1782/PKS/ITS/2022).

REFERENCES

[1] Napoleoncat, “Instagram users in Indonesia,” *https://napoleoncat.com/stats/instagram-users-in-indonesia/2021/08/*, Aug. 2021.

[2] Instagram, “What’s trending – info-social: Creativity of the Movement,” *https://business.instagram.com/blog/trends-info-social-posts*, Dec. 21, 2020.

[3] Y. Li and Y. Xie, “Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement,” Journal of Marketing Research, vol. 57, no. 1, pp. 1–19, Nov. 2019, doi:10.1177/0022243719881113.

[4] R. Rietveld, W. Van Dolen, M. Mazloom, and M. Worring, “What you Feel, Is what you like Influence of Message Appeals on Customer Engagement on Instagram,” Journal of Interactive Marketing, vol. 49, no. 1, pp. 20–53, Feb. 2020, doi: 10.1016/j.intmar.2019.06.003.

[5] A. M. Rafi, S. Rana, R. Kaur, Q. M. J. Wu, and P. Moradian Zadeh, “Understanding Global Reaction to the Recent Outbreaks of COVID-19: Insights from Instagram Data Analysis,” 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct. 2020, doi:10.1109/smc42975.2020.9283376.

[6] L. D. Hollebeek and K. Macky, “Digital Content Marketing’s Role in Fostering Consumer Engagement, Trust, and Value: Framework,

Fundamental Propositions, and Implications," Journal of Interactive Marketing, vol. 45, pp. 27–41, Feb. 2019, doi:10.1016/j.intmar.2018.07.003.

[7] Y. A. Argyris, Z. Wang, Y. Kim, and Z. Yin, "The effects of visual congruence on increasing consumers' brand engagement: An empirical investigation of influencer marketing on instagram using deep-learning algorithms for automatic image classification," Computers in Human Behavior, vol. 112, p. 106443, Nov. 2020, doi:10.1016/j.chb.2020.106443.

[8] S. Acar, M. Neumayer, and C. Burnett, "Social Media Use and Creativity: Exploring the Influences on Ideational Behavior and Creative Activity," The Journal of Creative Behavior, vol. 55, no. 1, pp. 39–52, Nov. 2019, doi: 10.1002/jocb.432.

[9] S. Ozcan, M. Suloglu, C. O. Sakar, and S. Chatufale, "Social media mining for ideation: Identification of sustainable solutions and opinions," Technovation, vol. 107, p. 102322, Sep. 2021, doi:10.1016/j.technovation.2021.102322.

[10] D. Chaffey, "Recommended tools and techniques to support content ideation," *https://www.davechaffey.com/digital-marketing-glossary/content-ideation/*, 2020.

[11] S. S. Madila, M. A. Dida, and S. Kaijage, "Tourism SME's Adoption of Social Media Analytics Tools and Technology," Vol 11 (1)2022, no. 11(1)2022, pp. 239–247, Feb. 2022, doi: 10.46222/ajhtl.19770720.223.

[12] Y. Wang, M. Rod, Q. Deng, and S. Ji, "Exploiting business networks in the age of social media: the use and integration of social media analytics in B2B marketing," Journal of Business & Industrial Marketing, vol. 36, no. 12, pp. 2139–2149, Mar. 2020, doi: 10.1108/jbim-05-2019-0173.

[13] P. Ducange, M. Fazzolari, M. Petrocchi, and M. Vecchio, "An effective Decision Support System for social media listening based on cross-source sentiment analysis models," Engineering Applications of Artificial Intelligence, vol. 78, pp. 71–85, Feb. 2019, doi:10.1016/j.engappai.2018.10.014.

[14] M. T. Ballestar, M. Cuerdo-Mir, and M. T. Freire-Rubio, "The Concept of Sustainability on Social Media: A Social Listening Approach," Sustainability, vol. 12, no. 5, p. 2122, Mar. 2020, doi:10.3390/su12052122.

[15] W. Pearce et al., "Visual cross-platform analysis: digital methods to research social media images," Information, Communication & Society, vol. 23, no. 2, pp. 161–180, Jun. 2018, doi: 10.1080/1369118x.2018.1486871.

[16] D. Akinbade, A. O. Ogunde, M. O. Odim, and B. O. Oguntunde, "An Adaptive Thresholding Algorithm-Based Optical Character Recognition System for Information Extraction in Complex Images," Journal of Computer Science, vol. 16, no. 6, pp. 784–801, Jun. 2020, doi:10.3844/jcssp.2020.784.801.

[17] J. Park, E. Lee, Y. Kim, I. Kang, H. I. Koo, and N. I. Cho, "Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter," IEEE Access, vol. 8, pp. 174437–174448, 2020, doi: 10.1109/access.2020.3025769.

[18] W. Suteddy, D. A. R. Agustini, A. Adiwilaga, and D. A. Atmanto, "End-To-End Evaluation of Deep Learning Architectures for Off-Line Handwriting Writer Identification: A Comparative Study," JOIV : International Journal on Informatics Visualization, vol. 7, no. 1, p. 178, Feb. 2023, doi: 10.30630/joiv.7.1.1293.

[19] Á. MacDermott, M. Motylinski, F. Iqbal, K. Stamp, M. Hussain, and A. Marrington, "Using deep learning to detect social media 'trolls,'" Forensic Science International: Digital Investigation, vol. 43, p. 301446, Sep. 2022, doi: 10.1016/j.fsidi.2022.301446.

[20] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," Procedia Computer Science, vol. 179, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[21] B. Jeong, J. Yoon, and J.-M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," International Journal of Information Management, vol. 48, pp. 280–290, Oct. 2019, doi:10.1016/j.ijinfomgt.2017.09.009.

[22] N. Hu, T. Zhang, B. Gao, and I. Bose, "What do hotel customers complain about? Text analysis using structural topic model," Tourism Management, vol. 72, pp. 417–426, Jun. 2019, doi:10.1016/j.tourman.2019.01.002.

[23] M. Reisenbichler and T. Reutterer, "Topic modeling in marketing: recent advances and research opportunities," Journal of Business Economics, vol. 89, no. 3, pp. 327–356, Sep. 2018, doi: 10.1007/s11573-018-0915-7.

[24] C. Thorat, A. Bhat, P. Sawant, I. Bartakke, and S. Shirsath, "A Detailed Review on Text Extraction Using Optical Character Recognition," Lecture Notes in Networks and Systems, pp. 719–728, 2022, doi:10.1007/978-981-16-5655-2_69.

[25] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," IEEE Access, vol. 8, pp. 142642–142668, 2020, doi:10.1109/access.2020.3012542.

[26] M. Zulqarnain, R. Ghazali, M. G. Ghouse, and M. F. Mushtaq, "Efficient processing of GRU based on word embedding for text classification," JOIV : International Journal on Informatics Visualization, vol. 3, no. 4, pp. 377–383, Nov. 2019, doi: 10.30630/joiv.3.4.289.

[27] U. Salimah, V. Maharani, and R. Nursyanti, "Automatic License Plate Recognition Using Optical Character Recognition," IOP Conference Series: Materials Science and Engineering, vol. 1115, no. 1, p. 012023, Mar. 2021, doi: 10.1088/1757-899x/1115/1/012023.

[28] J. R. Saura, D. Palacios-Marqués, and D. Ribeiro-Soriano, "Exploring the boundaries of open innovation: Evidence from social media mining," Technovation, vol. 119, p. 102447, Jan. 2023, doi:10.1016/j.technovation.2021.102447.

[29] A. Geissinger, C. Laurell, C. Öberg, and C. Sandström, "Social media analytics for innovation management research: A systematic literature review and future research agenda," Technovation, vol. 123, p. 102712, May 2023, doi: 10.1016/j.technovation.2023.102712.

[30] K. R. Fitzpatrick and P. L. Weissman, "Public relations in the age of data: corporate perspectives on social media analytics (SMA)," Journal of Communication Management, vol. 25, no. 4, pp. 401–416, Jun. 2021, doi: 10.1108/jcom-09-2020-0092.

[31] J. Choi, J. Yoon, J. Chung, B.-Y. Coh, and J.-M. Lee, "Social media analytics and business intelligence research: A systematic review," Information Processing & Management, vol. 57, no. 6, p. 102279, Nov. 2020, doi: 10.1016/j.ipm.2020.102279.

[32] H. Shahbaznezhad, R. Dolan, and M. Rashidirad, "The Role of Social Media Content Format and Platform in Users' Engagement Behavior," Journal of Interactive Marketing, vol. 53, pp. 47–65, Feb. 2021, doi:10.1016/j.intmar.2020.05.001.

[33] Instagram, "View Account Insights on Instagram," *https://help.instagram.com/1533933820244654*, 2023.

[34] Sproutsocial, "Social Media Analytics," *https://sproutsocial.com/features/social-media-analytics/*, 2023.

[35] Brandwatch, "Brandwatch," *https://www.brandwatch.com/*, 2023.

[36] U. Ruhi, "Social Media Analytics as a Business Intelligence Practice:Current Landscape & Future Prospects," Journal of Internet Social Networking & Virtual Communities, pp. 1–12, Jul. 2014, doi: 10.5171/2014.920553.

[37] Z. Mottaghinia, M.-R. Feizi-Derakhshi, L. Farzinvash, and P. Salehpour, "A review of approaches for topic detection in Twitter," Journal of Experimental & Theoretical Artificial Intelligence, vol. 33, no. 5, pp. 747–773, Jun. 2020, doi: 10.1080/0952813x.2020.1785019.

[38] A. R. Rahmanti, D. N. A. Ningrum, L. Lazuardi, H.-C. Yang, and Y.-C. Li, "Social Media Data Analytics for Outbreak Risk Communication: Public Attention on the 'New Normal' During the COVID-19 Pandemic in Indonesia," Computer Methods and Programs in Biomedicine, vol. 205, p. 106083, Jun. 2021, doi: 10.1016/j.cmpb.2021.106083.

[39] D. Amanatidis, I. Mylona, I. (Eirini) Kamenidou, S. Mamalis, and A. Stavrianea, "Mining Textual and Imagery Instagram Data during the COVID-19 Pandemic," Applied Sciences, vol. 11, no. 9, p. 4281, May 2021, doi: 10.3390/app11094281.

[40] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi," IOP Conference Series: Materials Science and Engineering, vol. 874, no. 1, p. 012017, Jun. 2020, doi: 10.1088/1757-899x/874/1/012017.

[41] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), Oct. 2019, doi: 10.1109/icecos47637.2019.8984523.

[42] S. J. Blair, Y. Bi, and M. D. Mulvenna, "Aggregated topic models for increasing social media topic coherence," Applied Intelligence, vol. 50, no. 1, pp. 138–156, Jul. 2019, doi: 10.1007/s10489-019-01438-z.

[43] J. A. Wahid et al., "Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response," Expert Systems with Applications, vol. 195, p. 116562, Jun. 2022, doi: 10.1016/j.eswa.2022.116562.

[44] K. Porter, "Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling," Digital Investigation, vol. 26, pp. S87–S97, Jul. 2018, doi: 10.1016/j.diin.2018.04.023.

[45] Instagram, "dibimbing.id," *https://www.instagram.com/dibimbing.id/*, 2023.