Analysis Of University Social Media User Engagement By Topic

Muhammad Aziz Pratama*1, Hasmawati2, Ade Romadhony3
1,2,3Informatics, Faculty of Informatics, Telkom University, Jl. Telekomunikasi No. 1 Terusan Buah Batu, Bandung, Indonesia
e-mail: *azizmap@student.telkomuniversity.ac.id, 2hasmawati@telkomuniversity.ac.id, 3aderomadhony@telkomuniversity.ac.id

Abstract
Social media is currently one of the primary platforms for spreading information about organizations or companies, including universities. To reach more users, we need to analyze user engagement. Topic is one of essential features to attract user interest on a post. Therefore, by understanding the topic preference, we can increase the user engagement. In this study, we identify topic posts using Latent Dirichlet Algorithm (LDA) method. Furthermore, we predict the engagement scores (number of likes and comments) using decision tree and random forest method. The experimental results show that there are 11 topics identified by LDA, and after analyzing further at three different engagement levels (views, likes, and comments), the most engaging topic is achievement. The experimental results on engagement prediction reveal that random forest performs better than decision tree, with the MAE score at 1074.

Keywords—user engagement, topic analysis, decision tree, random forest

1. INTRODUCTION
Social media is one of the essential aspects of daily life [1]. In addition to having their website, it is common for universities and colleges to have social media accounts to increase their popularity and share relevant information [2]. Some universities in Indonesia already have official social media accounts such as Facebook, Twitter, Instagram, and others to share the latest information and activities at the university.

Typically, social media users communicate by exchanging messages. Users can interact with posts on social media platforms like Instagram by leaving likes or comments. User engagement reflects the experience quality characterized by actor’s cognitive, temporal, emotional, and behavioral investment when interacting with digital systems [3]. User engagement, such as user growth (followers) and interactions (e.g., likes, comments, shares, etc.), are crucial indicators of an organization's platform success [4]. The bigger the interaction indicator on a social media post or account, the more popular the account and may get more attention from publics who interested to follow the account. Universities that have official social media accounts can consider the user engagement aspect to determine which types of posts are most attractive to public. Users tend to have different topic preferences when consuming information on social media [5]. Several research has used topic features to predict user preferences, such as product recommendations [6]. Topics are also helpful for understanding users’ preferences who aim to get higher engagement on Twitter [7]. It’s essential to understand which topics to post to increase user engagement on social media platforms.

This study examines social media user engagement on university Instagram social media posts based on the post topic. Each engagement metric has its own effects, so different metrics...
will generate unique types of engagement and user expressions [4]. On analyzing user engagement, we observed the engagement metric based on the number views, likes, and comments for each topic on the university's Instagram posts. Meanwhile to identify topic post, we use Latent Dirichlet Allocation (LDA), which is a statistical model that has been widely used in text analysis and to identify topic [8].

The post topic identification results are utilized as features in predicting user engagement. In this work, we use Decision Tree and Random Forest algorithms to predict the user engagement level. Previous research has demonstrated that Decision Trees and Random Forest are effective classification techniques with positive outcomes [4]. In this study, Decision Tree and Random Forest are used to predict the number of likes and comments on an Instagram post from a university by using the post's topic, textual feature, and sentiment scores obtained from the sum of each word's scores in a dictionary of sentiment words, and TF-IDF feature extraction from the post's caption.

2. METHODOLOGY

The data used in this research are Instagram posts made by Telkom University from 10 January 2021 until 31 December 2021. The total number of Instagram posts was 251. LDA is used to identify topics in university Instagram posts using the caption feature. The chi-square test was performed to determine if the difference between the number of posts with high and low engagement was statistically significant at three levels for each topic. Decision Tree and Random Forest were used to predict the engagement score (sum of likes and comments).

2.1. Topic Identification on University Instagram Account Posts

In this research, each post topic was manually labeled based on the Instagram caption of each university post, which can be categorized in general as follows:

- **Prestasi**: achievement of students, lecturers, or the university itself;
- **Trivia**: additional information about the university;
- **Peringatan**: national day or condolences;
- **Event**: activities or events organized by the university;
- **PMB**: new student admission;
- **Wisuda**: information around graduation;
- **Akademik**: information around academics.

Topics that have been manually labeled are used to evaluate the results of topic identification by LDA, where LDA is used to identify topics based on university Instagram captions automatically. Before being given to LDA, the Instagram caption data went through pre-processing steps as follows:

- Case folding;
- remove e-mail, URL, and punctuation;
- remove stop words;
- tokenize sentences into words and perform stemming (returning terms to their base form).

2.2. Engagement Analysis of University Instagram Posts by Topic

In this research, the topic analysis was conducted at three levels. According to previous research conducted by Aldous et al. [4], the levels of engagement were arranged from most private to the most public based on their expression. Table 1 displays the analyzed engagement levels.
At each engagement level, post data is sorted by the number of engagement metrics and divided into three equal parts of 33% of the overall data. The top 33% of the data was considered high engagement posts, and the bottom 33% of the data was regarded as low engagement posts. The middle 33% of the data was ignored to ensure the differences in the metrics were comparable. The chi-square test was used to determine if there was a statistically significant difference between the number of posts with high and low engagement on each topic across all three levels.

2.3. Engagement Score Prediction System

The Decision Tree and Random Forest methods are used to predict the engagement scores for university Instagram posts, which are the sum of the number of likes and comments. The overall system can be seen in Figure 1.

![Figure 1. System Flowchart](image-url)
2.3.1. Dataset
The data features used in the engagement score prediction systems are obtained from Telkom University’s official Instagram account. Data features are grouped into two categories:
- **Language**: the result of TF-IDF method feature extraction on Instagram caption;
- **Metadata**:
  - **Topic**: the topic of post by LDA;
  - **Textual**: the number of characters of the post caption;
  - **Sentiment**: the sum of the sentiment scores of all words in a caption obtained from the InSet sentiment words dictionary [9].

2.3.2. Pre-processing
The pre-processing of caption data uses the same step and method as pre-processing for LDA modeling.

2.3.3. Feature Extractions
Because machine learning models don’t accept strings, Instagram caption-text must be extracted first. At this stage, the TF-IDF method is used to weight each word in the caption text. TF-IDF can eliminate the most common terms and retrieve more relevant terms in a document corpus [10].

2.3.4. Data Splitting
Separate, non-overlapping datasets are required for the training and testing phases of the model. The dataset is separated into train set and test set with ratio of 8:2, where the train set data consists of 80% of the total data used for training the model, while the test set data is 20% of the total data used in the model testing stage.

2.3.5. Model Training
The methods used for predicting engagement scores are Decision Tree and Random Forest. Decision Tree has been widely used in several disciplines because of its easy use and can be interpreted easily [11]. A random forest consists of a collection of regression Decision Tree, where each Decision Tree forms a unit and then selects the most popular target [12]. Each model is trained with three data sets: language, metadata, and language + metadata. The language feature is a caption text that has been represented as a vector with TF-IDF and metadata containing topic, textual, and sentiment score features. Therefore, each method generates three models.

2.3.6. Model Testing
After the model has gone through the training phase, the testing phase is carried out with test data that has never been seen by the model before. The test data has gone through the same pre-processing phase as the train data. Each model that has been trained predicts the number of engagement scores for each test set data.

2.3.7. System Evaluation
MAE (Mean Absolute Error) calculation is used to evaluate the error rate of the engagement score prediction results that have been done by the models towards each test data. MAE is a method that calculates the error by giving equal weight to all data [13].

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|
\]  

(1)

Muhammad Aziz, et., al [Analysis of University Social Media User Engagement by Topic]
Where $y_i$ is the actual value, $\hat{y}_i$ is the predicted value of $y_i$, and $n$ is the number of data. The smaller the MAE of the prediction made by the model, the better the model's performance.

3. RESULT AND ANALYSIS

3.1. Summary of Data

The distribution of topics in the university Instagram post data that has been manually labeled is shown in Figure 2. It can be seen that the prestasi topic is the topic with the most amount in the dataset, followed by trivia, peringatan, and event.

![Figure 2. Manual Label Topic Distribution](image)

Figure 3 shows the mean and median values of views, likes, and comments for all post data. The number of views is only found on video posts, and video posts are only about 10% of the total Instagram upload data. It can be seen that akademik topic has the highest mean and median value on views, followed by trivia and wisuda. Akademik topic has the highest mean and median like followed by trivia and akademik topic is also has the highest mean and median comment followed by PMB topic.
3.2. **Topic and Engagement Analysis**

The result of the LDA model is the coherence value, where the coherence value is used to determine the correct number of topics for LDA based on existing data. According to the Figure 4, 11 is the number of topics with the highest coherence value of around 0.42.

![Figure 3. Mean and Median Views, Like, Comment across Topics](image)

![Figure 4. Coherence value for the number of topics LDA](image)

Using the LDA topic results, each university's Instagram post data is labeled based on the highest topic percentage. Figure 5 shows the distribution of 11 topics from LDA results on each
university's Instagram posts. It is known that topic 8 is the topic with the highest number of posts in the dataset.

![Topic Distribution](image)

Figure 5. Distribution of LDA topic results on the dataset

Table 2 shows the LDA topic results are analyzed based on the relationship of keywords and compared with topics that have been manually labeled on the dataset. Prestasi (achievement) is the most dominant topic from the LDA topic results. Topic 1 discusses trivial information related to covid and achievement. Topics 2, 4, 5, 6, 8, 9, and 10 mostly discuss the achievements and innovations that have been achieved by students, lecturers, and the university itself. Topic 3 discusses activities and condolences. Topic 7 discusses graduation activities and topic 11 discusses national day. According to the comparison between LDA and manual topic labels, only 154 (or 61 percent) data match based on the topic group.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords Analysis Topic Based On Manual Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mandiri, isolasi, tingkat, covid, pasien, rumah, tubuh, choir, aplikasi, gejala</td>
</tr>
<tr>
<td>2</td>
<td>selamat, bangsa, telkom, moga, alumni, sukses, kuliah, kerja, teknologi, didik</td>
</tr>
<tr>
<td>3</td>
<td>lomba, covid, alumni, mahasiswa, vaksinasi, keluarga, program, daftar, link, duka</td>
</tr>
<tr>
<td>4</td>
<td>fakultas, selamat, program, moga, raih, ucap, ilmu, terap, informatika, kucing</td>
</tr>
<tr>
<td>5</td>
<td>program, studi, unggul, akreditasi, teknik, selamat, indonesia, kasih, banpt, mahasiswa</td>
</tr>
<tr>
<td>6</td>
<td>indonesia, robot, kontes, raih, moga, tim, giat, prof, selamat, inovasi</td>
</tr>
<tr>
<td>7</td>
<td>wisuda, jiang, informasi, masker, daftar, periode, internasional, laksana, ayah, semangat</td>
</tr>
<tr>
<td>8</td>
<td>peringkat, indonesia, hasil, guru, university, baik,</td>
</tr>
</tbody>
</table>
Figure 6 shows the distance between the 11 topics generated by LDA. Some topics appear overlapping and close together, meaning that some topics share the same words and have the same level of similarity. It can be interpreted that most of the topic distributions from the LDA model on the dataset are still within the same scope.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Terms</th>
<th>Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>fakultas, didik, digital, seleggara, bidang, prof, indonesia, mahasiswa, bantu, wib</td>
<td>Prestasi, event</td>
</tr>
<tr>
<td>10</td>
<td>studi, mahasiswa, informasi, program, fakultas, sistem, indonesia, daftar, video, palapa</td>
<td>Prestasi</td>
</tr>
<tr>
<td>11</td>
<td>pancasila, indonesia, industri, maritim, pohon, selamat, ingat, semangat, trex, giat</td>
<td>Peringatan, trivia</td>
</tr>
</tbody>
</table>

The chi-square test was conducted to see which topics had posts with high and low engagement at three levels. The results of the chi-square test on 11 LDA topics are shown in Table 3. Low-engagement and high-engagement topics are represented as Low or High, respectively. It can be seen that at engagement level-1 and level-2, no topic has a significant number of posts with high engagement. At engagement level-3, only topic 8 (related to university achievements) has a significant number of posts with high engagement. This indicates that the engagement of each topic at three levels is equal or balanced.
3.3. Result of Engagement Score Prediction System

Decision trees and random forests are used to create models that can predict the number of engagement scores (number of likes and comments). Experiments were conducted on each model using three data sets: language, metadata, and language + metadata, on the same test data where the language feature is a caption text that has been represented as a vector with TF-IDF and metadata which contains topic features, textual features, and sentiment score.

As shown in Table 7, Decision Tree and Random Forest make the smallest errors in predicting test data when using language features. However, using a combination of language and metadata features, Random Forest predicts engagement scores more accurately than Decision Tree, with an MAE of 1074. Based on these results, the language feature has a higher impact on the engagement score prediction process than the metadata feature alone.

Table 4. Performance Evaluation Results of Each Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>Metadata</td>
<td>1508</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>1256</td>
</tr>
<tr>
<td></td>
<td>Language + Metadata</td>
<td>1358</td>
</tr>
<tr>
<td>Random Forest</td>
<td>Metadata</td>
<td>1508</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>1324</td>
</tr>
<tr>
<td></td>
<td>Language + Metadata</td>
<td>1074</td>
</tr>
</tbody>
</table>

4. CONCLUSION

This research uses LDA to help understand the topics in the university Instagram upload data. However, LDA has not been able to separate each topic well in the upload data because most of the topics identified by LDA have the same scope. Based on the results of topic identification with LDA and chi-square test analysis, it is found that topic 8 (related to university achievements) is a topic that has a significant number of posts with high engagement at level 3. In contrast, no topic has a significant number of posts with high engagement at levels one and 2. This indicates that the engagement of each topic generated by LDA at the three levels is equal or balanced. Decision Tree has the smallest error with an MAE of 1256 when predicting engagement scores using language features (TF-IDF), while Random Forest has the smallest error when using a combination of language and metadata features with an MAE of 1074. It can
be concluded that Random Forest has a better performance than Decision Tree in predicting engagement scores.

5. SUGGESTION

Based on the experimental results, to conduct a better analysis on topic identification, more dataset instances especially posts on after pandemic time will give better insights, not only limited to the pandemic time posts. The preprocessing steps also could be improved to reduce the unimportant words. Other features extractions and machine learning algorithms should be explored to find the most optimum method on this case study.

ACKNOWLEDGEMENT

The author would like to thank the university, family, and everyone who has provided support throughout this time.

REFERENCES


