Internship Facebook Project

Phi Thai Nhat

2022-10-12

I. PROJECT OVERVIEW

The Facebook Project was given with two data sets: Page Data and Post Data. The data have some trouble that one can not read them straight into R. Instead, the author had to transform the data manually and then read it. For the codes that read the data works smoothly, one can click to download the transformed data.

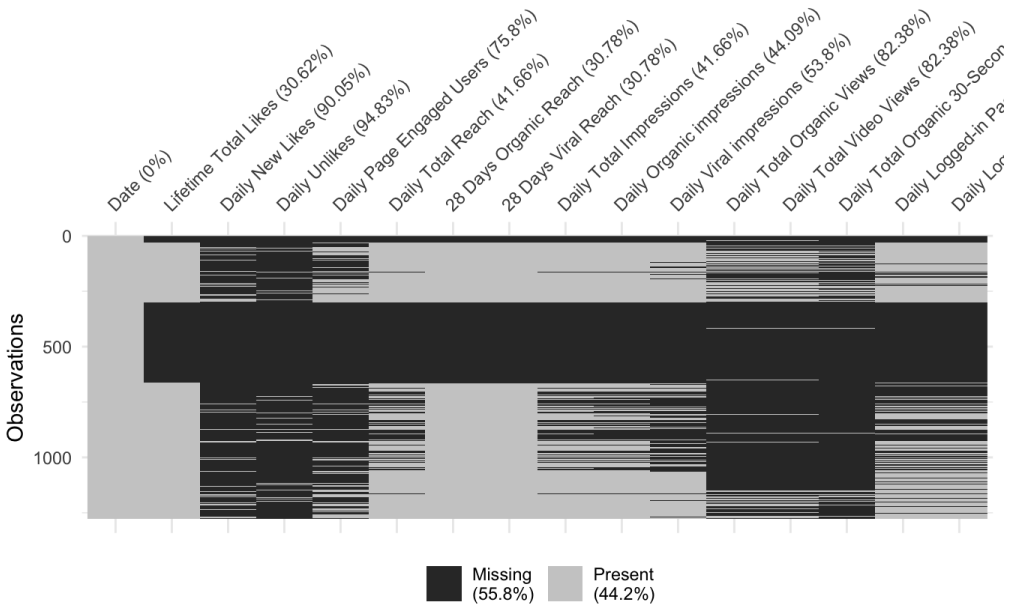
- Page Data (https://drive.google.com/drive/folders/1SMgtW8D2G1cZBxcx9EIvmAw4xVQNkEqt?usp=sharing): The information of posts published on Saigon A.I fan page on Facebook platform from February 2019 to July 2022.
- Post Data (https://drive.google.com/drive/folders/1TKbd6d8BpnPoJLWy4VWZwMdmFPI60Pjs?usp=sharing): The fluctuation of some metrics of Saigon A.I fan page on Facebook platform from February 2019 to July 2022.

The purpose of the project is to consolidate the author's knowledge during the last three months of learning in the company, mainly focusing on exploratory data analysis. It is anticipated that the outcomes will be data visualizations that offer some insights.

II. PAGE DATA ANALYSIS

For the Page Data, the author will read it in, check for missing values, subset the data to visualize some key metrics of SAIGON A.I Facebook's fan page.

1. Read the data and check for missing values

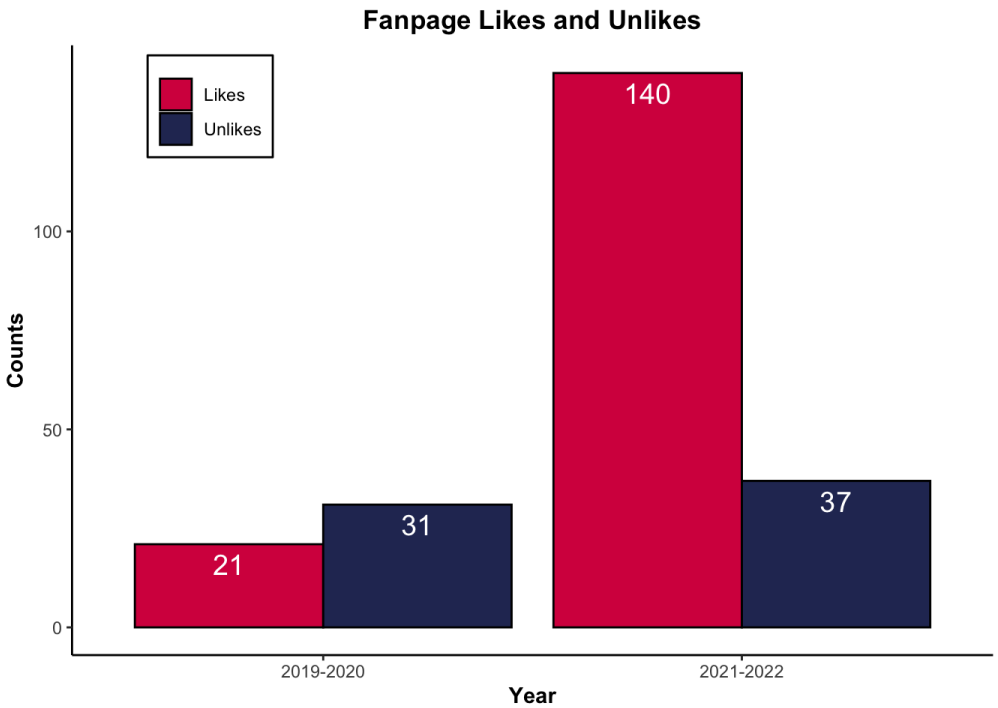


Page Data's missing values.

It can be easily seen that the missing values are more than half of the data frame. With suspicion, the intern double-checked with the transformed data set. It turned out that zero values in the Excel files are converted into NA values in R. So that, the intern replaced all the missing values with "0".

2.Visualize and interpret the data

1.The number of fan page likes and unlikes



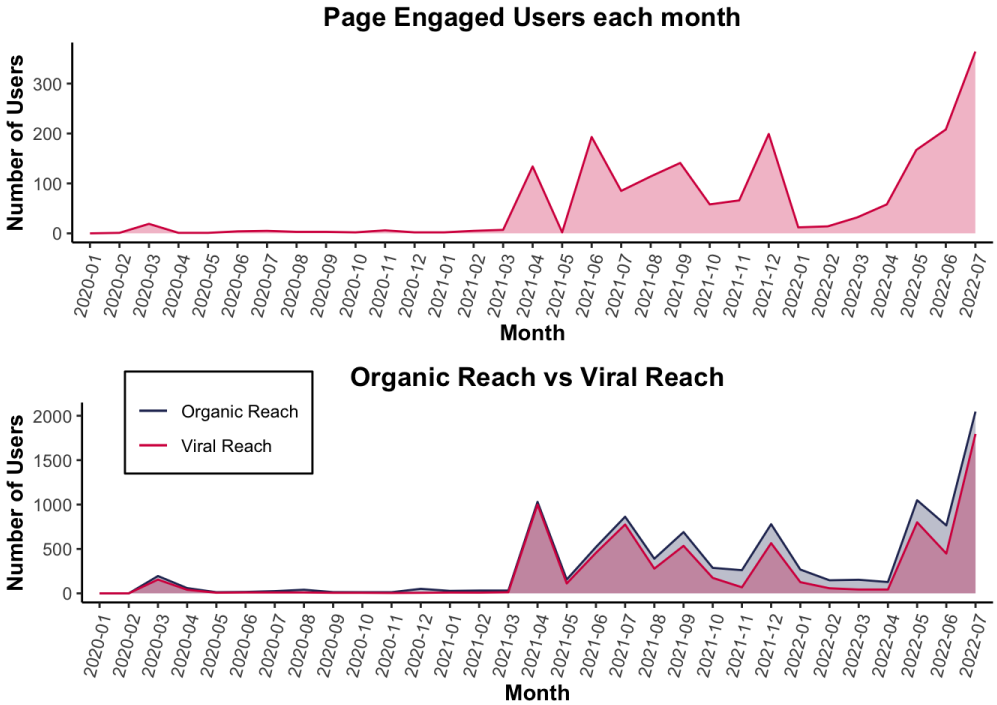
The fan page's likes and dislikes for the years 2021–2022 and 2019–2020 were represented by a bar chart. From 2019 to 2020, the number of people who unliked the fan page was bigger than the ones who liked it. In contrast,

between 2021 and 2022, the number of likes outweighed the number of unlikes.

It might be a good sign that the fan page's contens are more appealed to the audience. This figure might be connected to the metrics below.

2. Engagement, Reach, and Frequency

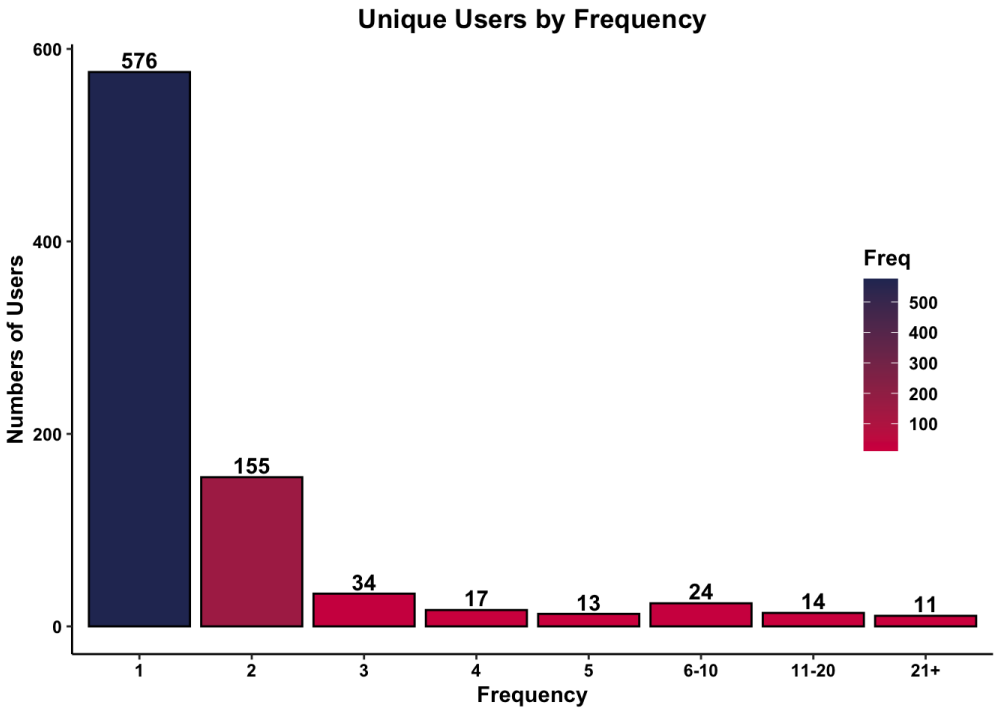
- **Engagement**: Simply put, Facebook engagement is defined as any action that someone takes on one of our posts or comments in pages. This includes any reactions, comments, shares, as well as link clicks.
- **Reach**: Reach is the number of people who saw any content from your Page or about your Page



Overall, there were similar trends of Engagement and Reach in the line graphs above. Both charts indicated that the key metrics of the Facebook fan page only rose after March 2020.

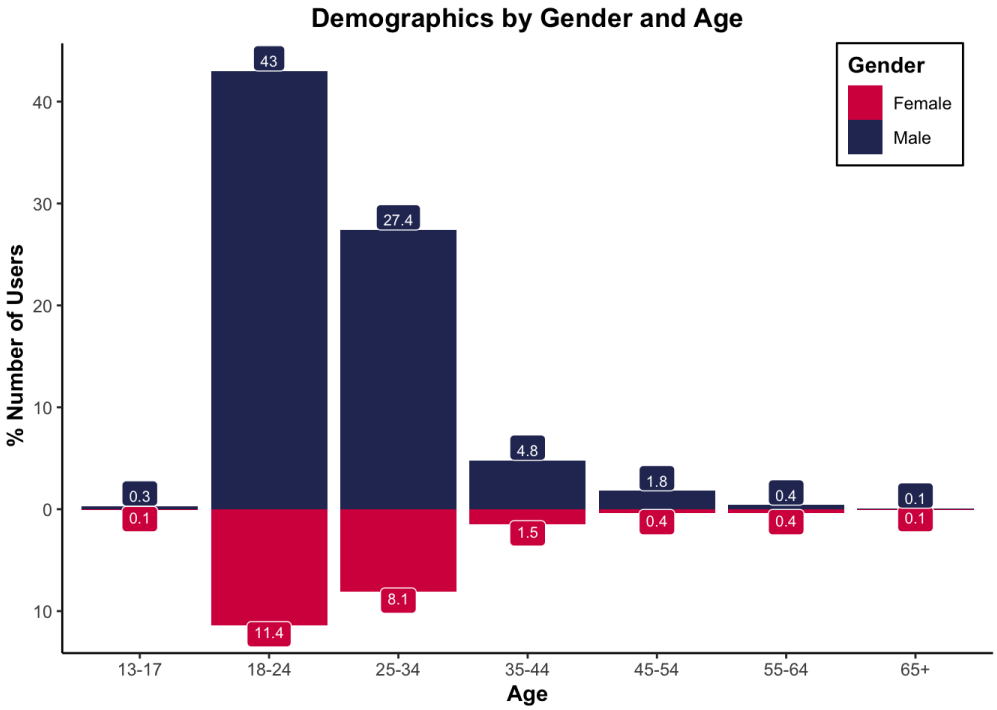
From January 2020 to early 2021, the number of users that the fan page could reach was so low at the bottom. Hence, the Engagement was also not significant. Afterwards, the lines went up and fluctuated, then peaked in July 2020 with more than 2000 reach and 300 engagements from users. Interestingly, the number was only low at the beginning of each year.

Since the metrics seemed brighter after the early 2020, the more the fan page posts reached audiences, the more the audiences would interact and found it interesting. Hence, the number of likes of the fan page was increasing significantly.

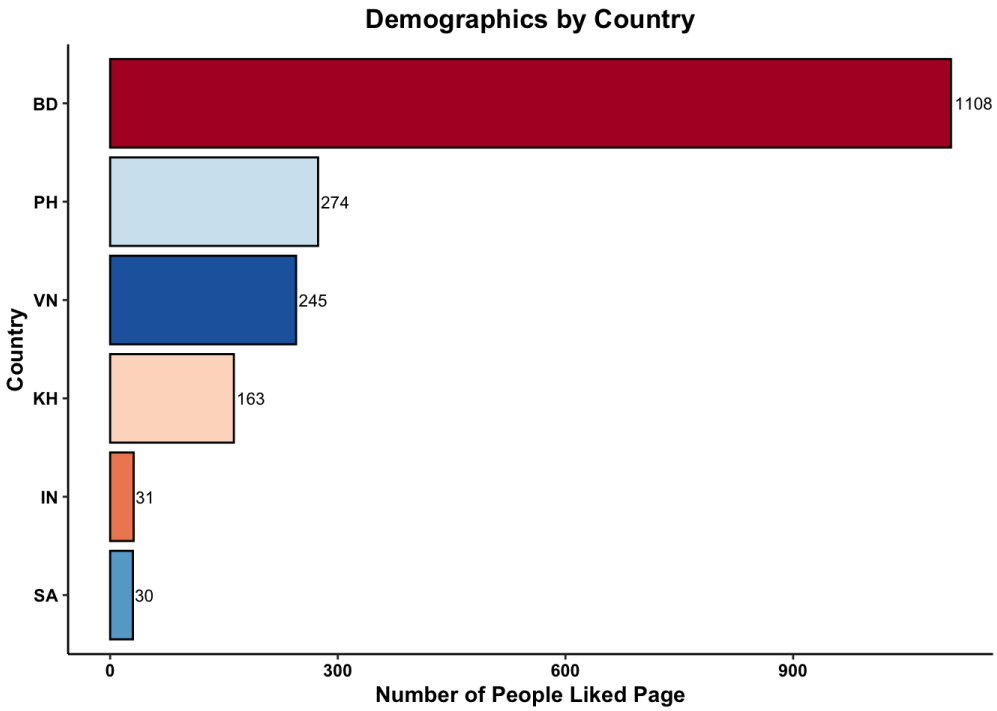


- **Frequency**: The number of people who saw the Page posts, broken down by how many times people saw the posts. The graph above plainly demonstrates that most of the people was only reached once.

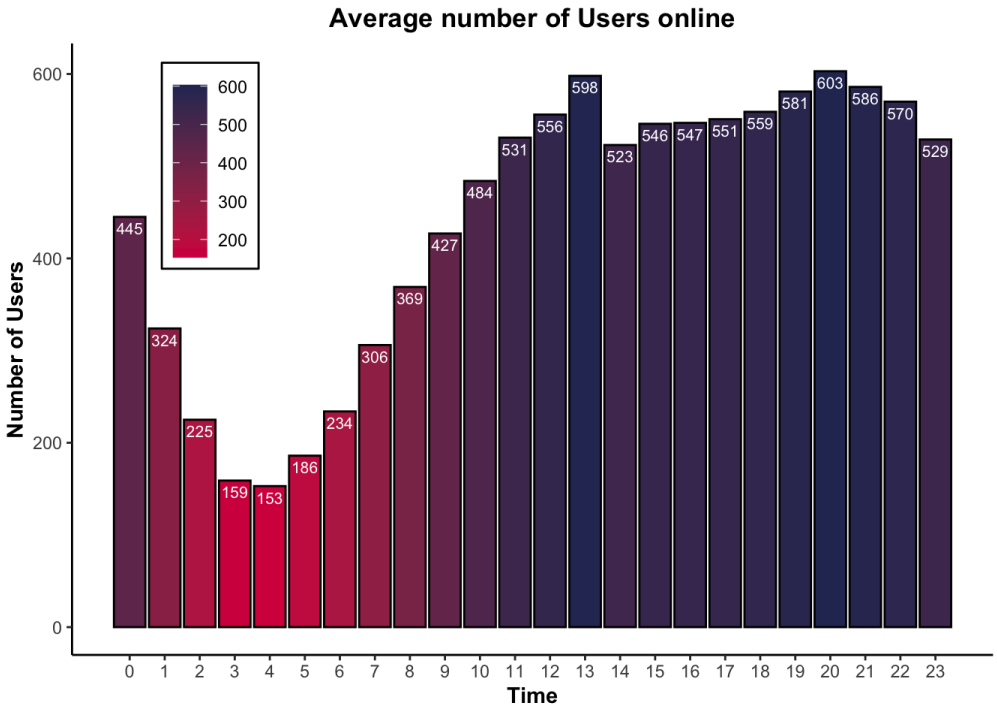
3. Demographics and Best time to post



There were 1978 fans in total on July 30, 2022. The gender distribution of the fans, which is the key demographic element, reveals that 77% of them are male and only 23% are female. Additionally, the graph below shows that compared to the other age groups, those between the ages of 18 and 24 and 25 and 34 are more inclined to interact with Saigon A.I.'s Facebook fan page. It could be said the fan page was hitting the right spot of the target audience in terms of age. However, with the aim of bringing more women into the industry, the fan page needs putting more efforts.



According to the graph above, it is surprising that most of the users who liked the fan page were not from Vietnam when the number of users is broken down by the nation where they reside. The majority of individuals that like the Facebook fan page are from several other Asian countries.

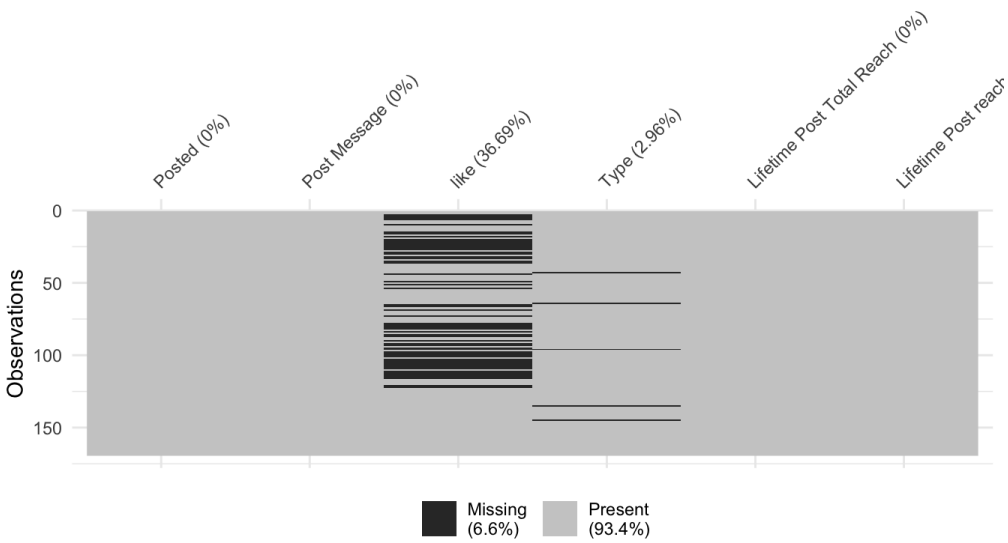


Some attempts have been done to change the timezone to fit with Vietnam hour in order to be able to plot the chart accurately. From the plot, we can see that the "golden hour" to post is 1 PM and 8 PM since these are the times that have the most users online, which could increase the probability of the fan page's posts reaching their newsfeed.

III. POST DATA

1. Pre-processing

Checking missing values



Punctuation removal and text lowercase.

For example:

Before

```
## [1] "How IoT And Artificial Intelligence Are the Perfect Partn
ers To Boost Business Productivity"
```

After

```
## [1] "how iot and artificial intelligence are the perfect partn
ers to boost business productivity"
```

Stopwords and non-English words removal

- **Stopwords**: Words that are very commonly used in a language but are not very informative.
- **Non-English words**: In this data set, it is Vietnamese texts because most of it is just the translation of English content.

Continue with the above example, after remove stopwords and non-English words, the text now becomes:

```
## [1] "artificial, boost, business, intelligence, partners, perf
ect, productivity"
```

Lemmatization and stemming

- **Lemmtization**: considers the context and converts the word to its meaningful base form. For instance, lemmatizing the word "Technologies" would return "Technology."

- **Stemming**: Stemming is a process that stems or removes last few characters from a word, often leading to incorrect meanings and spelling. For instance, stemming the word “Technologies” would return “Technologi.”

The aim of this step is to enhance the results of topic modelling by limiting redundancy and miss counts of words, which would affect the probabilities in the LDA.

2. LDA Topic Modelling

Latent Dirichlet Allocation (LDA) Topic Modelling

> **The main principle of LDA is these 2 concepts:**
>
> > *Every topic is a mixture of words.*
> >
> > *Every document is a mixture of topics.*

The LDA is a technique developed by David Blei, Andrew Ng, and Michael Jordan and exposed in Blei et al. (2003). The LDA is a generative model, but in text mining, it introduces a way to attach topical content to text documents. LDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, one

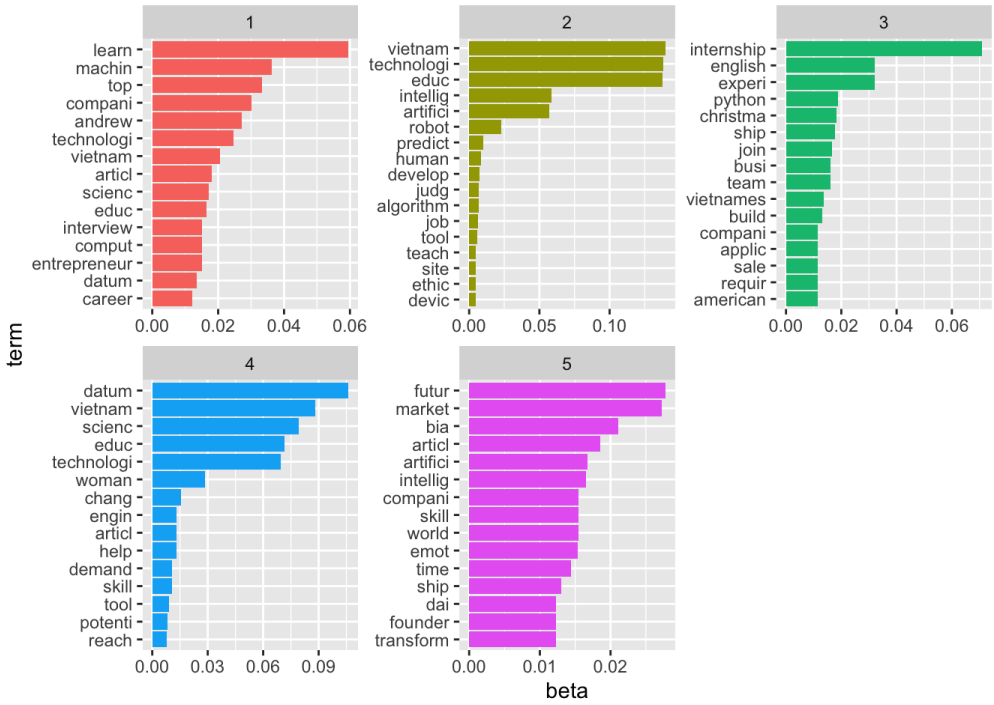
- Decide on the number of words N the document will have.
- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics).
- Generate each word w_i in the document by:
    - First picking a topic.
    - Using the topic to generate the word itself (according to the topic’s multinomial distribution).

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

The mathematics behind the LDA is beyond the scope of this work, however, if one wants to know deeper about LDA. This is the original paper: Link (https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)

3. Results

The top 15 terms that are most common within each topic

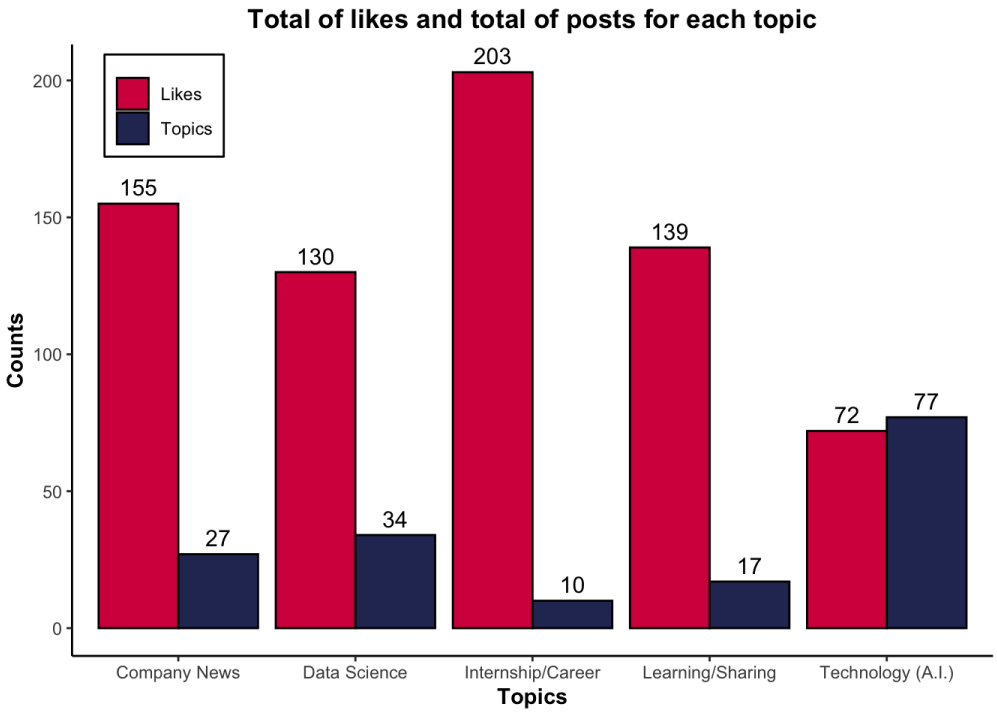


The main words of topic 1 are “learn,” “machine,” “top,” “company,” “vietnam,” “artificial,” “andrew” and “science,” which seem to be related to the learning of any who are interested in technology. Similarly, topic 2 is about Vietnam technology education with main words are “vietnam,” “technology,” “education.” The main words of Topic 3 are “internship,” “english,” “experience,” which clearly presents the internship in the company. Topic 4 seems to be related with data science since in Vietnam its words are composed of “datum,” “science,” “vietnam.” Topic 5 may be more about news of the company when checking with the post message. Hence, the inferred topics are determined as below:

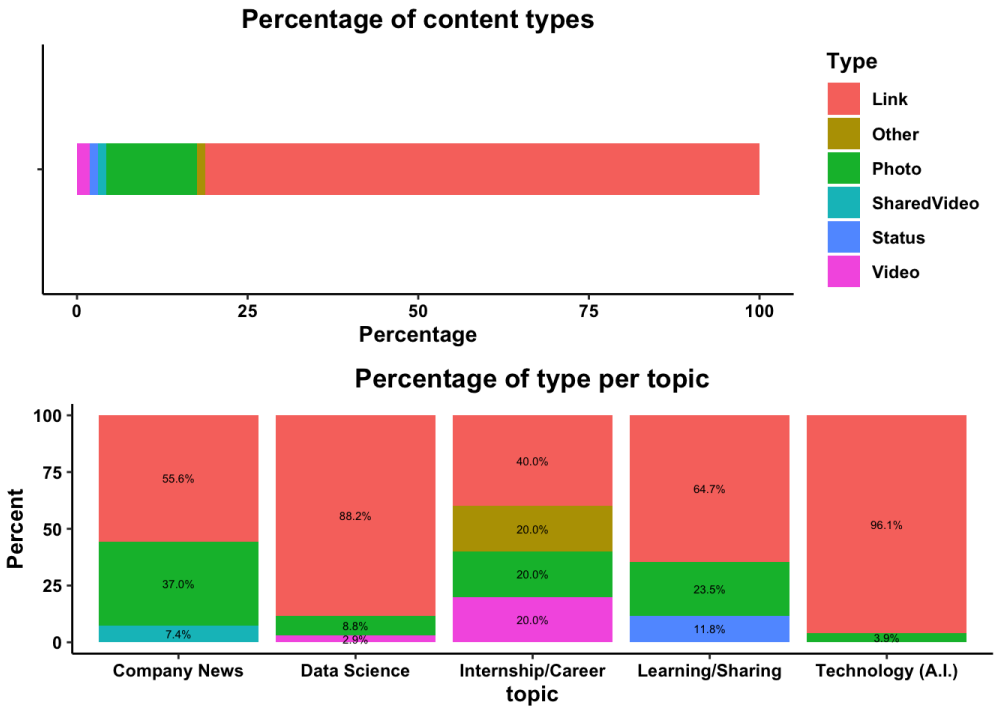
Topic	Top 5 terms	Inferred Topic
1	*learn, machin, top, compani, andrew*	**Learning/Sharing**
2	*vietnam, technologi, educ, intellig, artifici*	**Technology (A.I.)**
3	*internship, english, experi, python, christma*	**Internship/Career**
4	*datum, vietnam, scienc, educ, technologi*	**Data Science**
5	*futur, market, bia, artifici, intellig*	**Company News**

Topic, Types, and Like

The grouped bar graph below displayed the total number of posts and likes for each topic. When there were few posts and the total likes were frequently over 100, it was clear that the majority of themes were doing well.



With 203 likes after only ten posts, "Internship/Career" stood out as the most popular of those topics. Technology, which has 72 likes spread across 77 posts, is the topic that is least impressive at attracting followers.



With a percentage of more than 80%, "link" was the content type that the fan page concentrated on the most. But out of all content types, this one is the least appealing. As can be seen in the graph above, the topic "Technology" had 96% "link," which unintentionally had the fewest interactions despite having the most submissions. And among those topics, "Internship'Career," which only had 40% "link," received the most likes.

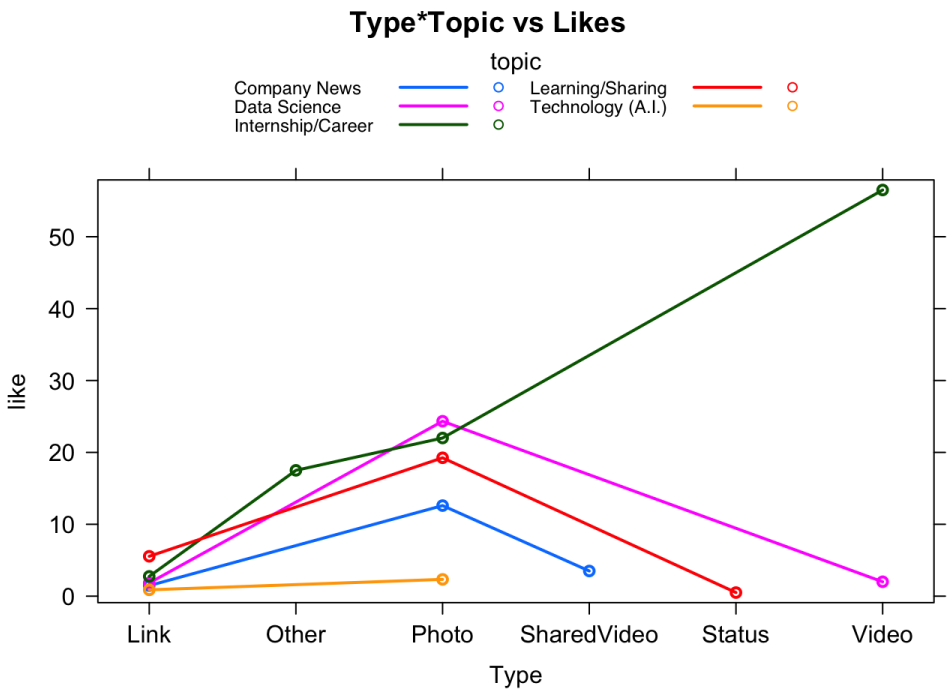
Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	5	7287	1457	44.63	4.838e-28
topic	4	862	215.5	6.6	6.41e-05
Type:topic	5	2305	461.1	14.12	2.581e-11
Residuals	150	4898	32.65	NA	NA

Type has a p-value <0.05 (significant), indicating that varied levels of Types are associated with varying numbers of likes.

The topic p-value <0.05 (significant), indicating that different topics are linked with substantial differences in the number of likes.

The interaction between Type and topic has a p-value < 0.05 (significant), indicating that the connection between topic and likes is influenced by type of the topic.



First, regardless of the post's topic, it is clear that "Link" was the least appealing content category.

Second, with the exception of the topic "Technology," "Photo" performed quite well when combined with other topics.

Third, the phenomena appears in the form of "Video" under the topic of "Internship/Career." The amount of likes was exceptional and far above the rest.

Finally, there were other combinations available to test for improved outcomes.

IV. CONCLUSIONS

With the support of recent content that has been going in the right direction, the main metrics for fan pages are generally increasing. “Photo” and “Video” are the content types that stand out among the rest. In contrast, the fan page should restrict update posts that contain links as this would reduce followers’ interest in the fan page.

Except for the topic “Technology,” other themes did a decent job of grabbing the audience’s interest. The topic “Internship/Career” stands out above the rest, so the fan page could provide more information regarding issues involving internships or careers at the company.

In addition, it is not required but recommended that the fan page should try out other different type and topic combinations to find out if there is any special occurrence emerge.

There also are some limitations in this report: 1) the sample is small and the types of posts are unequal, so that the results may be not representative; 2) the page data in 2019 is missing; 3) some visualizations could have been better. Hence, it is expected to have feed backs and updates from seniors.