Social Media Usage: A TikTok case study

Zohra Bouchamaoui – City, University of London

***Abstract (155 words)* —** The aim of this paper was to analyse TikTok's web scraped data and try to understand what drives the engagement and popularity of content. This analysis would allow content creators and businesses to identify whether certain topics are more popular than others and help them target larger audiences. For that, we used a variety of tools to determine what characterises the popularity of posts and whether these could be categorised into topic classes using hashtags. Furthermore, new features were derived from the older ones to try and give this analysis another dimension. In the end, we looked at topic modelling and classified the video captions using two unsupervised clustering algorithms, LDA and NMF, which proved to be pretty accurate with a coherence score of 67.1% for LDA. However, there were some limitations to the separation of topics due to the fact that some users add random hashtags or no hashtags at all to their captions.

Keywords — TikTok, User Engagement, Latent Dirichlet Allocation, Topic Modelling, Non-Negative Matrix Factorization

1. Introduction (160 words)

TikTok is a social media app that is built on sharing short videos. The platform offers a variety of filters, visual effects, and a large music library, which also gives the users the freedom to add their own sounds.

Since its global release in 2018, TikTok's growth has been explosive. The application has seen a massive influx of users, making it the number one most downloaded app of 2020 (Sehl, 2020). This digital transformation marked a change in the way users share information and interact with others. Today, TikTok is the 6th largest social network platform according to Sehl (2020).

As more and more companies are starting to turn to TikTok to target their market segments, it is convenient to analyse the users' behaviour and their interaction with the platform. This will allow creators and companies to get a better understanding of how to get the users to interact and engage with their products and services by using better content.

2. Research questions and data (289 words)

2.1. Data Source

The data we used to work through this analysis is focused on TikTok. However, it is hard to come across TikTok datasets from known repositories. Therefore, we used two resources to scrap the data necessary for our project:

- TikTok-Scraper(Nord, 2020)
- RapidApi TikTok(2020)

We scrapped the information about the trending videos using the first source and then scrapped the user profiles and the information about the hashtags using the second source.

2.2. Research Questions

TikTok is a platform that focuses on short-video sharing. Its main goal is to share entertainment videos where the most popular genres involve dancing, comedic skits, and or lip-syncing. However, the social platform started gradually to move towards a more commercially based usage, allowing the users to earn revenue from their videos and through advertisement. Moreover, as TikTok's reach increased, a lot of companies started to promote their brands on the platform.

As the platform gained more attention, its database grew with it. Therefore, we believe it would be interesting to put this data to use. On this basis, we perform this analysis with the aim to answer specific questions about the platform and to understand the engagement of the users with it.

- What characterises the popularity of a video?
- Can we generate topic classes? If so, how reliably?

By researching these questions, we aim to look for patterns which characterise the user's interaction and behaviour with the application which will help the content creators and companies to make better use of the platform as a marketing playground.

2.3. Methodology

- We connected to the API and extracted the data
- Pre-processing: we cleaned the data and looked at outliers.
- Derived new features
- Carried out an Exploratory Analysis of the data
- Performed classification and modelling

3. ANALYSIS (854 WORDS)

3.1. Data Preparation & Derivation

The dataset has 2081 rows and 39 columns. We dropped the columns that were of no use for our analysis and also decided to drop the rows with missing values. The data, after cleaning, has 1419 rows and 21 columns. Subsequently, we engineering new features using variables from the original dataset and added them to the updated table.

Then, we identified the outliers using the Mahalanobis distance as it is very efficient, mostly when there is a linear relationship between the variables (Ghorbani, 2019).

3.2. Exploratory Analysis of the data

3.2.1. Video Duration

One of the first features we looked at is the distribution of TikTok video durations on the Trend Feed. Videos on TikTok are limited to a duration length of 60 seconds, however, in Figure 1 we can see that most of the trending videos cluster around 10-15 seconds.

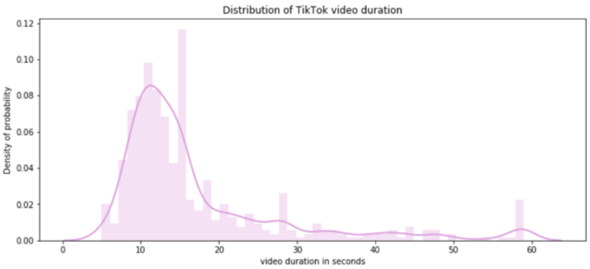


Fig. 1. Distribution of TikTok video duration

According to Ton (2019), the completion ratio of a video is the most important factor in the performance of video. When the ratio is above 0.5 then it means that the video is more engaging. However, when scraping the data, we could not access the average length of time spent by the users on the video. Therefore, we could say that in order to increase the chance of a video performing relatively well and retain the viewers' attention, users should aim to make 10-15 seconds video. This is because shorter videos are more likely to be watched in full.

3.2.2. Hashtags

The second feature we looked at are the hashtags. We started by setting a hashtag pattern and then, we compiled all the words or characters, from the caption text of each video, that matched that pattern. We can see in Figure 2 below that the most important hashtags represent the 'For You Page' ('fyp'), which is a page curated by TikTok where the application shows videos that the user might like. Each individual's 'fyp' is unique to them.

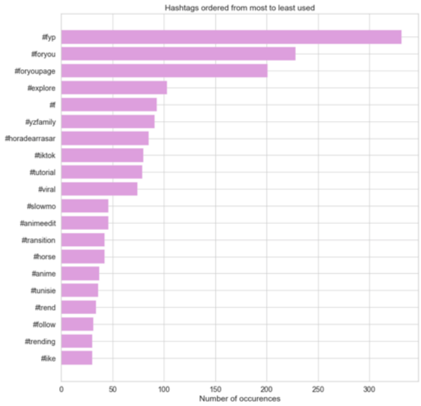


Fig. 2. Hashtags – from most to least used

3.2.3. Video Popularity

The popularity of a TikTok video can be defined in many different ways. The most reliable way to measure it is through the number of views. For the sake of our analysis, we engineered new features and looked at their correlation (*Figure 3*) before implementing them into a Multiple Linear Regression model.

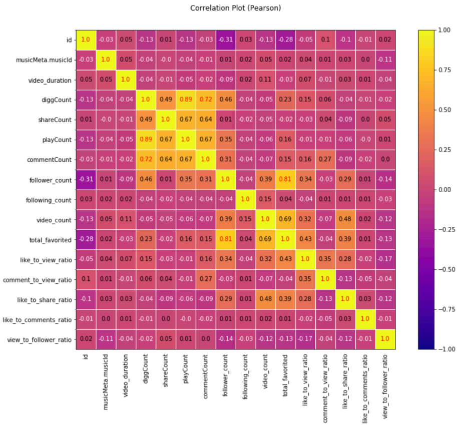
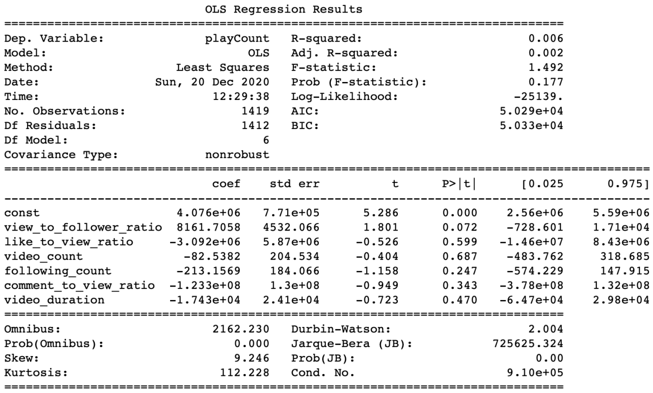


Fig. 3. Pearson Correlation Plot

We then performed an OLS regression using the least correlated features. The regression shows a skewness of 9.246 and an extremely small R-squared value of 0.006, which means that the model is a poor fit. Hypothetically, the number of views should increase as the users interact by sharing or commenting on the video, however, the OLS regression conveys that this hypothesis is wrong. This is probably due to the fact that videos with a smaller number of views could be shared more often, which is a circumstance that the regression would not be able to explain. The results for the number of videos and likes are the highest but are still insignificant. This can be explained by the fact that if a video has more views, it is more likely to reach a larger audience and maybe gain more likes than a video with a small audience.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              playCount   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     1.492
Date:                Sun, 20 Dec 2020   Prob (F-statistic):              0.177
Time:                        12:29:38   Log-Likelihood:                -25139.
No. Observations:                1419   AIC:                         5.029e+04
Df Residuals:                    1412   BIC:                         5.033e+04
Df Model:                           6
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                 4.076e+06   7.71e+05      5.286      0.000    2.56e+06    5.59e+06
view_to_follower_ratio 8161.7058  4532.066      1.801      0.072    -728.601    1.71e+04
like_to_view_ratio   -3.092e+06   5.87e+06     -0.526      0.599   -1.46e+07    8.43e+06
video_count            -82.5382    204.534     -0.404      0.687    -483.762     318.685
following_count       -213.1569    184.066     -1.158      0.247    -574.229     147.915
comment_to_view_ratio -1.233e+08    1.3e+08    -0.949      0.343   -3.78e+08    1.32e+08
video_duration       -1.743e+04   2.41e+04     -0.723      0.470   -6.47e+04    2.98e+04
==============================================================================
Omnibus:                     2162.230   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           725625.324
Skew:                           9.246   Prob(JB):                         0.00
Kurtosis:                     112.228   Cond. No.                     9.10e+05
==============================================================================
```

Fig. 4. OLS Regression results

3.3. Construction of models

Now let us look at the unsupervised labelling of classes using the textual caption from the videos, as well as their hashtags. Creating classes and labelling them manually would be extremely time-costly and inefficient as data is continuously growing and the videos' features keep on changing as users interact with them. We found that efficient ways to categorise this information into topics would be to use Non-Negative Matrix Factorization and Latent Dirichlet Allocation models. As we are clustering what can be considered as documents, we have to process each word individually to uncover topics.

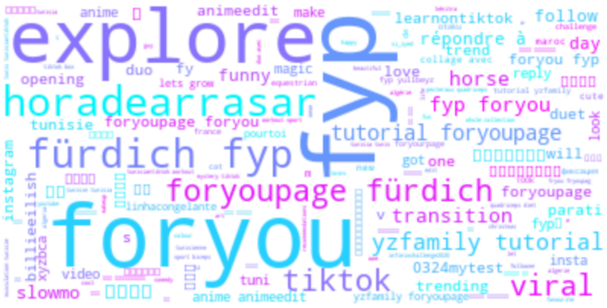


Fig. 5. Word Cloud

We started by cleaning the text, removing the punctuations, converting it to lowercase, etc. Then we printed the Word Cloud in Figure 5 make sure the cleaning was done properly. We then looked at the most common words in the caption as this will be important when creating the topics (Figure 6).

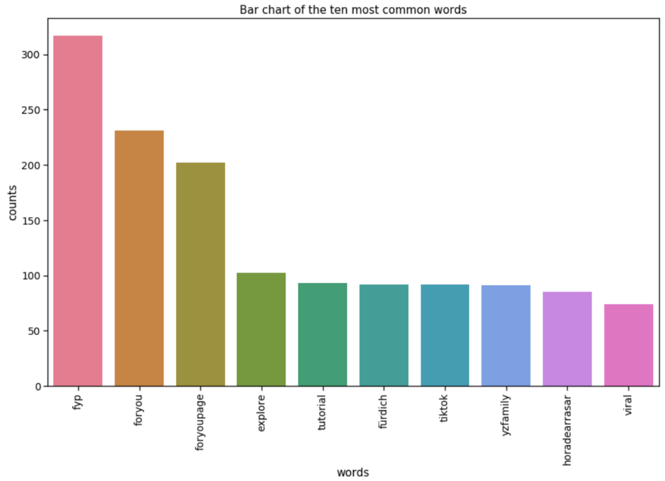
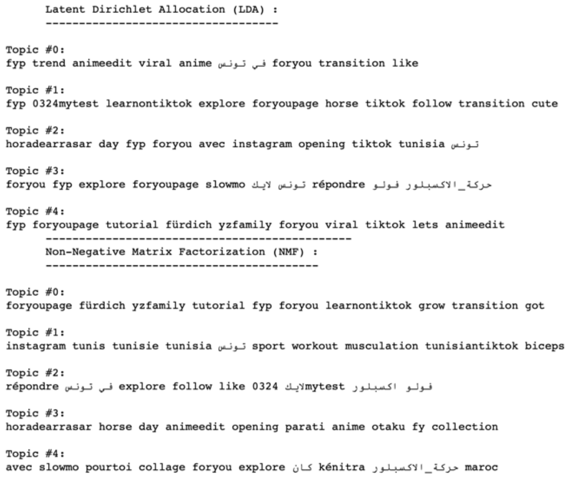


Fig. 6. Bar chat of the ten most common words

<u>3.1. Results</u>

Figure 7 shows displays five topics for each model. Additionally, when looking at Figure 8, we can see that the NMF model performs better than LDA as the separation of the clusters is clearer.



```
      Latent Dirichlet Allocation (LDA) :
      -----------------------------------

Topic #0:
fyp trend animeedit viral anime في تونس foryou transition like

Topic #1:
fyp 0324mytest learnontiktok explore foryoupage horse tiktok follow transition cute

Topic #2:
horadearrasar day fyp foryou avec instagram opening tiktok tunisia تونس

Topic #3:
foryou fyp explore foryoupage slowmo تونس لايك répondre حركة_الاكسبلور فولو

Topic #4:
fyp foryoupage tutorial fürdich yzfamily foryou viral tiktok lets animeedit
      -------------------------------------------
      Non-Negative Matrix Factorization (NMF) :
      -----------------------------------------

Topic #0:
foryoupage fürdich yzfamily tutorial fyp foryou learnontiktok grow transition got

Topic #1:
instagram tunis tunisie tunisia تونس sport workout musculation tunisiantiktok biceps

Topic #2:
répondre في تونس explore follow like 0324 لايكmytest فولو اكسبلور

Topic #3:
horadearrasar horse day animeedit opening parati anime otaku fy collection

Topic #4:
avec slowmo pourtoi collage foryou explore كان kénitra حركة_الاكسبلور maroc
```

Fig. 7. LDA and NMF results

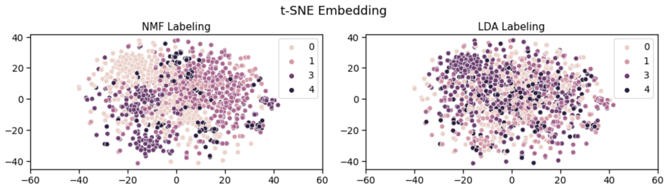


Fig. 8. t-SNE Embedding plot of NMF and LDA labelling

In Figure 9, we can see the perplexity and coherence metrics for the Latent Dirichlet Allocation. The perplexity measures how well a probability model predicts a sample. However, this measure is vain when not used in association with another model's value. (Kapadia, 2019)

We encountered some issues when it came to the metrics of the Non-Negative Matrix Factorization model and could not compute them in the given timeframe.

Metrics	LDA	NMF
Perplexity	-7.487981888	-
Coherence Score	0.670597278	-

Fig. 9. Table of validation results for LDA and NMF

4. Findings and Discussion (577 words)

At the beginning of this paper, we have asked ourselves on what characterizes the popularity of a video and whether we would be able to generate topic classes. We have concluded that the best way to analyse the popularity of a video is through number of views that it has. The more views a video has, the larger the audience it reached. Therefore, we have used this feature as our dependent variable. We have then used features derived from the original dataset, as well as other regressors to determine whether it would be possible to increase the number of views by focusing on specific statistics (*Figure 4*).

Therefore, if companies or users would wish to expand their popularity, they would have to do it manually. This would mean creating a great content or collaborating with more popular users on the platform. This is also why a lot of companies do reach out to users with high follower counts in order to collaborate and promote their products.

Although, we saw in the data that videos that are shorter and use commonly known hashtags could potentially improve the reach by a very insignificant amount. The Word Cloud in Figure 5 shows that using a variety of languages except English could potentially increase the range of audience viewers from other countries. This is due to the fact that, when a user uses different languages in their captions and hashtags, i.e., German, Arabic, English, and French, the TikTok Algorithm would assign their videos to different countries where those words are used more often.

During the validation process, we have realised that the results for the Non-Negative Matrix Factorization could not be made due to the deprecation of the necessary function on Python and thee lack information that could be found online. This creates a possible issue when one would like to use metrics to compare the two models. Due to the time restrictions, we were unable to make a deeper research to find a better fitting model and therefore, no metrics were added for that model.

Summurasing, there is a variety of characteristics that can characterise a video's popularity but there are no guarantees that focusing on these features will improve the popularity of the video. Furthermore, as the data obtained was unlabeled, it appeared to be difficult to classify the dataset using k-fold cross-validation. This is because, TikTok does not classify its videos by topic but mainly by the way users interact with the app. Also, many videos did not have hashtags, or had typos in their hashtags which could be one of the reasons why the topic modelling was not very accurate. That being said, it is still possible to categorise some of the videos using this variable, as long as the hashtags are accurate and spelled correctly. In a further study, it would be interesting to look at the number of posts and views for each hashtag as this could give hashtags more weight.

Lastly, another aspect that would be interesting to analyse in the future would be to consider whether music that is used in the videos could impact the number of views significantly. This is because, a lot of users who like the background music in a video will go look at other videos using that sound. Although, it is commonly known that one does not like too much repetition and therefore, this strategy only works for videos that are at the top of the application.

References

[1] City, University of London (2020). 'Lab 2 Feedback'. In3061 Principles of data science

[2] City, University of London (2020). 'Lab 3 Feedback'. In3061 Principles of data science

[3] City, University of London (2020) Lab feedback 05 - In3061 Principles of Data Science

[4] City, University of London (2020) Lab 08 - Text - Feedback. IN3061 Principles of Data Science

[5] CR (2020). 'Topic Modeling using Gensim-LDA in Python'. [Online]. Available at: https://medium.com/analytics-vidhya/topic-modeling-using-gensim-lda-in-python-48eaa2344920 [Accessed on 20 December 2020]

[6] Ghorbani, H. (2019). '*MAHALANOBIS DISTANCE AND ITS APPLICATION FOR DETECTING MULTIVARIATE OUTLIERS'*. [online]. Available at: <https://core.ac.uk/download/pdf/233075917.pdf> [Accessed on 20 December 2020]

[7] Kapadia, S. (2019). 'Topic Modeling in Python: Latent Dirichlet Allocation (LDA)'. [online]. Available at: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0 [Accessed on 20 December 2020]

[8] Nord, A. (2020). 'TikTok Scraper'.[online]. Available at: https://github.com/drawrowfly/tiktok-scraper [Accessed on 16 December 2020]

[9] RapidAPI (2020). 'TikTok'. [online]. Available at: https://rapidapi.com/logicbuilder/api/tiktok/endpoints [Accessed on 16 December 2020]

[10] Scikit-learn (2020). 'Topic extraction with Non-negative Matrix Factorization and Latent Dirichlet Allocation'. [online]. Available at: https://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py [Accessed on 17 December 2020]

[11] Sehl, K., 2020. '*20 Important Tiktok Stats Marketers Need To Know In 2020'*. [online] Social Media Marketing & Management Dashboard. Available at: <https://blog.hootsuite.com/tiktok-stats/> [Accessed 16 December 2020]

[12] Tafasca, S. (n.d.). 'On Online Media Consumption : A Youtube Case Study'

[13] Ton, H. H. (2019). 'How to Crack the TikTok Algorithm to Create Viral Videos'. [Online]. Available at: https://medium.com/@henryhienton/how-to-crack-the-tiktok-algorithm-to-create-viral-videos-d8a00e38e5ae [Accessed on 20th December 2020]

[14] Zornoza, j. (2019). 'Visualisation of Information from Raw Twitter Data — Part 1'.[online]. Available at: https://towardsdatascience.com/visualization-of-information-from-raw-twitter-data-part-1-99181ad19c [Accessed on 20th December 2020]