NOTTINGHAM UNIVERSITY BUSINESS SCHOOL

Academic Year: 2020 – 2021

ANALYTICS SPECIALISATIONS & APPLICATIONS

2nd Coursework

BRAND ANALYSIS USING TWITTER DATA

Student ID: 20243144

Jun 2021

CONTENT

A. EXECUTIVE SUMMARY 3

B. TECHNICAL APPROACH 3

C. DATA COLLECTION & PROCESSING 4

D. PERFORMANCE ANALYSIS 5

Overall statistics

Sentiment Analysis

Topic Modelling

E. MICRO INFLUENCER 7

Network analysis

Micro influencer recommendation

F. CONCLUSION 8

A EXECUTIVE SUMMARY

The aim of this snapshot study is to provide Marvel Studio insights on their audience via Twitter discussion. Based on that, recommendation on appropriate micro-influencers and other improvements have been made to help the label strengthen its branding. In providing a performance benchmark, Marvel's key competitor, DC cinematic universe (Warner's Bros.) was also investigated in parallel.

Through the process of data collection, transformation, and modelling, in total 167,786 qualified tweets over a week of tracking were harvested and analysed. Overall, Marvel Studio led the competition in both quantity and quality, when its audience size, user engagement and positive sentiment were significantly higher than the rival's. This was thanks to the robustness and well-established of the label's character network in the cinema industry, that provides numerous topics for fans to discuss on.

A quick look on users' network revealed that Marvel fandom was more loyal and condensed. Their interactions mostly happened between fan accounts, with less appearances showed on community page. Based on these findings, the recommendation for the brand is to have not only a micro-influencer to keep up the hype, but also a community partner to reach competitor's audience and non-fan.

B TECHNICAL APPROACH

To achieve mentioned objective, the analysis went through 3 keys stages.

First, in data collection and processing stage, tweets were loaded using search algorithm and a try and except approach to maintain the streaming without disruption. The collected data was then examined to identify useful columns and sample base for analysis (Section C).

Second, overall performance of the two brands was investigated using key statistics such as volume and unique values of different features (Section D.1). The report then provided a more in-depth understanding of users' opinions using sentiment analysis and topic modelling. A range of different techniques in cleaning the text (NLTK, Spacy, TweetPreprocessor) and deriving key themes (CountVectorizer, LDA, NMF) was examined (Section D.2).

Finally, network analysis based on the replies and retweets was made to understand the dynamic and engagement level of each fanbase, fuelling more insights to identify suitable micro-influencers. The parameters extraction involved 'NetworkX' package, while visualisation was supported by Gephi software (Section E).

C DATA COLLECTION & PROCESSING

The nearly 7 day-streaming record started from '2021-05-10 17:35:41' to '2021-05-17 07:14:24'.

The search-term used for Marvel Studio was 'mcu' (Marvel Cinematic Universe, referred as MCU) due to its uniqueness but commonness within the fandom. In providing more insights on the landscape, 'dceu' (DC Extended Universe – DCEU), representing MCU's direct competitor, was also tracked.

The collection process relied on search algorithm from Twitter API package, with the involvement of Try and Except statements to avoid disruption of data stream (Jupyter notebook – Data collection-MCU/DECU). The collected batches were then then merged and processed to clean up missing, duplicated values.

Throughout the period, 139,794 tweets were harvested for MCU from 82,501 Twitter accounts, while DCEU received 27,992 tweets from 19,296 users. The data table included in total 39 features. A quick statistical check was made to determine useful data fields for further analysis.

Overall, except columns that have meaningful null value (for example, 'reply_to_status_id' or 'quote_status_id' values are null when tweet is an original post), the features with highest missing values are those related to location (99.99%). Thus, no specific geographical area was restricted in analysis.

The truncated ('text') and extended tweet ('extended_tweet') were merged into a new column for full content ('full_text') while the 'user' feature contains account profile was unpacked. The key information selected for analysis were user's ID and screen name. Similarly, 'retweeted_status' field with information on the IDs and owners of shared posts was also decoded.

Final columns, their meaning and origins were included in Table 1.

Feature	Meaning	Original column
'id'	Tweet ID	'id'
'created_at'	Tweet created time	'created_at'
'in_reply_to_status_id'	ID of original post this tweet replies to	'in_reply_to_status_id'
'in_reply_to_user_id'	ID of original post owner this tweet replies to	'in_reply_to_user_id'
'lang'	Language detected for this tweet	'lang'
'user_id'	ID of this tweet's user	from 'user'
'user_name'	Screen name of this tweet's user	from 'user'
'in_retweet_to_status_id'	ID of original post this tweet retweets	from 'retweeted_status'
'in_retweet_to_user_id'	ID of original post owner this tweet retweets	from 'retweeted_status'
'full_text'	Full content of the tweet	'text' + 'extended_tweet'
'network'	ID of original post owner of this reply/retweet	'in_reply_to_user_id' + 'in_retweet_to_user_id'

Table 1. Summary of included features

D PERFORMANCE ANALYSIS

Overall Statistics

From the number of tweets and accounts collected as mentioned in Section 3, MCU was superior not only in the discussion volume and user base but also engagement level of audience (1.7 versus 1.5 tweets per user for DCEU).

MCU's audience also appeared to be more active and willing to voice their opinions when original posts and replies accounted for around 31% and 15% of total tweets respectively, while for DCEU the figures were approximately 25% and 10%.

The most retweeted (4530 times) and responded (110 times) threads for MCU were fandom-oriented (a meme and a friend-making post), while for DCEU the hottest posts (2139 retweets and 63 replies) were more movie content related.

Figure 2 illustrates the contribution of top 10 languages ('lang') returned for the search terms. English was dominant in the list, accounting for 69% of MCU's tweets and 67% of DC's, with Portugal came second of around 7.5% contribution for both. However, the 3rd and 4th positions differed between 2 labels with Japanese and Indian appeared more popular for MCU, and Korean and Thailand for DCEU.

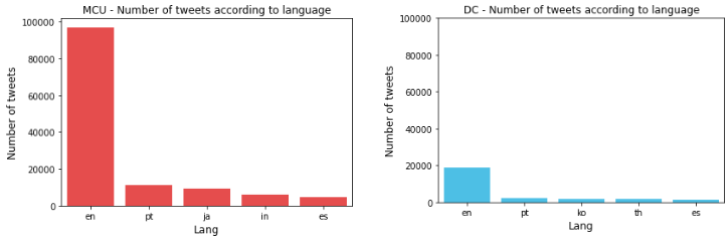


Figure 2. Number of tweets by language ('lang')

Though the absolute numbers of tweets in Korean and Thailand for DCEU were higher versus MCU (1788 vs. 631 and 1788 vs. 484 respectively), a cross check using Google Trend showed that these differences only meant on Twitter platform, not truly reflect the popularity of the 2 studios in mentioned countries.

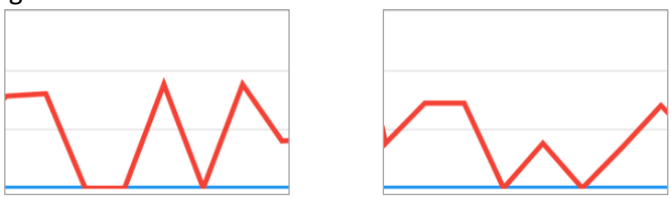


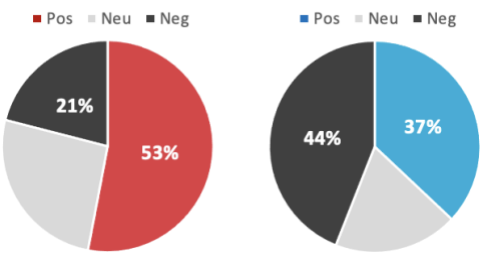
Figure 3. Number of search in Google for MCU (red) and DCEU (blue) from May 10 – May 17

Due to its large proportion, English verbatims was used for further analysis.

Sentiment Analysis

The analysis involved identifying the percentage of positive, neutral, and negative elements from each tweet using Sentiment Analyzer (NLTK package). The final decision on sentiment was then made by comparing the positive and negative scores ('positive' if pos > neg; 'negative' if pos < neg, 'neutral' is pos = neg). The feature used was original 'full_text' since the lexicon dictionary used in evaluation ('vader_lexicon') contains ranking for emoticons and all forms of word. Retweet samples were included as they represented the volume of agreements or disagreement towards the matter.

Figure 4. Tweets' sentiment proportion for MCU (red) and DCEU (blue)



The result showed that MCU received better feedback from its fandom versus DCEU, with over 50% positive and around 21% negative comments. On the other hand, the unfavourable proportion from DCEU search term doubled MCU's figure at 44%. What were the factors causing this phenomenon?

Topic modelling

To answer above question, key themes of the tweets were identified, then being mapped with the sentiment result to understand which were the more controversial topics.

Before modelling, comments went through a cleaning process, from removing hyperlink, emoticon, punctuation, to tokenizing and lemmatizing. The stop words list for each brand was added with respective search term to avoid confusion to model.

Two different unsupervised methods including Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) were examined. Exhaustive search on LDA showed that 8 topics for MCU and 6 topics for DCEU were the most optimal options delivering highest log-likelihood scores, and manual trials with NMF further supported this argument. Overall, NMF performed better in interpretation. This might result from the fact that LDA relies on calculating the distribution of words in each document and therefore work better with longer text. While in this case, tweets are short and often single topic-oriented, allowed NMF to optimise its capacity.

The topics, their interpretation, associated words, volume, and sentiment proportion were demonstrated in Figure 5.

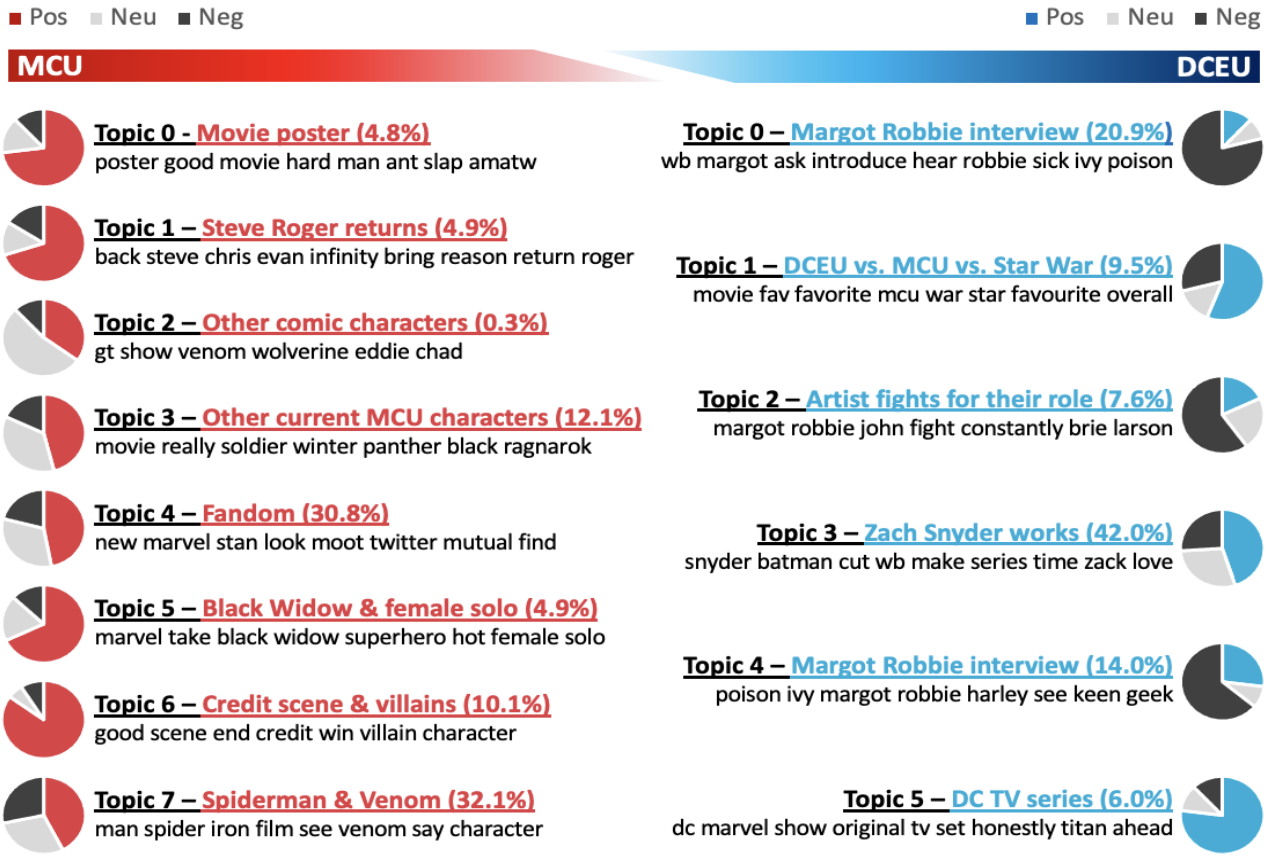


Figure 5. Topics, their sentiment (left pie chart), volume proportion and key associated words

For MCU, the hottest topic included discussion around Spiderman and Venom (originally a support role in Spider Man series) (Topic 7). Surprisingly this was also the topic with highest proportion of negative comments. A qualitative investigation into unfavourable tweets showed that the discussion mostly evolved around the comparison between MCU's Spiderman and previous Sony Picture versions, and whether Venom will make guest appearance in MCU. This conversation boomed up as Venom 2 coming to theatre in June, and Spiderman 3 in December.

Apart from that, the feedbacks towards MCU were mostly positive, especially regarding credit scenes, villain characters (Topic 6) and movie poster (Topic 0). There was also a hype for upcoming projects of Black Widow, the hinted female superhero movie (Topic 5), and the hope that Chris Evan (playing Captain America) will come back to MCU with more explanation on how he returned the Infinity Stones in Endgame (Topic 1).

Regarding DCEU, the main factor led to its high proportion of negative comments arisen from recent Margot Robbie's interview (Topic 0, 4). In the article, the actress shared on how she actively asked Warner Bros. to include a related anti-hero character (Poison Ivy) in DCEU movies but not yet succeeded. Most of the complaints directed toward the studio, triggering discussion on other artists' '*fights for their character to be in queue*' (Topic 2). The most positive theme recorded for DCEU was regarding their TV series that seemed ahead of MCU's (Topic 5), while the hottest thread was about Zach Snyder and his works.

Another interesting observation was that while DCEU rarely appeared on MCU fanbase discussion, the opposite trend recorded on the competitor side. This not only showed that MCU fans were more solely loyal, but also indicated the brand's potential to tackle competitor's audience.

E MICRO INFLUENCER

Network analysis

To understand users' interaction and from that determining appropriate influencers, Network analysis was applied. In a more technical explanation, the considered data included replies and retweets, while the people who responded or retweeted ('user_id) are source nodes and the owners of original tweets ('network') are target nodes. User ID was used due to its permanent and unchangeable characteristic. The data was then analysed using NetworkX package and visualised using Gephi software (Force Atlas 2 layout algorithm).

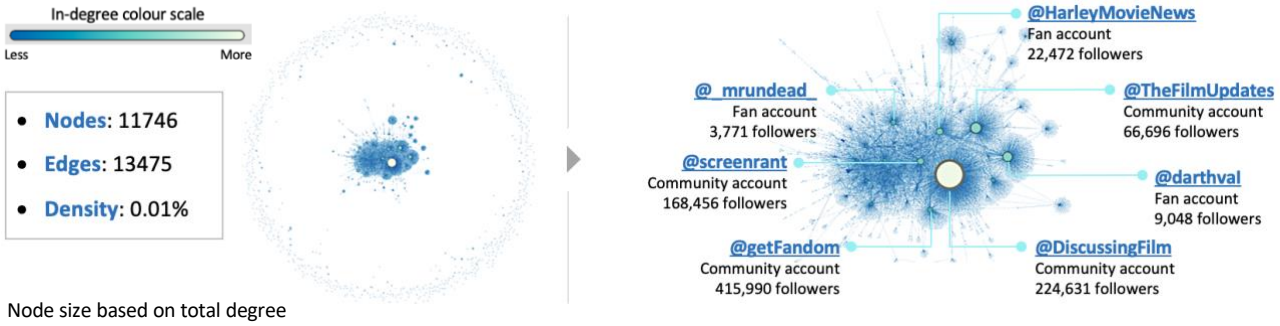


Figure 6. DCEU users' network, parameters and key nodes

Looking at the network (Figure 7), communication within DCEU fanbase was highly centralised, mostly towards big and neutral communities. There was also an absence of Official studio in the largest nodes, showing less engagement between fandom and produced label.

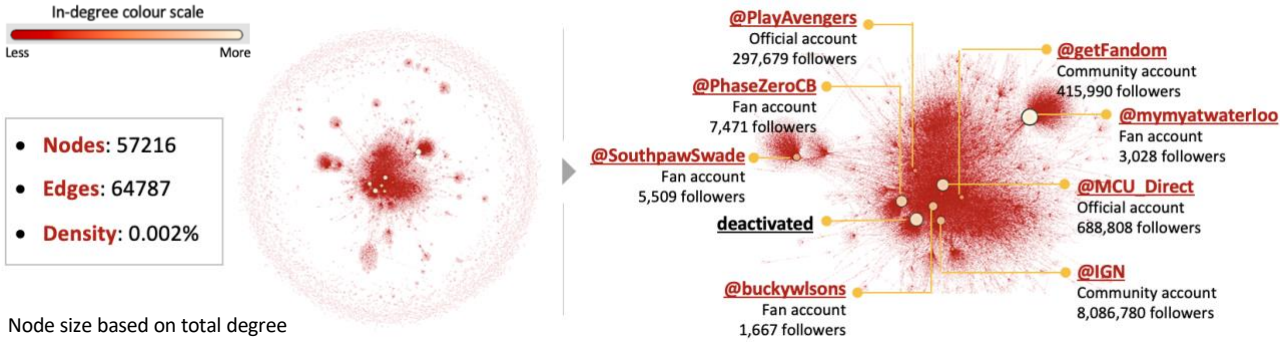


Figure 7. MCU users' network, parameters and key nodes

On the other hand, though having lower density due to its popularity among mass audience, the core fanbase of MCU showed a more solid shape. Biggest targeted nodes include several fan accounts and micro influencers, while studio's official accounts including MCU_Direct and Marvel's Avenger still appeared in Top 15 in-degree nodes (Figure 8).

Micro-influencer recommendation

Effective influencers were evaluated based on 4 key aspects:

- **Strengthening loyalty:** their impact to the network represented by the total in-degree interactions they received.
- **Tackling competitor:** their ability to reach competitor's audience showed in the number of interactions with DC fanbase they had.
- **Approaching non-users:** their overall popularities to reach the non-fan audience based on number of followers.
- **Quality of content:** the sentiment of their tweets towards MCU and the replies they received. Afterall, the influencers should obtain and spread-out positive brand image.

Based on that, it is recommended for MCU to develop partnership with 2 groups of influencers, the micro-ones that further increase fandom engagement, and the semi-micro KOLs who have large audience and reachable to DCEU fanbase. The profile of the most potential candidates can be found in Figure 8. Depend on the project budget, more accounts can be considered via detailed list on Jupyter notebook.

Note: Manual sanity check was carried out on Top accounts to select the most relevant

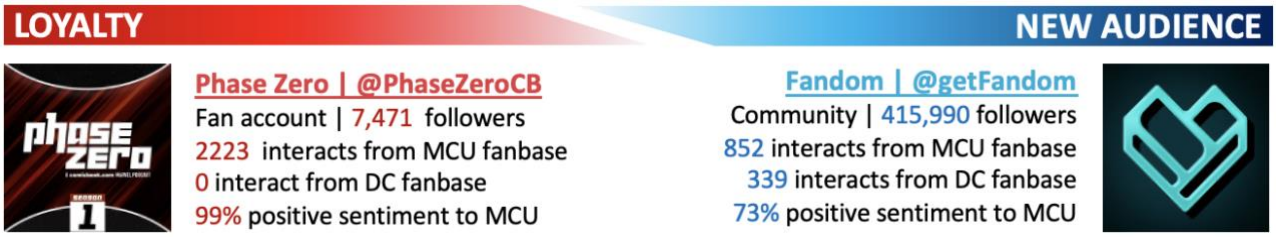


Figure 8. Recommended influencers for MCU and their key stats

F CONCLUSION

On the bright side, MCU has proven its success and popularity, not only versus DCEU but in general by attracting large conversation volume and user base on Twitter. The sentiment was also highly positive across different topics.

Having said that, the brand still has room for further improvement.

First, a major proportion of the tweets still focused on previous projects, with much smaller space left over for the current on-air characters such as Wanda, Vision (March), Falcon (May), and coming-soon Loki (June). This showed that the new roles and series have not yet drawn significant hype and would need more time and effort to build up engagement with audience.

Second, network analysis revealed that due to its robust fandom, MCU presence on community channels was still limited versus its size. Hence, more effort in engaging with these pages would help the brand to reach more competitor's audience and non-fans. This could include sponsoring mini-game and give-away events.

Third, in leveraging this snapshot study, it is recommended that MCU continues the data tracking every week to capture the trend in sentiment and discussion. The multi-period data would then be analysed to have a more reliable look into users' opinions. Search terms regarding specific characters, movies and series, especially on-air ones could be investigated for portfolio development. Other languages and platforms, wherever possible, should be added on future study to fuel insights on different parts of the world.