Social Media Analytics: Pink Venom Album

BA2 Group 17

Han, Joy; Sun, Haichao; Zhou, Sophie

MBAN, Sauder School of Business

University of British Columbia

BAIT 508: Business Analytics Programming

Dr. Gene Moo Lee

October 7th, 2022

Han, Joy: qyjoy.han@gmail.com; 81278178;

Main: Part E and analysis; Support: Part A to D

Sun, Haichao: sunhc.2008@gmail.com; 25089153;

Main: Part C, D and E; Support: Part A, B, and analysis

Zhou, Sophie: sophie.littlestar@gmail.com; 21984133;

Main: Part A, B; Support: Part C, D, E, and analysis

Introduction

Blackpink, one of the world's most popular girl groups, is formed by South Korean YG entertainment. Members of the group include Jisoo, Jennie, Rosé, and Lisa. Motivated by our interest in Blackpink, we chose "pink venom album" as our keyword for the project. Pink Venom is one of the theme songs in Blackpink's second studio album Born Pink, released on September 16, 2022. Therefore, we wanted to take this opportunity to conduct social media analysis of the public's opinion about this song and the newly released album using our Python skills.

There are multiple steps to complete this project: First, we gathered 10K tweets related to the keyword we picked and collected author information. Second, we cleaned the data and performed a preliminary analysis of the information collected. Then, by utilizing word clouds, we elaborated data visualization and interpreted the Keywords related to this topic. Later, sentiment analysis was used to determine the polarity. Finally, data-driven insights will be concluded at the end of this report.

STEP A: Keyword Selection and Data Collection

As discussed above, "pink venom album" is our keyword. Then we used the Python script (tweet_collecton_example, Gene Moo Lee, Jaecheol Park and Xiaoke Zhang for BAIT 508) that the instructor team provided to collect 10K recent tweets on the selected keyword. Tweety package and Twitter collector were used. The tweets collected are in dictionary type and some major keys include: "id", "name", "description", "source", and "text". To collect the author IDs, we defined

a function: get_twitter_author_info(author_id). Both tweets and the author ID collected are stored in a json file.

STEP B: Preliminary Analysis

We used multiple packages such as pprint, counter, pickle, ntlk, plt, numpy and pandas to analyze the data. Results are illustrated as follows:

1. Ten most popular words with and without stop words:

By using a for loop to collect words in “text”, we find out the most popular words with stop words are: “RT”, “Pink”, “#1”, “200”, “Shut”, “Venom”, “Down”, “the”, “on”, “100”, shown as Screenshot 1. It is obvious that some words are meaningless such as “the” and “on”. Therefore, we created a stop words list containing the words we would like to omit. We created a new list (words2) which excludes the stop words. The most popular words without stop words are: “Pink”, “#1”, “200”, “Shut”, “Venom”, “Down”, “100”, “Hot”, “Global”, “▫Billboard”, shown as Screenshot 2. These words are highly related to our keyword: “pink venom album”. The numbers such as “#1” and “200” were not filtered out since they are related to the album’s global ranking. “#1” is the ranking on multiple platforms, “100” and “hot” is the Hot 100 rank, and “200” is the Billboard 200 ranking.

```
from collections import Counter

words = []
for tweet in data['tweets']:
    txt = tweet['text']
    words.extend(txt.split())

Counter(words).most_common(10)
```

```
[('RT', 4313),
 ('Pink', 2832),
 ('#1', 2221),
 ('200', 2106),
 ('Shut', 1924),
 ('Venom', 1886),
 ('Down', 1835),
 ('the', 1825),
 ('on', 1821),
 ('100', 1691)]
```

Screenshot 1

```
words2 = [] # our accumulator list

for w in words:
    if w not in stopwords and len(w) > 1:
        words2.append(w)

c2 = Counter(words2)
c2.most_common(10)
```

```
[('Pink', 2832),
 ('#1', 2221),
 ('200', 2106),
 ('Shut', 1924),
 ('Venom', 1886),
 ('Down', 1835),
 ('100', 1691),
 ('Hot', 1618),
 ('Global', 1522),
 ('▫Billboard', 1507)]
```

Screenshot 2

2. Ten most popular hashtags

We also used the for loop to search for hashtags in our created list (words2). Initially, some ranking numbers, such as #1 and #10, are misleading. Therefore, we excluded ranking numbers and the results are: “#BORNPINK”, “#BLACKPINK”, “#PinkVenom”, “#Official_Audio”, “【#BLACKPINK】”, “#BLINK”, “#PINKVENOM”, “#JENNIE:”, “#블랙핑크”, “#JISOO”, shown as Screenshot 3. These hashtags are closely related to the album, the group and the 4 group members.

```
In [5]: hashtag_blackpink = []
        for w in words2:
            if "#" in w:
                if not ((w.replace("#", "")[0]).isnumeric()):
                #Ranking Number is used in tweet text, so filtering out the case with #number
                    hashtag_blackpink.append(w)

        Counter(hashtag_blackpink).most_common(10)

Out[5]: [('#BORNPINK', 694),
         ('#BLACKPINK', 291),
         ('#PinkVenom', 88),
         ('#Official_Audio', 71),
         ('【#BLACKPINK】', 54),
         ('#BLINK', 54),
         ('#PINKVENOM', 14),
         ('#JENNIE:', 14),
         ('#블랙핑크', 11),
         ('#JISOO', 8)]
```

Screenshot 3

3. Ten most frequently mentioned usernames

Similarly, for loop was used to search for “@” in the words2 list. The results are: “@BLACKPINK”, “@BLACKPINKSTATS5:”, “@ZEXIONOXIOUS:”, “@cryforthepinks:”, “@chartsblackpink:”, “@minsebornblonk:”, “@BLACKPINKGLOBAL:”, “@pinkIovesick:”, “@UK_BLINKS:”, “@lakrandhi95:”, and the usernames and frequencies could be found as Screenshot 4. Notice that most accounts are Blackpink official accounts and fan accounts.

```
In [6]: at_blackpink = []
        for a in words2:
            if "@" in a:
                at_blackpink.append(a.replace(".", ""))
                #Due to using "." with @ directly, replace "." with ""

        Counter(at_blackpink).most_common(10)

Out[6]: [('@BLACKPINK', 2449),
         ('@BLACKPINKSTATS5:', 979),
         ('@ZEXIONOXIOUS:', 752),
         ('@cryforthepinks:', 412),
         ('@chartsblackpink:', 360),
         ('@minsebornblonk:', 308),
         ('@BLACKPINKGLOBAL:', 270),
         ('@pinkIovesick:', 164),
         ('@UK_BLINKS:', 138),
         ('@lakrandhi95:', 103)]
```

Screenshot 4

4. Three most common sources

Then, a loop and counter were used to count the common source in the “source” key. The results are: “Twitter for Android,” “Twitter for iPhone,” and “Twitter Web App”, categories and frequencies are shown in Screenshot 5.

3 most common sources of the tweets

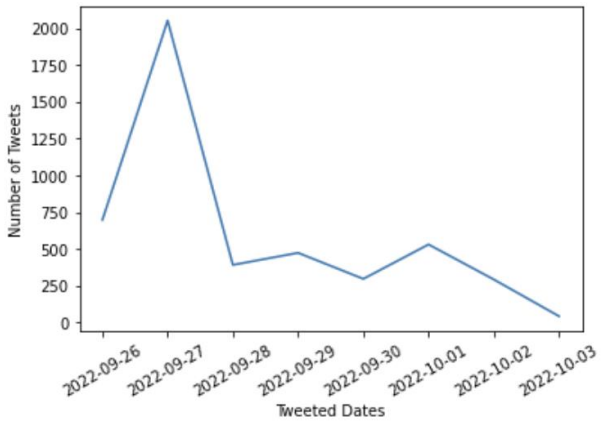
```
In [7]: platform = []
        for tweet in data['tweets']:
            source = tweet['source']
            platform.append(source)

        Counter(platform).most_common(3)

Out[7]: [('Twitter for Android', 2905),
         ('Twitter for iPhone', 1409),
         ('Twitter Web App', 360)]
```

Screenshot 5

5. Time trend line chart

We created a line chart to show the relationship between the number of tweets and the tweeted date. Firstly, we collected date data from the key “created_at”. Then we counted the tweets for each date and sorted them from oldest to newest. Finally, by applying “matplotlib.pyplot”, we created a line chart to show the trend. As the chart (Screenshot 6) indicates, there are more tweets

when the date is closer to the Album release date, Sep.16. Especially for Sep.27, when Blackpink officially announced their “Comeback” event, the number of tweets has peaked.



Screenshot 6

6. Three most influential tweets

We used the pandas package and the for loop to find the top three influential tweets. Firstly, we collected the text in the tweets. Then we assessed the influence score by summing the “quote_count”, “reply_count”, “retweet_count”, “like_count” for the tweets. The result is shown below as Screenshot 7.

Out[9]:

	Tweet Text	Influential Score
267	RT @AppleMusic: The new era of @BLACKPINK has ...	27752
272	RT @BLACKPINK: [🎧]\n#BLACKPINK 의 선공개 싱글 'Pink ...	20381
269	RT @BLACKPINK: BLACKPINK 2nd Album ‘BORN PINK’...	17253

Screenshot 7

7. Three most vocal authors

Using “author_info_pink_venom_album.json”, we created separate lists for Username, Profile Name, Author ID, and Public Metrics to calculate User Score. As we already collected the counted tweeting activities on the topic, we used the count method to collect the number of tweets posted for each author ID. Further, we combined such information onto a data frame structure.

The author names corresponding with the IDs can be shown in Screenshot 8.

3 most vocal authors on Pink Venom

```
#Display DataFrame of Author data with focus on Pink Venom tweet topic count
all_author_data_df.sort_values('Tweeted Topic Count', ascending=False)[:3]
```

	Username	Author ID	Profile Name	Influential Score	Tweeted Topic Count
2529	sirsimonCwll	1566707541043982337	SirSimon Cowell	3299	33
1719	amiinno1	1415227479518707715	amiinno	78354	10
3461	dd12315	969681145	ABC[9327]	294506	8

Screenshot 8

8. Most influential authors

Building on question 7 we sorted the values in the data frame according to their “influential score”. The influence score is the sum of “followers_count”, “following_count”, “listed_count”. “tweet_count” in the author data frame. The author with the highest influence score is shown as Screenshot 9.

```
#Display DataFrame of Author data with focus on User Inflencial Score
all_author_data_df.sort_values('Influential Score', ascending=False)[:3]
```

	Username	Author ID	Profile Name	Influential Score	Tweeted Topic Count
2367	inquirerdotnet	15448383	Inquirer	4775745	1
2708	soslalisa	260390523	jon	2123837	1
1906	ZSinein	1446021164	ŠⓘⓝⓔⓘN	1387229	1

Screenshot 9

STEP C: Word Cloud

To show the popularity of words and phrases in a more vivid and audience-friendly way, we used a word cloud to show the most popular words in the collected tweets. Packages used include word cloud, matplotlib and nltk. First, we used the for loop to accumulate words in words2, to exclude the stop words. The accumulated words are stored in a variable: text2. Then, we used text2 to create the word cloud and set the form of the word cloud. The result is illustrated below as Screenshot 10.

Screenshot 10

The word cloud indicates that it is the most popular word in Pink Venom, which meets our expectations since this is our keyword. Other words such as Born Pink (the concert will be held at the end of 2022), Billboard, Hot, Global are all closely related to our keyword and show this Album's strong influence.

STEP D: Sentiment Analysis

1. Average polarity and subjectivity scores

To calculate the polarity and subjectivity scores, textblob is used. Firstly, we calculated the polarity for all tweet text and append them into a list called “tweets_text”. Then we determined the average by dividing the sum of polarity by the number of tweets. The average polarity is -0.00671 (rounded to 5 digits). Polarity spreads from -1 to 1, from most negative to most positive. The average polarity is close to 0, which means the polarity is approximately neutral. We used similar procedures for the subjectivity analysis. The average subjectivity is 0.364. Subjectivity spreads from 0 to 1.0 is non-subjective and 1 is exceptionally subjective. So, the average represents that the tweets are mainly objective. The results are shown as Screenshots 11 and 12 below.

Average polarity score

```
avgpol=sum(tweets_text_polarity)/len(tweets_text_polarity)
print("Average of polarity score of all tweets is", round(avgpol,5))
```

```
Average of polarity score of all tweets is -0.00671
```

Screenshot 11

Average subjectivity score

```
avgsub=sum(tweets_text_subjectivity)/len(tweets_text_subjectivity)
print("Average of subjectivity score of all tweets is", round(avgsub,5))
```

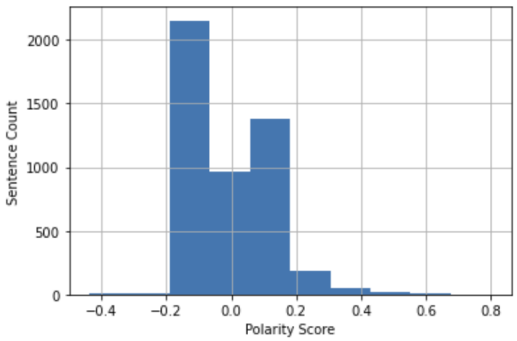
```
Average of subjectivity score of all tweets is 0.36446
```

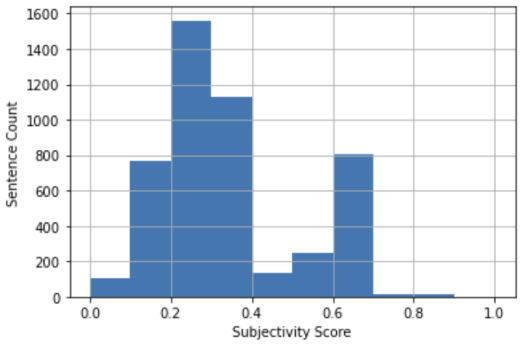
Screenshot 12

2. Sentiment analysis histogram

By applying the histogram function, we have the graph shown below. For the polarity histogram (Screenshot 13), the x-axis is the polarity score, and the y-axis means the number of

tweets. For the subjectivity histogram (Screenshot 14), the x-axis is subjectivity, and the y-axis is also amounts of tweets.





Screenshot 13 *Screenshot 14*

From the graphs above, we conclude that a high number of tweets have a low polarity (-0.2-0) and low subjectivity (0.2-0.4). There are multiple reasons leading to this result. It seems counter-intuitive, but some popular words like “pink” have a low polarity in the program. Hence, these popular words drag the average polarity to a negative value. Furthermore, since most tweets are facts about Album and concert information, the subjectivity score is also low.

3. Most positive and negative tweets

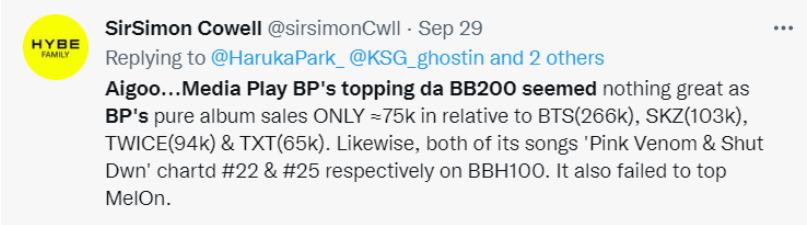
By applying the “sort” function, we acquired the results for the most positive and negative tweets on the keyword. The most positive tweets with the highest polarity score are unexpected. It is actually a very negative comment (retweeted multiple times) toward our keywords, the author expressed their disappointment toward this album (Screenshot 16). We tested multiple phrases and figured out that the program is misled by some particular words such as “great”. Even if the phrase is “nothing great” the polarity score is still as high as 0.8 (Screenshot 17) (SirSimon Cowell, 2022).

3 highest polarity scored tweets

```
Tweet_Score.sort_values(by='Polarity', ascending=False)[0:3]
```

	Tweet Text	Polarity	Subjectivity
369	RT @sirsimonCwll: @Legenkillerm2 @KSG_ghostin ...	0.8	0.75
368	RT @sirsimonCwll: @HarukaPark_ @KSG_ghostin @c...	0.8	0.75
375	RT @sirsimonCwll: @listeningRFL @boramussu @ch...	0.8	0.75

Screenshot 15



Screenshot 16

```
sentense="nothing great"
tb = TextBlob(sentense)
pol = tb.sentiment.polarity
print(pol)
```

```
0.8
```

Screenshot 17

And the most negative tweets with the lowest polarity score are mainly websites and links (Screenshot 18). Therefore, the program will assign them a 0-polarity score. Although this text analysis may not be accurate, it can represent extreme situations. Then we can imply the polarity and their attitude between these extreme situations.

3 lowest polarity scored tweets

```
Tweet_Score.sort_values(by='Polarity', ascending=True)[0:3]
```

	Tweet Text	Polarity	Subjectivity
101	@gretchenzenvox https://t.co/N0Efh7VbR1	0.0	0.0
388	RT @walenpink: Ya pero este BOP? https://t.co/...	0.0	0.0
556	はやく帰りたい… https://t.co/FiLESv5qAv	0.0	0.0

Screenshot 18

STEP E: Insights

This project analyzes the related information about the latest album “Pink Venom” on Twitter. Significant insights are gained from this project. During the project we mainly learned about text analysis by utilizing python. First, python skills we learned including how to collect data from social media and cleaning the data before analysis. Then we learned multiple methods of text data analysis that can be applied in future projects, such as frequency, source, sentiment analysis, etc. After reviewing our procedure, we noticed some limitations and improvements that we can make. For example, polarity analysis can be improved by training with more datasets to increase accuracy.

In the real world, this project result is meaningful for YG Entertainment, especially in their later direction on advertising and future genre on the publications. Possible applications from this project can be on Amazon and YouTube, for example. We could get the most influential singer among Blackpink, and publish a few peripherals of whom, to attract more revenue from potential groups of buyers. Or we could also get the most popular works and publish on YouTube to attract more clicks. By doing so the influence of Black Pink would be even higher.

However, there are a few limitations on our project. From a broader perspective, we will notice that majority of fans of Black Pink are young people, who are more likely using Instagram and TikTok, instead of Twitter. Future analysis will be open to more platforms to gather more data. Another limitation is about the data itself. Since the project is focused on an art topic, tons of pictures, videos and audio can be analyzed. Better understanding could be illustrated by also including these data.

Reference

SirSimon Cowell [@@sirsimonCwll]. (2022, September 29). *N.B. Aigoo...Media Play BP's topping da BB200 seemed nothing great as BP's pure album sales ONLY ≈75k compare to* [Tweet]. Twitter. https://twitter.com/search?q=Aigoo...Media%20Play%20BP%27s%20topping%20da%20BB200%20seemed&src=typed_query