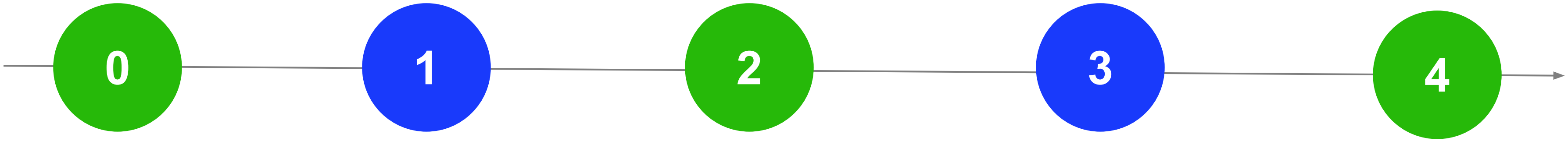
OPINION MINING FOR BRAND REPUTATION MANAGEMENT

DATA MINING

A.y. 2020/2021

Irene Bondanza
Matteo Emanuele
Alessandra Monaco

Table of Contents



INTRO

- Motivations

DATA COLLECTION

- Scraping from Amazon.us
- Scraping from Twitter

TOPIC EXTRACTION : Amazon

- Methodology
- Results

TOPIC EXTRACTION : Twitter

- Methodology
- Results

HORIZONTAL ANALYSIS

Motivations

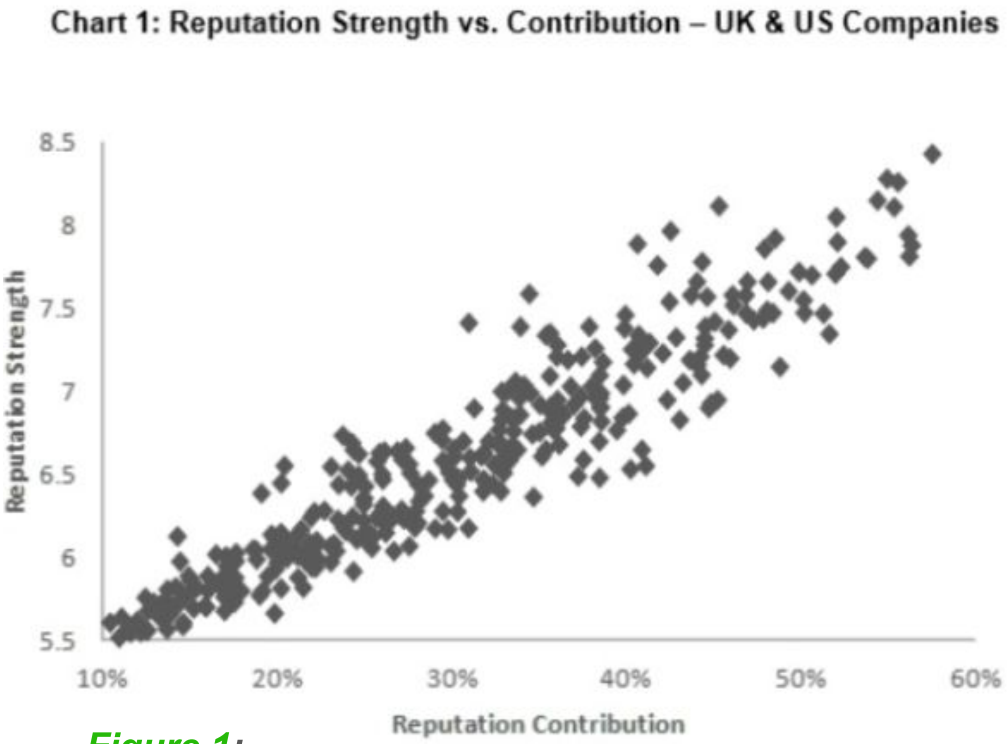


Figure 1:

"The Impact of Reputation on Stock Market Value", from "https://www.world-economics-journal.com"

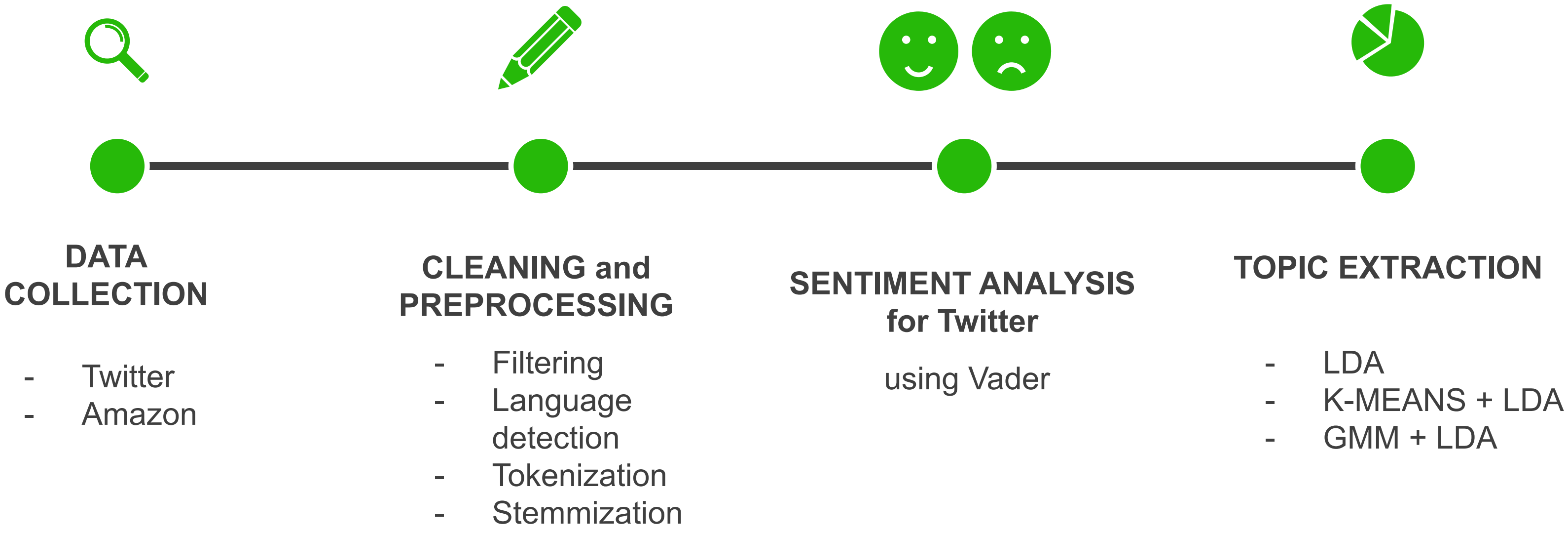
Opinion drives decision making in companies:

Any brand, nowadays, knows that one of the keys for success in their brand management rely on the opinion that the single customer has about the company

Reputation strength and its contribution:

Investors care about the popular opinion of the brand they are investing in. Stock price tends to be volatile with respect people's opinion.

Overview



DATA COLLECTION

- Twitter
- Amazon

CLEANING and PREPROCESSING

- Filtering
- Language detection
- Tokenization
- Stemmization

SENTIMENT ANALYSIS for Twitter

using Vader

TOPIC EXTRACTION

- LDA
- K-MEANS + LDA
- GMM + LDA

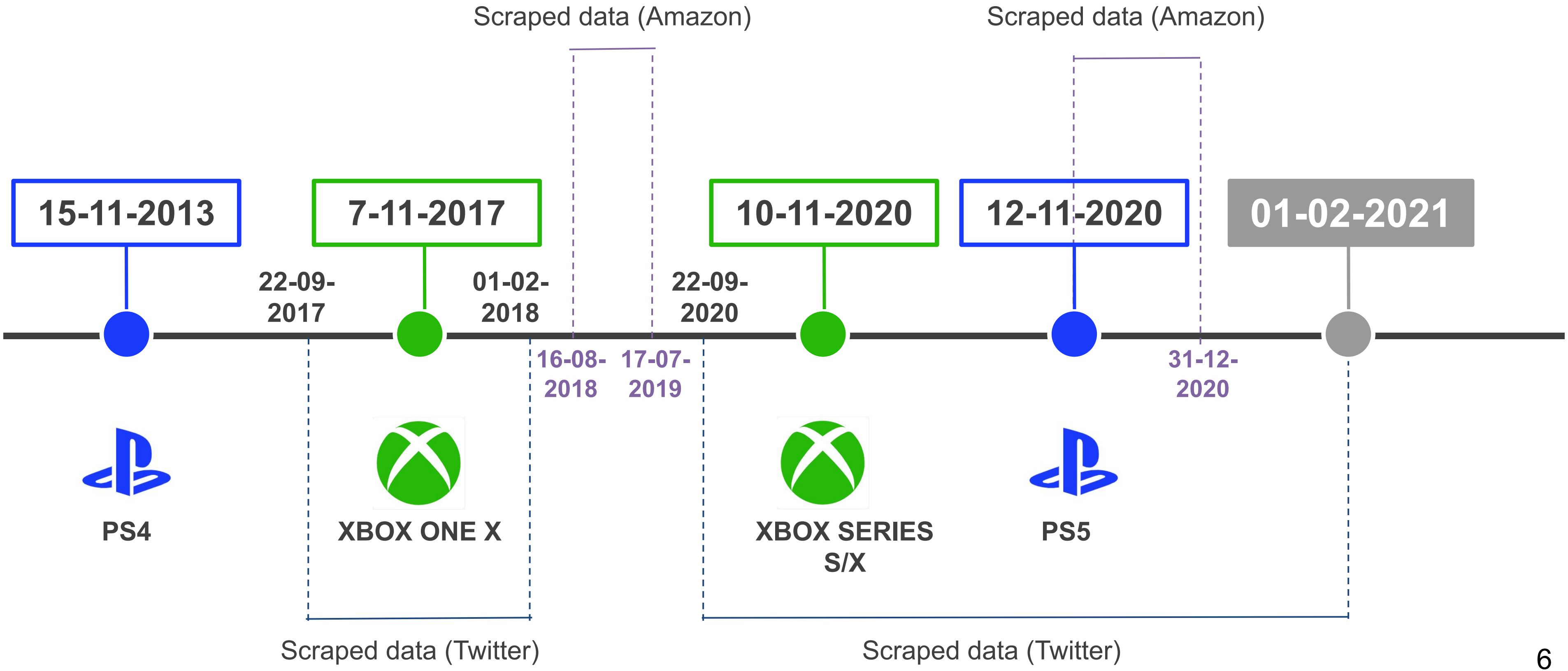
Challenges

Mostly about data.

- Amazon:
  Data presents both long and structured reviews with different topics and short reviews composed of few words

- Twitter:
  Data present hashtags in the middle of the corpus, together with **slangs-based expressions** and eventually **grammatical errors**. Retweets may be out of context if taken individually. Plenty of **spam** and/or **ads** related tweet with random hashtags.

Console timeline



Scraped data (Amazon)

Scraped data (Amazon)

15-11-2013

7-11-2017

10-11-2020

12-11-2020

01-02-2021

22-09-2017

01-02-2018

22-09-2020

16-08-2018

17-07-2019

31-12-2020

PS4

XBOX ONE X

XBOX SERIES S/X

PS5

Scraped data (Twitter)

Scraped data (Twitter)

The datasets

	CONSOLE/BRAND	N SAMPLES	TOTAL SAMPLES
Amazon 2017/18	Playstation 4	5.297	6.986
	Xbox One	1.689	
Twitter 2017/18	Playstation	4.029	8.699
	Xbox	4.670	
Amazon 2020/21	Playstation 5	2.221	3.420
	Xbox series	1.199	
Twitter 2020/21	Playstation	21.480	51.542
	Xbox	30.062	

The datasets

The different nature of datasets implies distinct collection of data

Amazon

- Author name
- Review’s text
- Number of stars
- Date
- Place
- Verified purchase
- Upvotes
- Language

Twitter

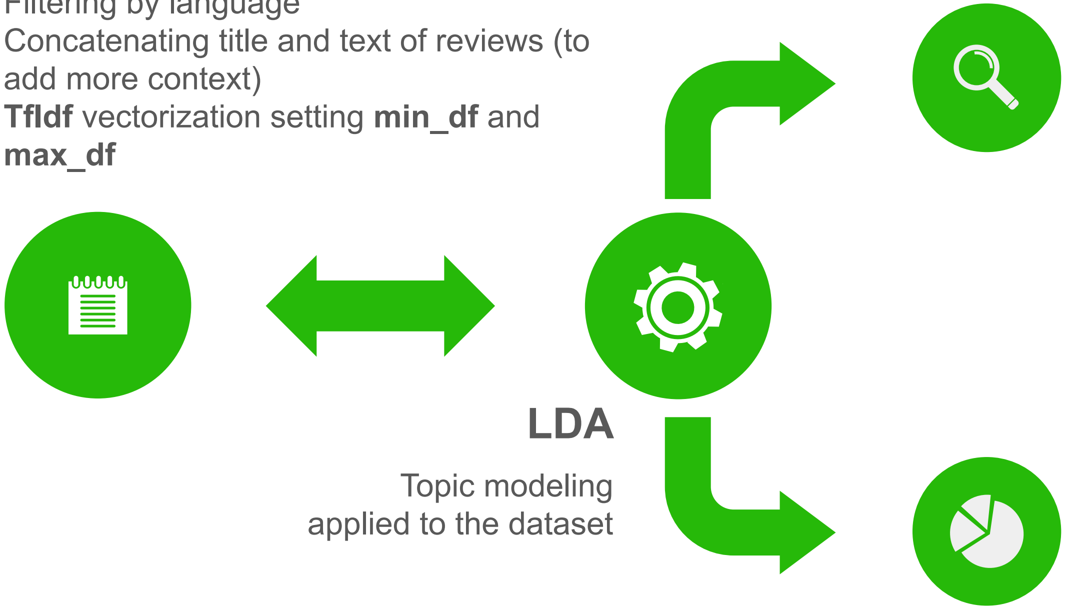
- Username
- Date
- Tweet text
- Retweet text
- Likes
- Retweets
- Comments
- Language

Mining Amazon reviews

Analysis : the methodology

PREPARING DATA

- Removing missing values
- Filtering by language
- Concatenating title and text of reviews (to add more context)
- **TfIdf** vectorization setting **min_df** and **max_df**



LDA

Topic modeling
applied to the dataset

GRIDSEARCH

In order to select best parameters for LDA (perplexity and likelihood)

VISUALIZATION AND INTERPRETATION

- Word cloud
- Most representative reviews for each topic
- Search by word
- Topic distribution
- Rating time series and distribution

Results PS4 - XboxOne

- The majority of samples are **positive** (e.g. “*awesome*”, “*great*”, “*love*”, “*perfect*”)
- **Technical problems** (e.g. broken charger)
- Negative aspects detected only using a high number of topics
- Different languages underline similar topics, an example below:

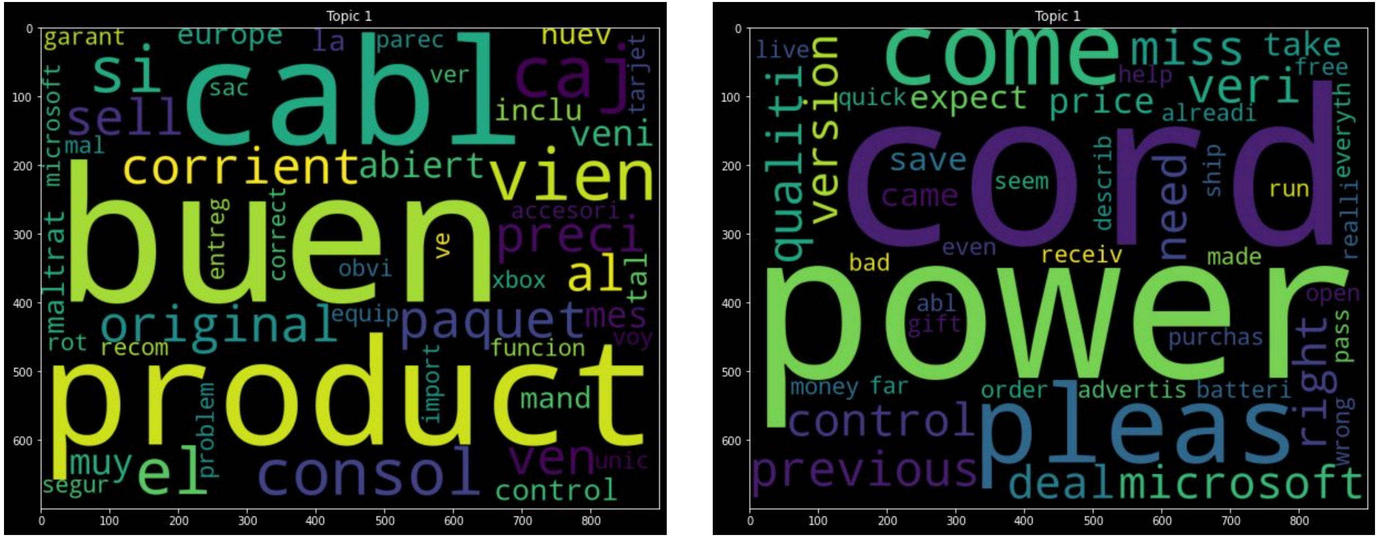


Figure 2: Wordclouds of one of the topics discovered for XboxOne : english and spanish reviews respectively

Results PS5 - XboxSeries

- Main problem detected: **scalpers**
- Despite the majority of reviews are positive, the algorithm was able to detect both positive and negative aspects using just two topics

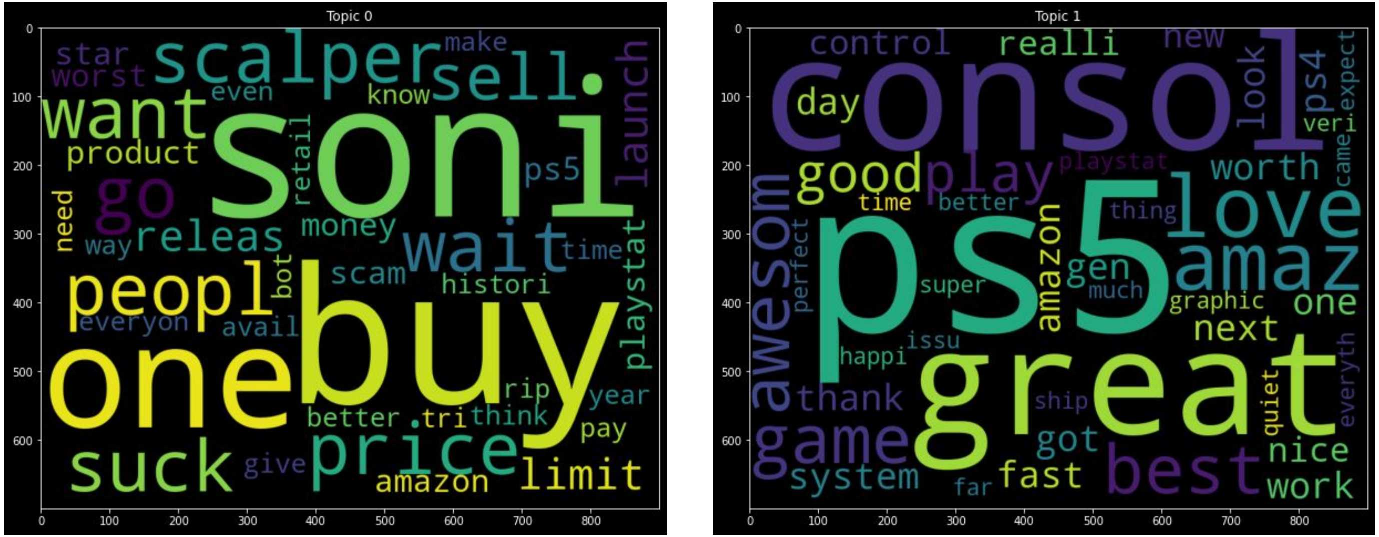
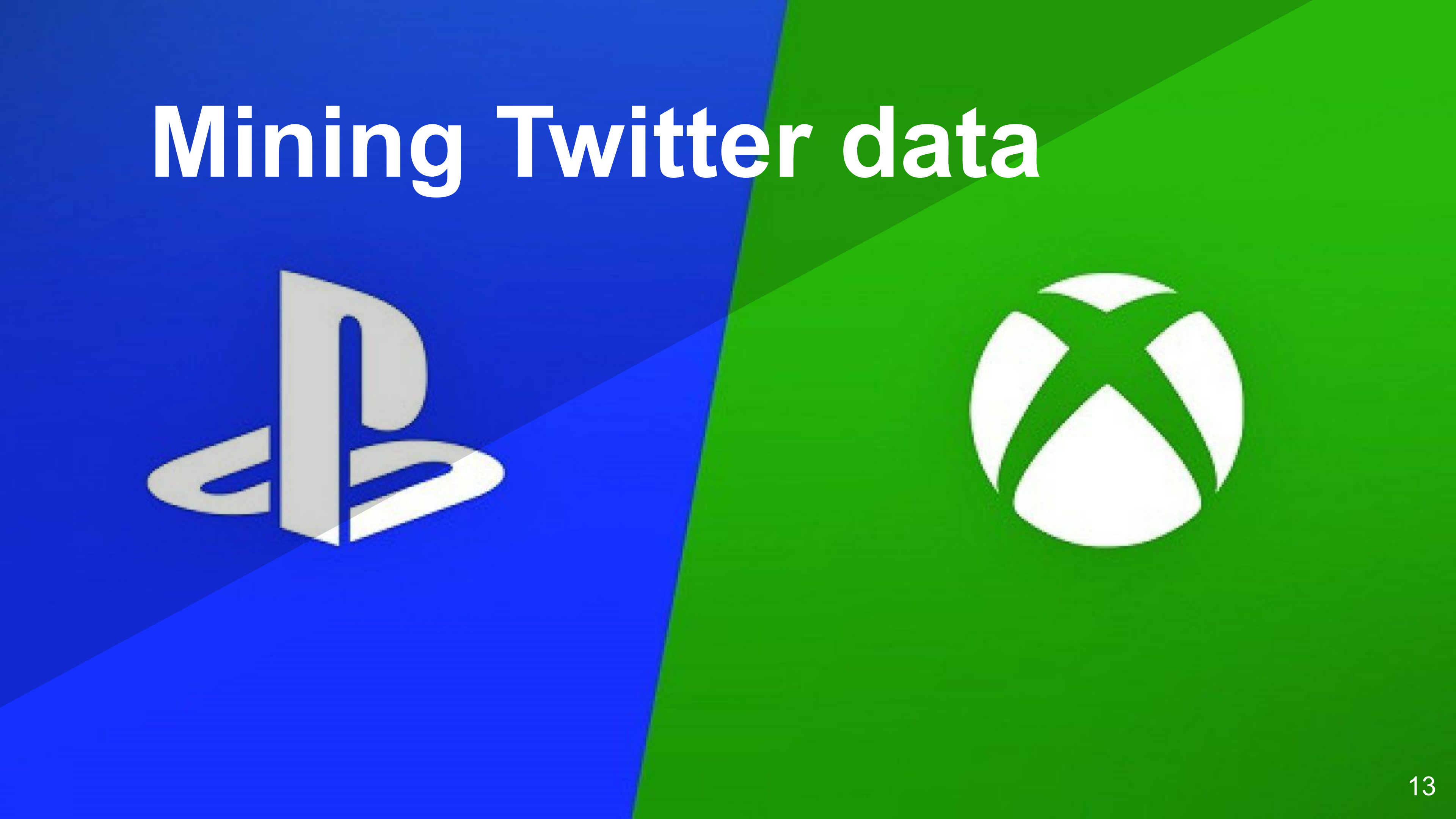


Figure 3: Wordclouds of topic model for PS5 : negative and positive topic respectively.

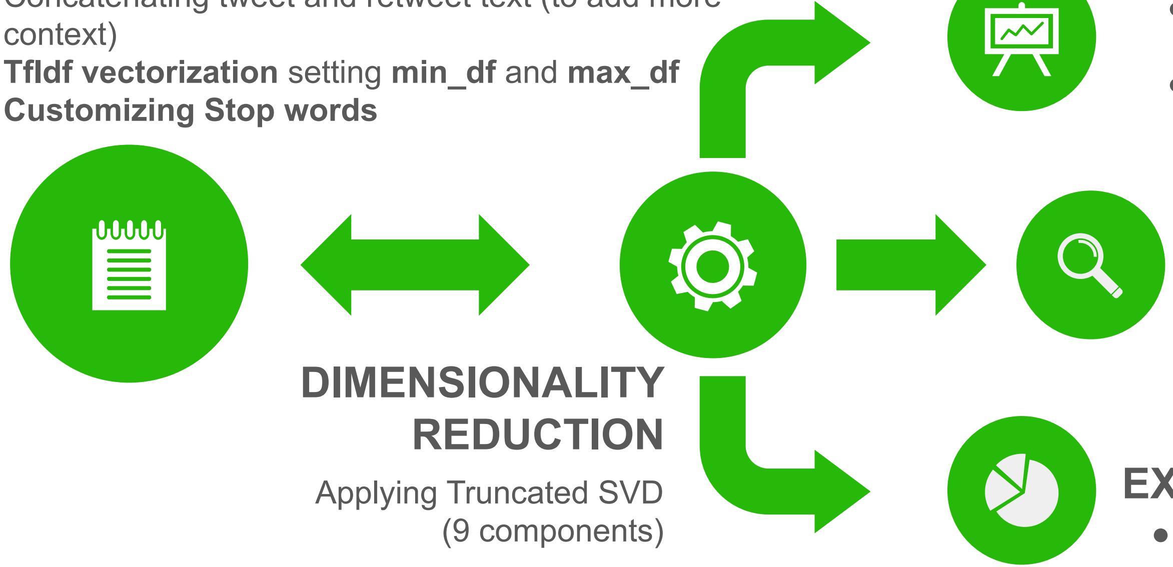
Mining Twitter data



Analysis : the methodology

PREPARING DATA

- Filtering spam
- Filtering by language
- Concatenating tweet and retweet text (to add more context)
- **TfIdf vectorization** setting **min_df** and **max_df**
- **Customizing Stop words**



DIMENSIONALITY REDUCTION

Applying Truncated SVD
(9 components)

ELBOW METHOD

- For K-MEANS : distortion, inertia, silhouette
- For GMM : BIC, AIC

CLUSTERING

- Hard clustering with **K-MEANS**
- Soft clustering with **GMM**
- **Cluster analysis** (wordcloud of centroids, timeseries, countplot, sentiment timeseries)

EXTRACTING LOCAL TOPICS

- **LDA** is applied on each cluster to find subtopics
- Topic analysis (wordcloud)

Results 2017-2018

Independently of the number of topics, we always found:

- a cluster about competitors
- a cluster about most trendy games (ex: Uncharted 4 for Ps4)
- considering just xbox, a cluster about launch day

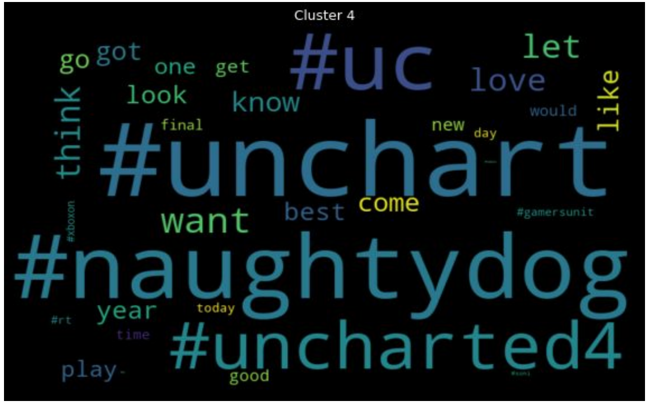
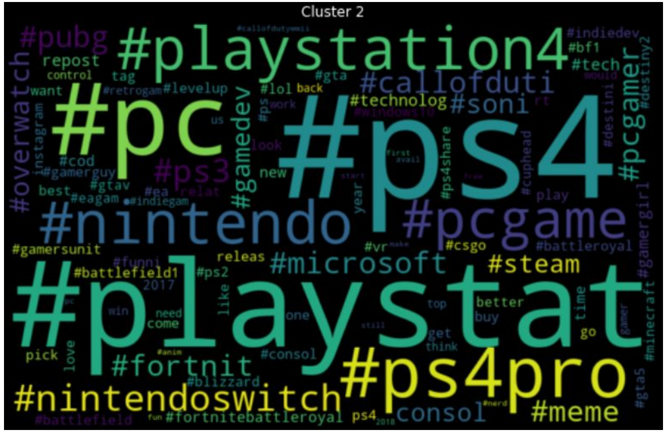
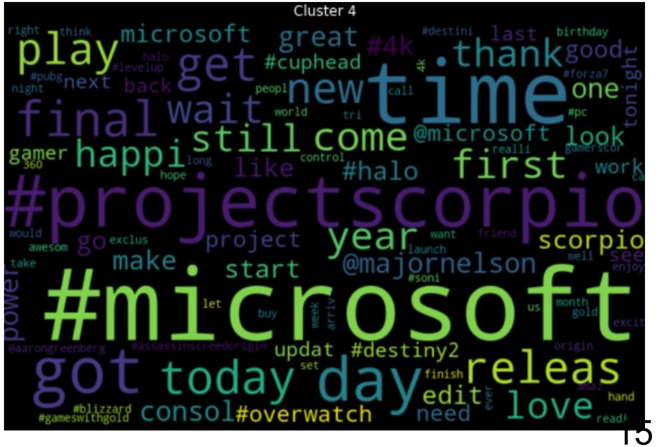

Cluster 4

Figure 4:
From above going to the right: a cluster from Ps4 tweets dedicated to Uncharted 4; a cluster filled with competitor's consoles; another cluster talking about the launch of the new Project scorpio's Xbox limited edition


Cluster 2

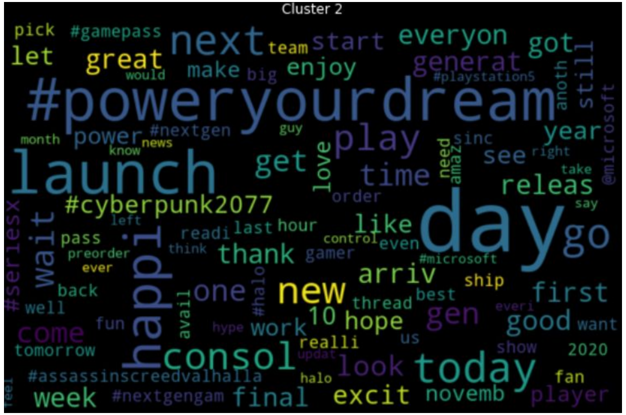

Cluster 4

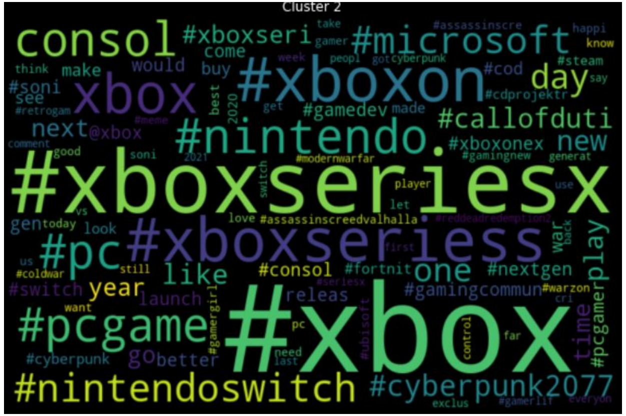
Results 2020-2021

Independently of the number of topics, we always found:

- For both the consoles:
  - a cluster about the competitor
  - an entire cluster about the launch day

- Only for Xbox:
  - a cluster about most trendy games(i.e. Call of Duty)





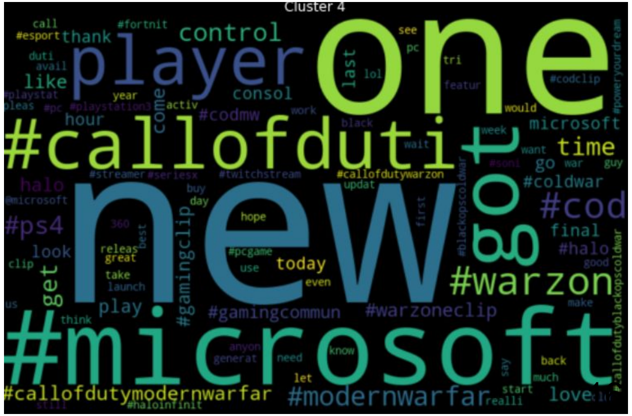
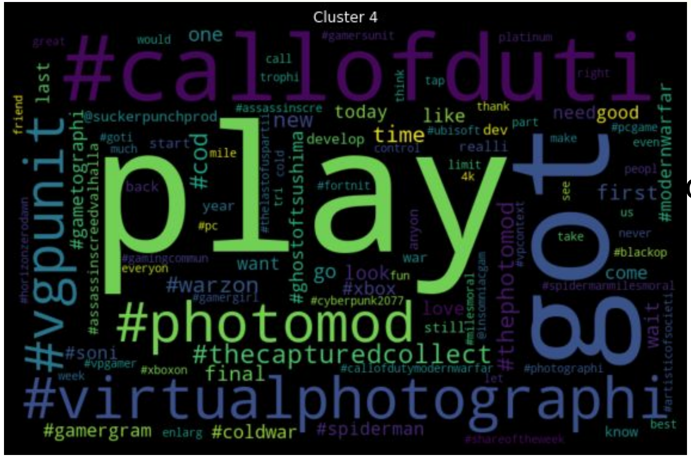


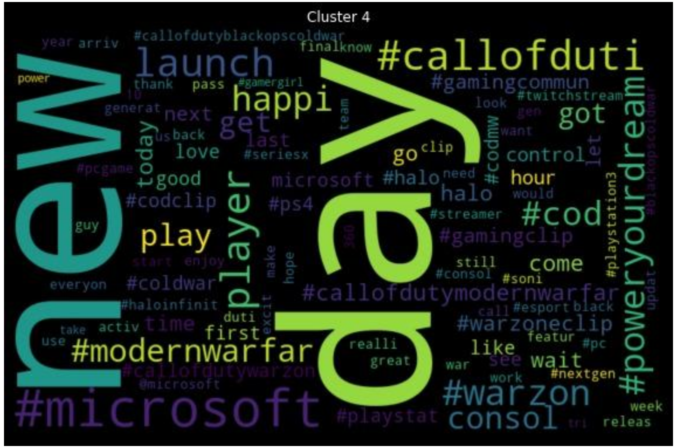
***Figure 5*:**

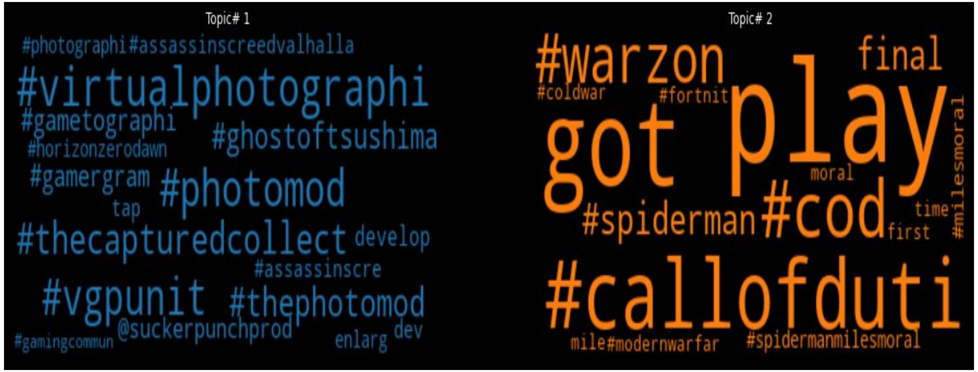
From above going to the right: a cluster from XboxSeries tweets dedicated to the launch day; a cluster filled with competitor's consoles for Ps5; another cluster about Call of Duty's related tweets on Xbox Series.

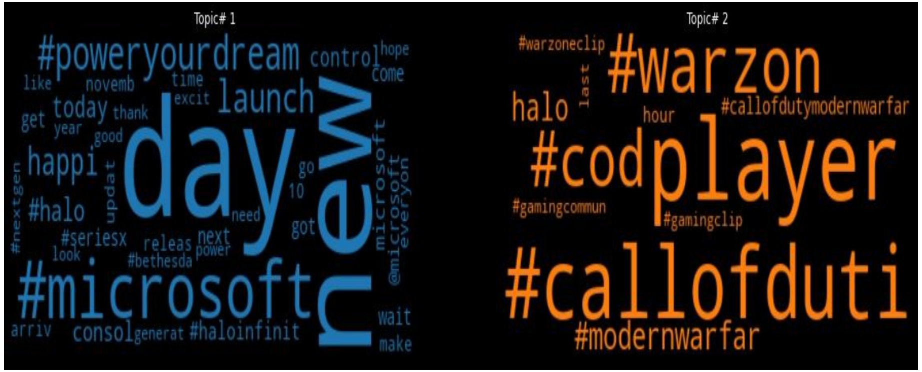
Results 2020-2021



Thanks to LDA we are able to detect sub-topics. For example, on the top are showed one cluster from Playstation and one from Xbox dataset and on the bottom are displayed the two sub-topics resulting from the application of LDA on that cluster.

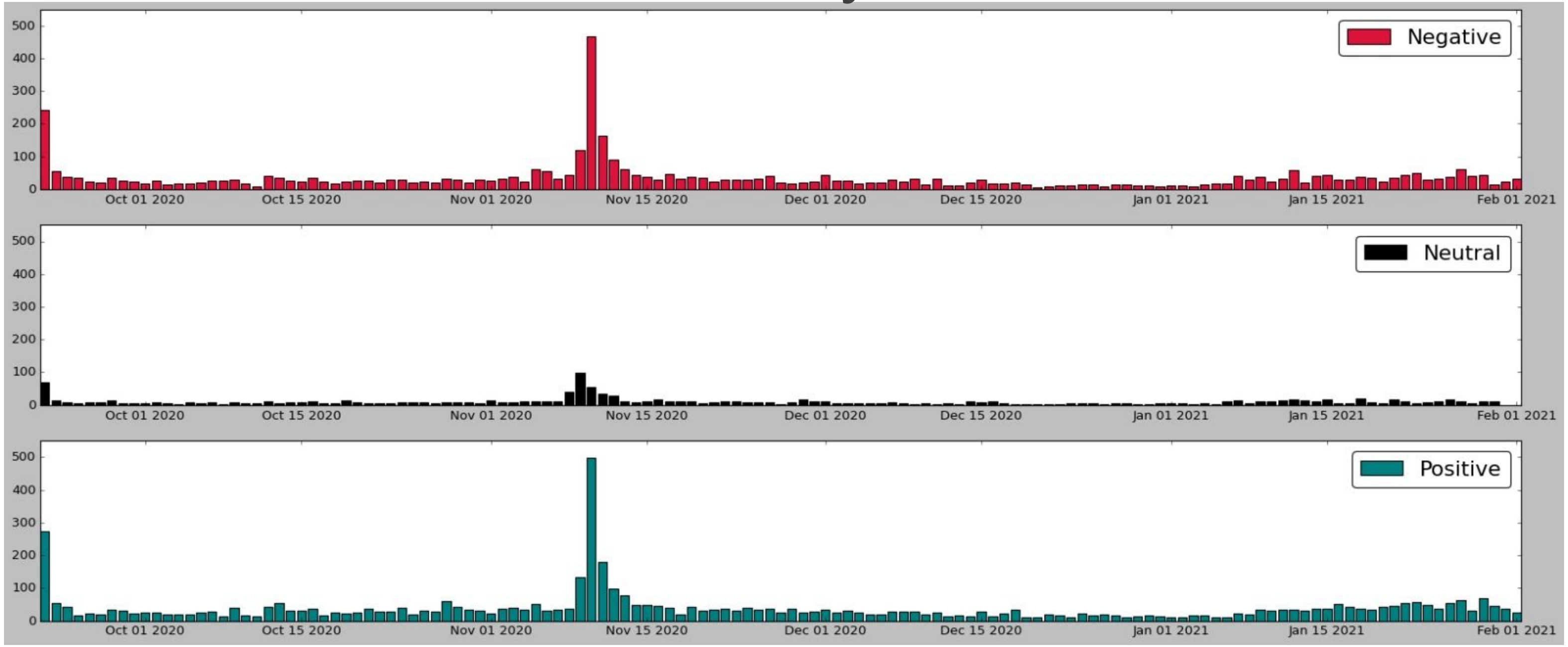






Cluster sentiment 2020-2021

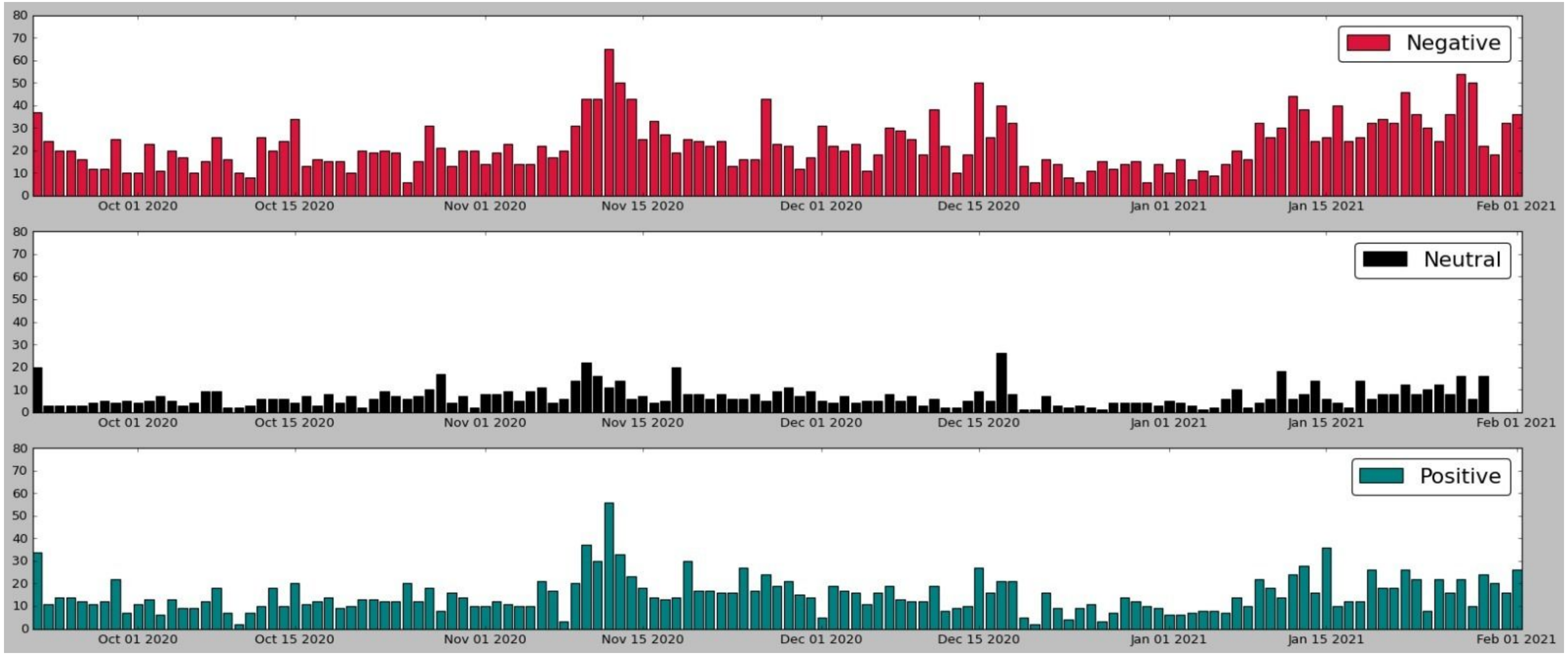
Xbox Series Launch day: cluster sentiment



Analyzing the sentiment of the cluster related to Xbox Series launch day, it’s easy to see that there is a peak of tweets on release date, both positive and negative. Negative tweets regard people that were not able to get the new console, while positive tweets express excitement about the release.

Cluster sentiment 2020-2021

Competitor cluster

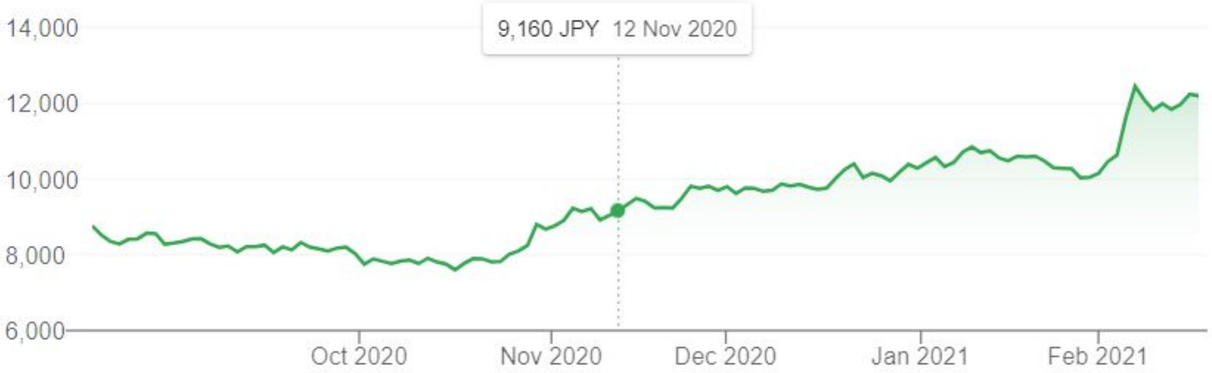


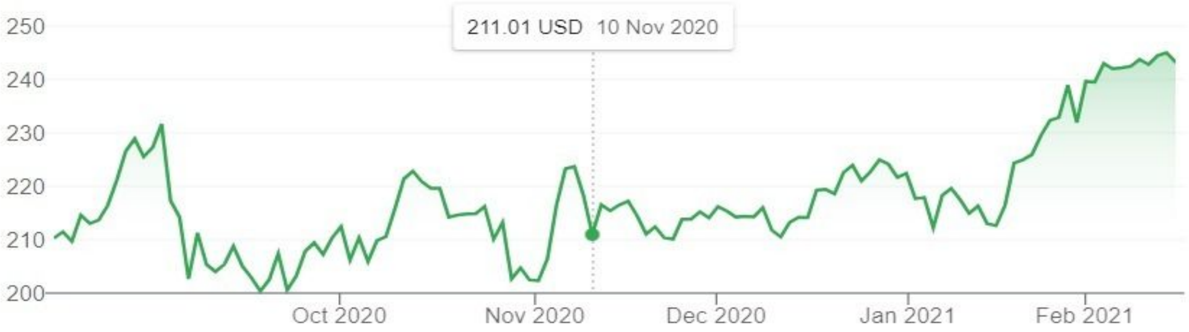
(Empirical proof of how hateful console gamers are between each other!)

Horizontal Analysis

How did the public opinion on the console releases influence the stock price for Sony and Microsoft?

Why same problems, but different behaviors?

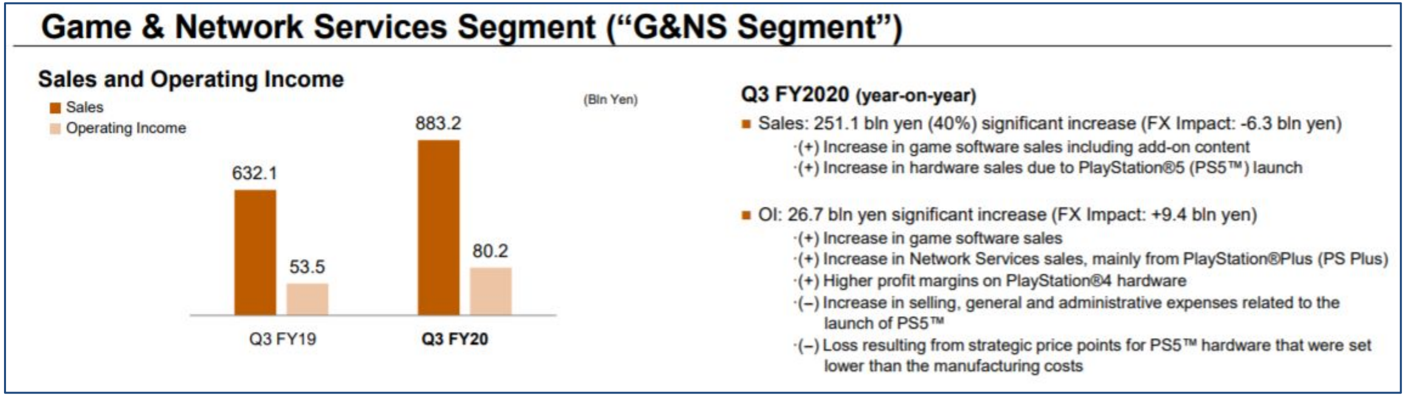




Sony stock market price did not flinch on the week of the release.

Microsoft stock market price **dropped the day of the release**, experiencing a local minima.
The rest of the months, the stock price appeared to be more volatile

Sony’s gaming segment:



Game & Network Services Segment (“G&NS Segment”)

Sales and Operating Income

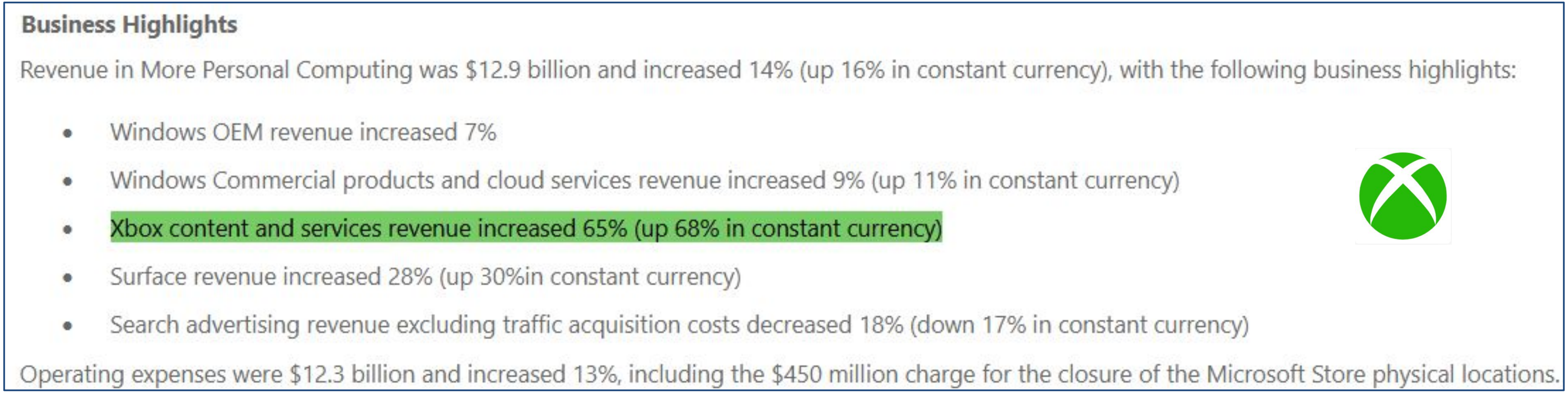
Q3 FY2020 (year-on-year)

- Sales: 251.1 bln yen (40%) significant increase (FX Impact: -6.3 bln yen)
  - ·(+) Increase in game software sales including add-on content
  - ·(+) Increase in hardware sales due to PlayStation®5 (PS5™) launch
- OI: 26.7 bln yen significant increase (FX Impact: +9.4 bln yen)
  - ·(+) Increase in game software sales
  - ·(+) Increase in Network Services sales, mainly from PlayStation®Plus (PS Plus)
  - ·(+) Higher profit margins on PlayStation®4 hardware
  - ·(–) Increase in selling, general and administrative expenses related to the launch of PS5™
  - ·(–) Loss resulting from strategic price points for PS5™ hardware that were set lower than the manufacturing costs

Picture from “ https://www.sony.net/SonyInfo/IR/library/presen/er/archive.html “

To understand how the stocks varies, we need to look at investors relation with the company! Sony revenue during the year of covid, increased of 40% respect to the past FY’s respective quarter

Microsoft’s gaming segment



Business Highlights

Revenue in More Personal Computing was $12.9 billion and increased 14% (up 16% in constant currency), with the following business highlights:

- Windows OEM revenue increased 7%
- Windows Commercial products and cloud services revenue increased 9% (up 11% in constant currency)
- Xbox content and services revenue increased 65% (up 68% in constant currency)
- Surface revenue increased 28% (up 30%in constant currency)
- Search advertising revenue excluding traffic acquisition costs decreased 18% (down 17% in constant currency)

Operating expenses were $12.3 billion and increased 13%, including the $450 million charge for the closure of the Microsoft Store physical locations.

Xbox and cloud-based services (which are the backbone of their gaming service) **increased up** to 65% the revenues respect the passed FY, because of the state-at-home policy.

So what?

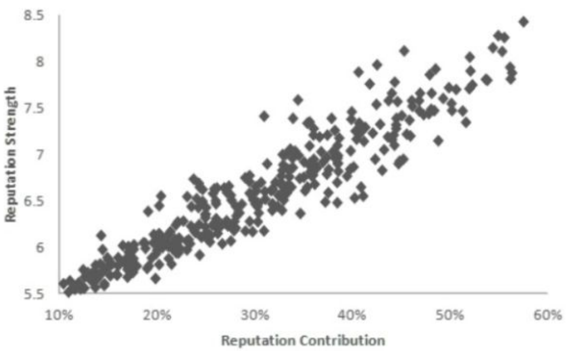
The reason why Microsoft suffered way more the release of their gaming console it’s because it was forecast to have a bigger growth in its gaming segment over time!

On the other hand, Sony had just a minor attention on this aspect, leading to a more steady behavior in the stock market price.

Furthermore, it’s reasonable to think that Sony will have ***less dependance by its reputation*** compared to Microsoft!

Thus, Sony will be an outlier respect to the distribution on the right!

Chart 1: Reputation Strength vs. Contribution – UK & US Companies



Related Works

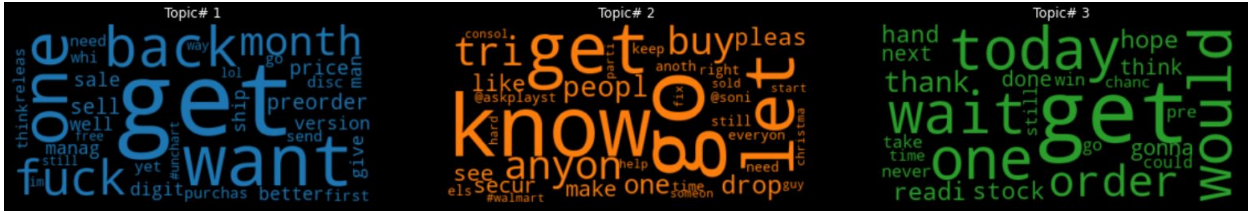
[1] R. A. Wayasti, I. Surjandari and Zulkamain, "***Mining Customer Opinion for Topic Modeling Purpose: Case Study of Ride-Hailing Service Provider***," 2018 6th International Conference on Information and Communication Technology (ICoICT), Bandung, 2018, pp. 305-309, doi: 10.1109/ICoICT.2018.8528751

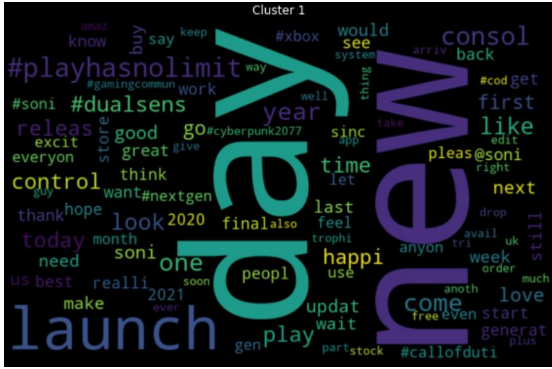
[2] K. Nur'aini, I. Najahaty, L. Hidayati, H. Murfi and S. Nurrohmah, "***Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter***," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2015, pp. 123-128, doi: 10.1109/ICACSIS.2015.7415168

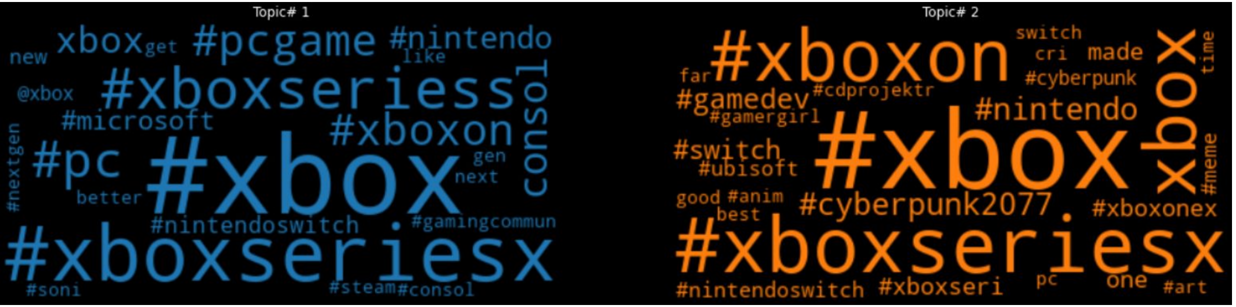
THANKS FOR YOUR ATTENTION

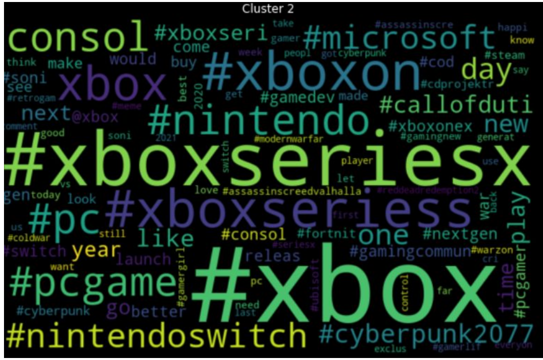
APPENDIX

Some subtopics Ps5 on Twitter

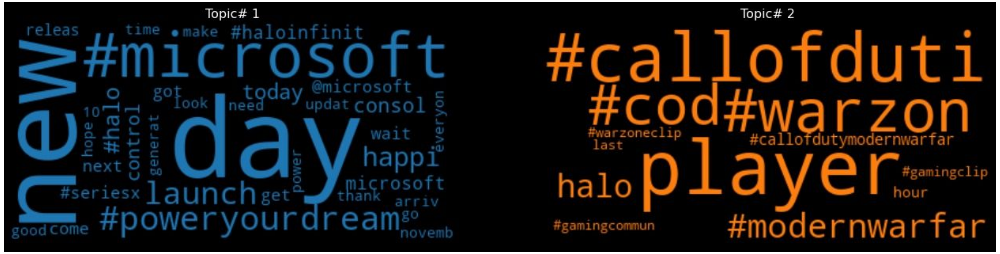


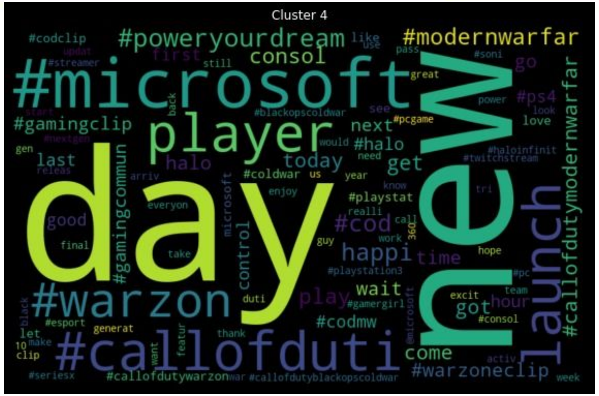




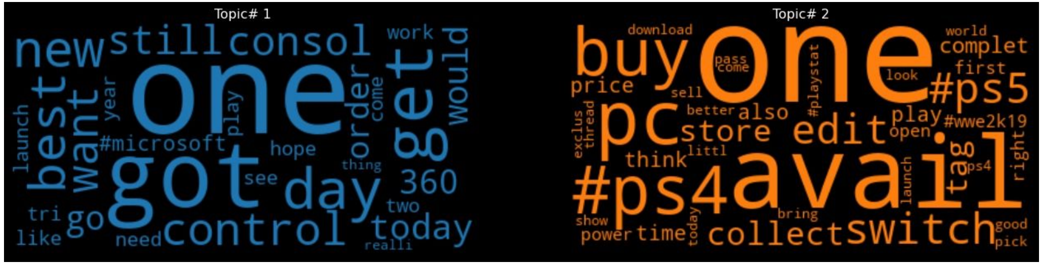


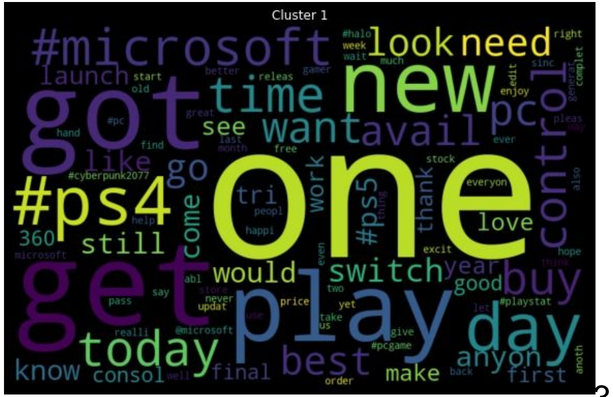
Some subtopics Xbox on Twitter



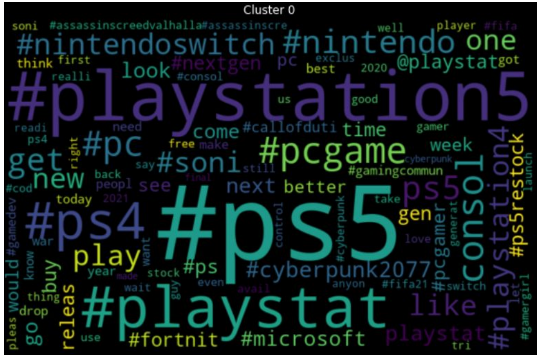


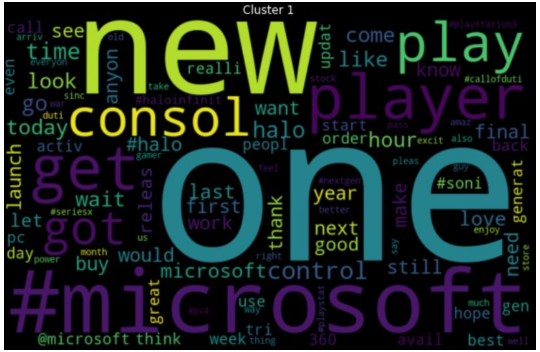
Nice and clean split of the topics that are visible from the macro cluster

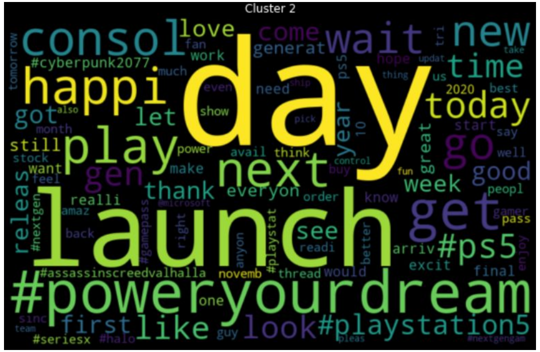


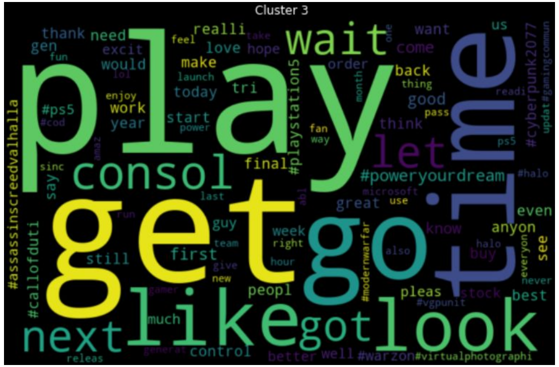


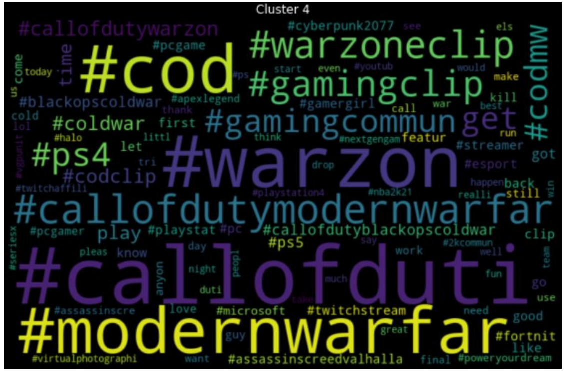
GMM results for Ps5


Cluster 0


Cluster 1


Cluster 2


Cluster 3


Cluster 4

K = 5 for GMM was chosen, mostly empirically.
Overall, it performed in the best case scenario as well as K-means, but averagely was slightly worse in terms of performance.