Network analytics and aspect-based sentiment analysis on Italian Amazon reviews and products related to the video game industry

Data Analytics final project

Claudio Rota, 816050
Simone Paolo Mottadelli, 820786

Contents

1 Introduction 1
2 Dataset description 1
3 Research questions 4
4 Network analytics on Amazon products 5
4.1 RQ1: Which are the most recommended products? . . . 5
4.1.1 Graph description and motivations 5
4.1.2 Methodology 8
4.2 RQ2: Which products can be used to create new retail opportunities? 11
4.2.1 Graph description and motivations 11
4.2.2 Methodology 12
4.3 RQ3: Which are the user preferences with respect to video game genres? 14
4.3.1 Graph description and motivations 14
4.3.2 Methodology 15
5 Aspect-based sentiment analysis on Amazon reviews 19
5.1 RQ4: What do people think about “FIFA 19”? 19
5.1.1 Document corpus and motivations 19
5.1.2 Preprocessing 20
5.1.3 Most common words 21
5.1.4 Dealing with ASUM limitations 23
5.1.5 Sentiword seeds 23
5.1.6 Experimentation with ASUM 25
6 Conclusions 28

1 Introduction

The video game industry is growing so fast that some people believe it will reach over $300 billion by 2025 [1]. With such a profitable market, video game industries invest a lot of money to understand the user preferences, thus creating more and more quality video games.
The main goal of this project was to extract valuable insights about the video game world. To accomplish this task, we used an Italian Amazon dataset containing both products and reviews related to video games of different consoles. We tried to answer 4 research questions, whose answers could be useful for business purposes, by exploiting network analytics and aspect-based sentiment analysis techniques. The first three questions regard the product network, while the fourth one concerns the reviews of the vary famous game "FIFA 19". In this report, we describe the work that we carried out for this project, describing both the methodologies and the motivations behind them.
This report is organized as follows: Section 2 describes the dataset used for this study; Section 3 introduces the four research questions we wanted to answer; Section 4 presents the study conducted using different product networks to answer the first three research questions; Section 5 describes how we managed to extract valuable insights from the "FIFA 19" reviews in order to answer the fourth research question and, finally, Section 6 concludes this report by briefly summarizing the work described in the previous sections.

2 Dataset description

The dataset contains several products extracted from Amazon IT and belonging to different categories, such as beauty, video games and office, with also their reviews. Since these products belong to the Italian Amazon market, most of their descriptions and reviews are written in Italian. Overall, it contains 20460 products and 1988855 reviews.

Each product is provided with the following fields:

- *_id*: the unique identifier of the product;
- *title*: the title of the product;
- *category*: the category of the product;
- *avg_rating*: the average rating of its reviews in stars, from 1 to 5;
- *reviews_number*: the number of reviews on this product;
- *question_number*: the number of questions made on the product;
- *pictures*: the list of links to the product images;
- *description*: the description of the product;
- *features*: the list of characteristics of the product;
- *versions*: the list containing other versions of the same product;
- *bought_together*: the list of products often bought together with this product;
- *also_bought*: the list of products often bought by users who bought this product;
- *also_viewed*: the list of products often viewed by users who viewed this product.

Instead, each review is provided with the following fields:

- *_id*: the unique identifier of the review;
- *product*: the identifier of the product the review refers to;
- *title*: the title of the review;

- *author-id*: the unique identifier of the author of the review;
- *author-name*: the name of the author of the review;
- *date*: the date of the review;
- *rating*: the rating of the review in stars, from 1 to 5;
- *helpful*: the number of users who rated the review useful;
- *verified*: *true* if the user bought the product the review refers to, *false* otherwise;
- *body*: the content of the review.

Since the dataset contains a large variety of products, we made the decision to focus just on the video game category, because we considered it more interesting and fun, thus reducing the dataset dimension to 558 products and 40097 reviews written from December 2010 to April 2019. As shown in

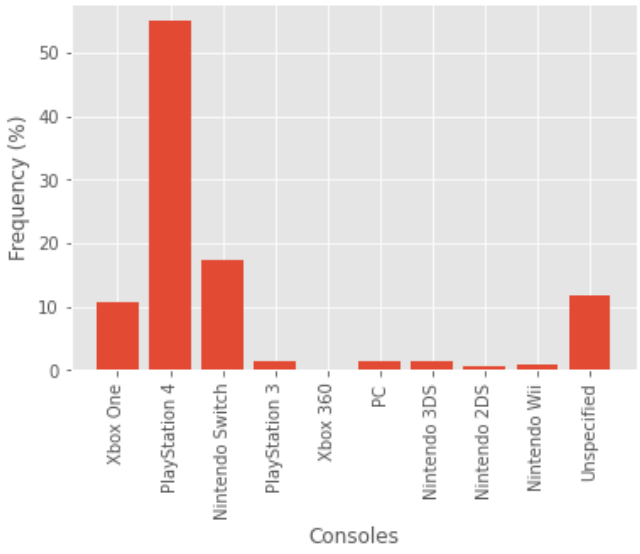


Figure 1: Console distribution considering the video game dataset

Figure 1, the dataset contains products belonging to different consoles and the majority of the products belong to PlayStation 4, Nintendo Switch and Xbox One. As 12% of the products did not specify the relative console, we labeled them as *Unspecified*.
Finally, Figure 2 shows the relationship distribution with respect to their types according to the video game dataset.

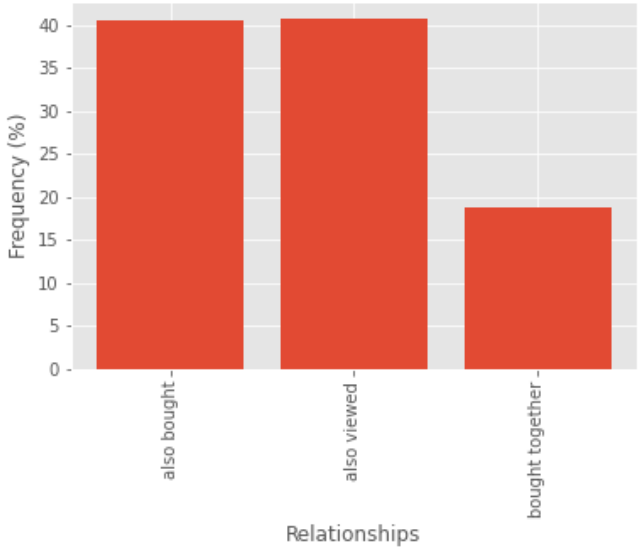


Figure 2: Relationship distribution considering the video game dataset

3 Research questions

The goal of this project was to extract insights that may turn helpful for business purposes. In particular, we wanted to answer the following questions by using network analytics and aspect-based sentiment analysis techniques:

- *RQ1*: Which are the most recommended products?
- *RQ2*: Which products can be used to create new retail opportunities?

- *RQ3*: Which are the user preferences with respect to video game genres?

- *RQ4*: What do people think about "FIFA 19"?

RQ1 can be useful to understand how to sort the products, for example within a website, in order to show users first the ones they are most likely looking for.
RQ2 has the purpose of finding out whether there exist products that are frequently bought together and this could be useful for many reasons, for example, suggesting new products to buy related to the ones that are currently in the cart or for better managing the warehouses.
Providing an answer to *RQ3* could be useful to discover user preferences with respect to video game categories, such as war, football and cars. For instance, discovering some highly correlated video game categories may be of great importance to design targeted advertisements.
In conclusion, *RQ4* aims to discover which are the positive and negative characteristics of the very popular soccer game "FIFA 19" through its Amazon reviews. The answer to this question can be of great interest for the video game company who developed the game and wants to add improvements for the next version. Similarly, a competitive company might take advantage of both the positive and negative aspects of the game to create a new appealing product.

4 Network analytics on Amazon products

4.1 RQ1: Which are the most recommended products?

4.1.1 Graph description and motivations

In order to answer *RQ1*, we modeled the dataset as a graph, where the nodes represented the products and the edges represented the recommendation relationships between products. We extracted the relationships from

the *bought_together*, *also_bought* and *also_viewed* fields of the dataset and we weighted them according to their importance. In particular, the weights of the edges have been assigned as follows:

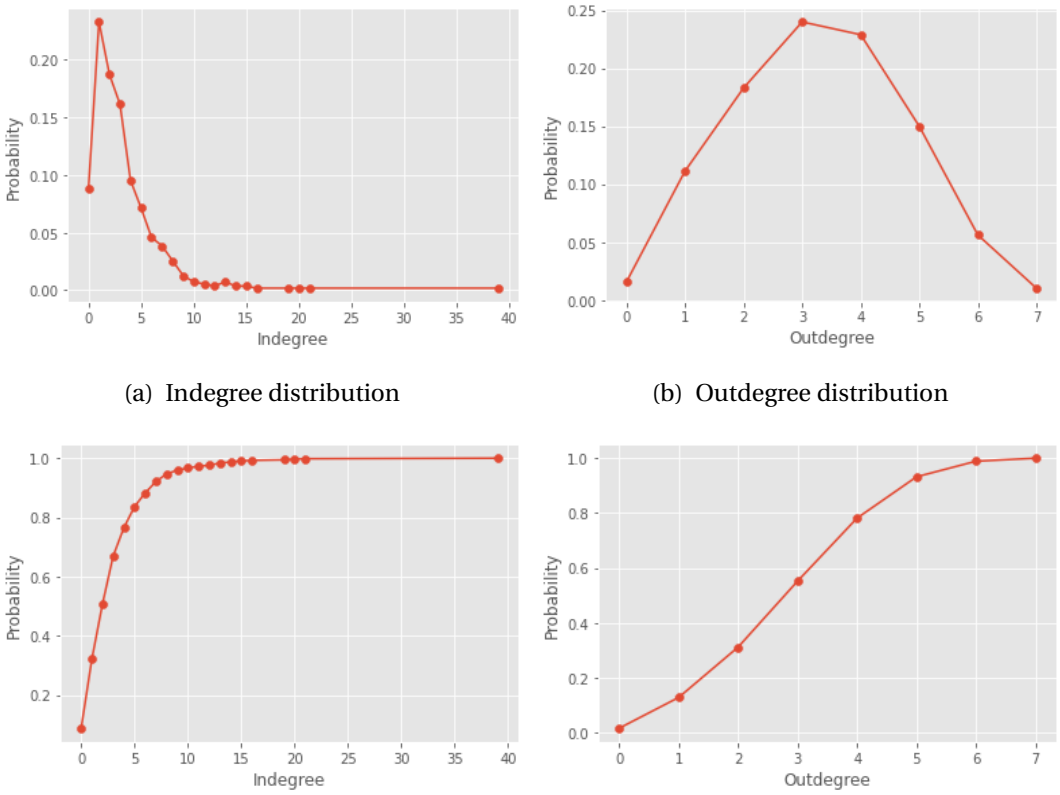
- 0.5 if the products were linked together in a *bought_together* relationship;
- 0.3 if the products were linked together in a *also_bought* relationship;
- 0.2 if the products were linked together in a *also_viewed* relationship.

Since the relationships were not mutually exclusive, that is, a product could be recommended by another product considering more than one relationship (e.g., *also_viewed* and *also_bought*), we weighted the edges connecting two products considering all the relationships between them. Note that the minimum weight for an edge was 0.2, while the maximum weight was 1. The greater the weight of the edge, the stronger the relationship. We chose these weights because, intuitively, two products frequently bought together have a stronger recommendation relationship than two products frequently viewed together.

Nodes	545
Edges	1792
Directed	Yes
Density	0.006
Reciprocity	0.53
Assortativity	0.08
Average degree	3.29
Max Indegree	39
Max Outdegree	7
Diameter	26
Clustering coefficient	0.28
Connected components	4
Giant component dimension	532

Table 1: Statistics of the graph used for *RQ1*

In addition, 35 nodes were isolated from the rest of the network. For this reason, we removed these products from the graph and the resulting network was composed of 545 nodes and 1792 edges.



(a) Indegree distribution

(b) Outdegree distribution

(c) Cumulative indegree distribution

(d) Cumulative outdegree distribution

Figure 3: Degree distributions of the graph used for *RQ1*

Table 1 reports some statistics for this graph. As we can see, the graph was composed of 4 connected components and the giant component included 98% of the nodes. Since each product recommended just a few other products, the graph was very sparse. Nevertheless, the reciprocity was quite high and this means that the majority of the products recommended each other. Moreover, as shown in Figure 3, only about 5% of the nodes had an

indegree higher than 10 and, in particular, we can notice some hubs, where the most important one was “FIFA 19 - PlayStation 4”, which had 39 incoming edges. Instead, considering the outdegree distribution, we can see that the average degree was 3.29 and there were no hubs. This confirms the fact that each product recommended just a few other products.

4.1.2 Methodology

We computed the *Page Rank* centrality for every node using the *Page Rank* algorithm, which is designed to find the most recommended nodes in a network. Note that the weights on the edges have been taken into account because they represented the strength of the recommendation relationship. The top 10 results are shown in Table 2 and takes into account the entire video game network.
In addition, we extracted three subgraphs from the entire video game network by filtering the products considering the related consoles and we computed the *Page Rank* algorithm on each of them. Tables 3, 4 and 5 show the results considering the three main consoles available in the market, that is, PlayStation 4, Xbox One and Nintendo Switch. From the results, it is possible to observe that in all the cases the top 10 products are well known products. In fact, the games about “Super Mario”, “Pokemon” and “Astral Chain” are very famous in the Nintendo world. Also in the case of PlayStation 4 and Xbox One the results are very good, because among the top 10 results we can find “FIFA 19”, “SEKIRO: Shadows Die Twice” and “Devil May Cry 5”, which also won “The Game Award for Best Action Game”. Interestingly, concerning the Xbox One and the Nintendo Switch results, it is possible to notice that not only video games appear among the most recommended products, but we can see the consoles themselves and various accessories.

Rank	Product
1	Mario Kart 8 Deluxe - Nintendo Switch
2	FIFA 19 - PlayStation 4
3	Lego Avengers - PlayStation 4
4	Astral Chain - Nintendo Switch
5	FIFA 19 - Xbox One
6	LEGO Jurassic World - PlayStation 4
7	Pokémon Spada - Nintendo Switch
8	MotoGP 18 - PlayStation 4
9	Dragon Quest XI Echi di un'era perduta -Definitive Edition - Nintendo Switch
10	Lego Star Wars: Il Risveglio della Forza - PlayStation 4

Table 2: The top 10 most recommended products of the entire network

Rank	Product
1	Lego Avengers - PlayStation 4
2	FIFA 19 - PlayStation 4
3	MotoGP 18 - PlayStation 4
4	Devil May Cry 5 - Special Lenticular Edition - PlayStation 4 [Esclusiva Amazon.it]
5	SEKIRO: Shadows Die Twice - PlayStation 4
6	Kingdom Hearts III - PlayStation 4
7	PlayStation 4 Slim 500GB F Chassis, Jet Black + 2° Dualshock 4 [Edizione: EU]
8	Catherine Full Body - Day-One Edition - PlayStation 4
9	F1 2018 Headline Edition - PlayStation 4
10	Lego Marvel Super Heroes 2 - PlayStation 4

Table 3: The top 10 most recommended products for PlayStation 4

Rank	Product
1	FIFA 19 - Xbox One
2	Xbox One - Xbox One S 1 TB, Bianco
3	Xbox One S 1TB Minecraft Creators Pack + 1M GamePass [Bundle]
4	FIFA 18 - Xbox One
5	Xbox One 1708, Controller Bluetooth, Bianco
6	SEKIRO: Shadows Die Twice - Xbox One
7	Metro Exodus Standard - Xbox One
8	Anthem - Xbox One
9	Xbox One: Wireless Controller, Rosso
10	XBOX ONE S 1TB + 2 Controller

Table 4: The top 10 most recommended products for Xbox One

Rank	Product
1	Mario Kart 8 Deluxe - Nintendo Switch
2	Nintendo Switch: Set Da Due Joy-Con, Verde/Rosa Neon - Limited
3	Super Mario Odyssey - Nintendo Switch
4	Nintendo Switch - Blu/Rosso Neon
5	Lioncast Joy-Con Quadrupla Ricarica per Nintendo Switch, Controller per Charging Station con porta USB-C e Indicatore di carica LED - Nero
6	The Legend of Zelda: Breath of the Wild - Nintendo Switch
7	Nintendo Switch,Mario kart 8 Deluxe Volante Gamepad Volante Gioco Volante Supporto Pack di 2
8	Pokémon Spada - Nintendo Switch
9	Astral Chain - Nintendo Switch
10	Set da due Joy-Con Grigi per Nintendo Switch

Table 5: The top 10 most recommended products for Nintendo Switch

4.2 RQ2: Which products can be used to create new retail opportunities?

4.2.1 Graph description and motivations

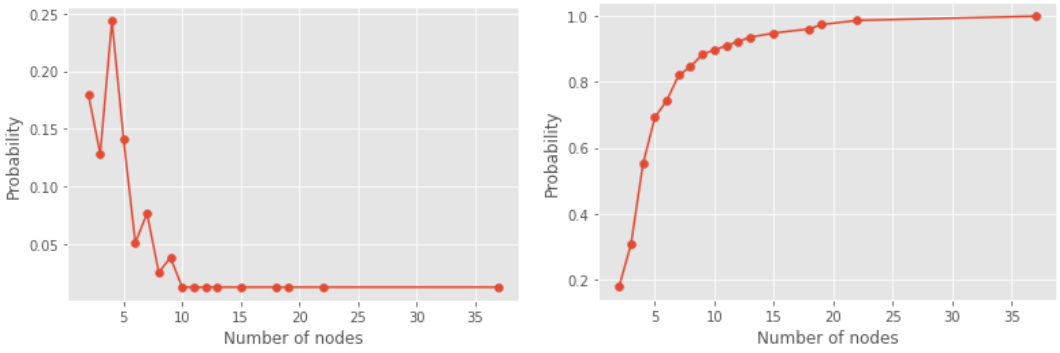
To answer *RQ2*, we used a variant of the graph used in Subsection 4.1. In particular, since we wanted to discover groups of products that recommended each other considering only the purchase relationships, we removed the *also_viewed* relationships, because we considered them useless to answer this question, and we kept only the purchase relationships that resulted reciprocal, thus transforming the graph into undirected. In other words, if there existed the directed edge from X to Y and the directed edge from Y to X, then we connected X and Y with an undirected edge in the new graph. We knew that this process would have not removed all the purchase relationships from the graph used in Subsection 4.1, because it had about 15% of the purchase relationships that were not reciprocal.
This process generated 90 new isolated nodes that we subsequently removed, obtaining a graph with 455 nodes and 476 edges. Such decrease in terms of number of edges is justified by the fact that we discarded the *also_viewed* relationships and the non-reciprocal relationships.

Nodes	455
Edges	476
Directed	No
Density	0.005
Assortativity	-0.07
Average degree	2.09
Max degree	6
Diameter	17
Clustering coefficient	0.19
Connected components	78
Giant component dimension	37

Table 6: Statistics of the graph used for *RQ2*

By analyzing the graph with respect to the statistics reported in Table 6, we

can notice that the average degree decreased from 3.29 to 2.09, while the density remained unchanged. We also observed that the number of connected components considerably increased, creating 78 connected components. We further investigated the number of nodes for each connected component and the results are shown in Figure 4.



(a) Nodes per connected components distribution

(b) Cumulative nodes per connected components distribution

Figure 4: Connected component distribution of the graph used for *RQ2*

4.2.2 Methodology

We used a node centric community detection method based on cliques to answer this question. Since the graph was not so big, this approach did not result computationally expensive. The main idea under this choice is related to the fact that if a customer buys a product X, that is known to be frequently bought together with another product Y, with high probability such customer will buy Y too.
Answering *RQ2*, we tried to simulate the mechanism with which Amazon suggests to add multiple products to the cart with just a click when visiting the page of a specific product. For example, if you want to buy a new smartphone, Amazon suggests you to buy it together with a cover and a screen protector and gives you the possibility to add all of them to your cart.
Despite the graph was composed by several connected components, as

already shown in Figure 4, most of them were composed of more than 3 nodes, which is the minimum number of nodes required to create a clique. This fact should not have negatively impacted the results because we were interested in finding cliques with just 3 or 4 products, as it is unlikely that a customer buys more than 4 video games at once.

Some of the cliques found are reported in Table 7, which shows the most

Product
LEGO Jurassic World - PlayStation 4, Lego Avengers - PlayStation 4, Lego Star Wars: Il Risveglio della Forza - PlayStation 4
PlayStation Camera - PlayStation 4, PlayStation 4 - PlayStation Move Twin Pack, Just Dance 2018 - PlayStation 4
Lego Marvel: Super Heroes 2 - Nintendo Switch, Lego Worlds - Nintendo Switch, Lego City Undercover - Nintendo Switch
Uncharted Collection - Classics - PlayStation 4, Uncharted: L'Eredità Perduta - PlayStation 4, Uncharted: The Nathan Drake Collection - PlayStation 4
Prey - PlayStation 4, Dishonored 2 - PlayStation 4, Dishonored - Definitive Edition
Yakuza Kiwami 2 - PlayStation 4, Yakuza Kiwami - Steelbook Day-one - PlayStation 4, Yakuza 0 - PlayStation 4
Death Stranding - PlayStation 4, Cyberpunk 2077 - PlayStation 4, Ghost of Tsushima - PlayStation 4
Pokémon Spada - Nintendo Switch, Astral Chain - Nintendo Switch, Dragon Quest XI Echi di un'era perduta - Definitive Edition - Nintendo Switch
Assassin's Creed The Ezio Collection - HD Collection - PlayStation 4, Assassin's Creed Syndicate - PlayStation 4, Assassin's Creed Rogue HD - PlayStation 4, Assassin's Creed: Unity - PlayStation 4
EasySMX Controller Joystick per Windows, EasySMX 2.4G Wireless Controller PC, FIFA 19 - PC
Xbox One - Xbox One S 1 TB, Bianco, Xbox One 1708, Controller Bluetooth, Bianco, FIFA 18 - Xbox One
Mario Kart 8 Deluxe - Nintendo Switch, SUPER MARIO PARTY - Nintendo Switch, New Super Mario Bros. U Deluxe - Nintendo Switch
The Legend of Zelda: Breath of the Wild - Nintendo Switch, Nintendo Switch - Blu/Rosso Neon, Super Mario Odyssey - Nintendo Switch

Table 7: The most relevant cliques of products found

significant cliques composed by 3 or 4 products. The first consideration we can do about the results is that for each clique the products belong to a specific console. This result is significant because it would make no sense suggesting products of different consoles, since in the majority of the cases users own only a console. In addition, the cliques suggest different versions of the same game, such as the games about "LEGO" or "Assassin's Creed", and this is also a positive result because this means that people who like a game are likely to buy also its next versions. Finally, there are some interesting cliques containing both video games, consoles and other accessories. For example, the clique containing "Just Dance - PlayStation 4", "PlayStation 4 Camera" and "PlayStation 4 Move" is very interesting because to play at "Just Dance" you must have both the other two products. Hence, it is

very likely that these three products are frequently bought together.

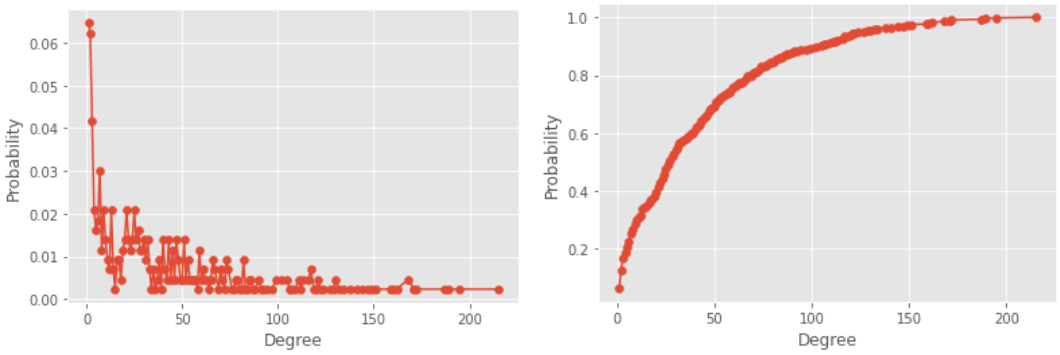
4.3 RQ3: Which are the user preferences with respect to video game genres?

4.3.1 Graph description and motivations

To answer question *RQ3*, we built a new graph because the ones in Subsections 4.1 and 4.2 modeled Amazon purchase recommendations without providing quantitative information about how many times the products were actually bought together. Such quantitative information was essential to answer *RQ3* because just knowing that two products have been purchased together was not as relevant as knowing also how many times this happened. In other terms, this information was necessary to discover new pattern within products able to capture the user preferences.
For this reason, we built a new undirected graph exploiting the data within the reviews in such a way that the nodes represented the products and the edges represented the purchase relationships. For example, if a user reviewed two products X and Y, then we connected X and Y with an edge. We also weighted the edges to represent how many users bought both X and Y, allowing to represent the quantitative information we needed.
Subsequently, we removed the 125 isolated nodes obtaining a graph with 433 nodes and 8825 edges.
Some of the statistics of this graph are reported in Table 8. As we can see, the average degree was much higher than the ones of the graphs used for *RQ1* and *RQ2* and there was just a connected component. In addition, since the clustering coefficient was 0.5, by definition this means that given a product X and two other products Y and Z that have been bought together with X, then Y and Z have been purchased together with a probability equal to 0.5.
Finally, Figure 5 shows the degree distribution of the new graph. The cumulative degree distribution shows that about 90% of the nodes had a degree less than 100 and there were 4 hubs having a degree higher than 180.

Nodes	433
Edges	8825
Directed	No
Density	0.1
Assortativity	-0.11
Average degree	40.76
Max degree	215
Diameter	6
Clustering coefficient	0.5
Connected components	1

Table 8: Statistics for the graph used for *RQ3*



(a) Degree distribution

(b) Cumulative degree distribution

Figure 5: Degree distribution of the graph used for *RQ3*

4.3.2 Methodology

In order to answer *RQ3*, we based on the concept of modularity with the idea that our network was very distant from a randomly generated one. In fact, relationships binding a subset of products together through heavily weighted edges were very unlikely to be originated randomly. In other words, since the relationships between two products modeled real purchases, it was expected that people bought products following a logic, contributing to increase the weights of those edges that connected products

highly correlated (e.g., they belonged to the same console, series or category).
We used a network centric community detection method to find the clusters with the objective of maximizing the modularity score. However, finding the global optimal solution was impractical with our graph, because it was too expensive in terms of computational and spatial complexity. Indeed, even allocating 25 GB of RAM, we did not manage to find the solution. For this reason, we opted to use a heuristic approach to approximate the optimal solution. As a result, we obtained that the number of clusters that maximized the modularity score was 7, achieving a modularity score equal to 0.17. However, we managed to extract interesting insights from the obtained results, even if this score was quite low.
We exploited our domain knowledge to validate the results, assigning to each cluster a representative label:

- *PlayStation 4*: this cluster contains various video game genres belonging to PlayStation 4, because 128 out of 150 products (85%) belong to this console;

- *Nintendo*: this cluster contains various video game genres belonging to Nintendo Switch, because 75 out of 90 products (83%) belong to this console;

- *Xbox + PC*: this cluster contains various video game genres belonging to Xbox One and many accessories for PC. In fact, 42 out of 78 products (54%) belong to Xbox One, while the others are related to PC;

- *Fantasy + Anime*: this cluster contains mainly fantasy games or games that are inspired to Japanese anime. In fact, 22 out of 25 products (88%) belong to these two categories;

- *Driving*: this cluster contains 20 out of 27 (74%) products belonging to the driving category;

- *Kids*: this cluster contains 40 out of 60 products (67%) that are generally for kids;

- *Ubisoft Starlink*: it contains 3 accessories that are related to "Starlink: Battle for Atlas", which is a flight simulation game.

For every node we computed the sum of the weights of its incident edges in order to select the 5 main representative products within each cluster and we reported them in Table 9. By observing the products within every

Cluster Name	**Representative Products**
PlayStation 4	"Crash Bandicoot N. Sane Trilogy - PlayStation 4", "Grand Theft Auto V (GTA V) - PlayStation 4", "FIFA 19 - PlayStation 4", "Uncharted: L'Eredità Perduta - PlayStation 4", "Red Dead Redemption 2 - PlayStation 4"
Nintendo	"Super Mario Odyssey - Nintendo Switch", "The Legend of Zelda: Breath of the Wild - Nintendo Switch", "Mario Kart 8 Deluxe - Nintendo Switch", "Splatoon 2 - Nintendo Switch", "Super Smash Bros Ultimate - Nintendo Switch"
Driving	"Assetto Corsa - PlayStation 4", "Project Cars 2 - Limited - PlayStation 4", "F1 2018 Headline Edition PlayStation 4", "Project Cars 2 - PlayStation 4", "MotoGP 18 - PlayStation 4"
Kids	"Lego Avengers - PlayStation 4", "Astro Bot - Classics - PlayStation 4", "LEGO Jurassic World - PlayStation 4", "Lego Marvel Super Heroes 2 - PlayStation 4", "Rocket League - PlayStation 4"
Xbox One + PC	"FIFA 19 - Xbox One", "FIFA 18 - Xbox One" , "Red Dead Redemption 2 - Xbox One", "Grand Theft Auto V (GTA V) - Xbox One", "Battlefield 1 - Xbox One"
Fantasy + Anime	"Spyro Trilogy Reignited - PlayStation 4", "Kingdom Hearts HD 1.5 + 2.5: ReMIX - PlayStation 4", "Kingdom Hearts HD 2.8 Final Chapter: Prologue - PlayStation 4", "Pokémon Y - Nintendo 3DS", "Pokémon Zaffiro Alpha - Nintendo 3DS"
Ubisoft Starlink	"Ubisoft Starlink Weapon Pack Shockwave + Gauss 3", "Ubisoft Starlink Pilot Pack, Razor 2", "Ubisoft Starlink Pilot Pack, Razor 2"

Table 9: The most representative products for each cluster

cluster, we can affirm that there exist 3 separate video game genres whose games are played by people who rarely play other video game categories. Such genres are:

- Driving games: it includes different sport car and motorbike games as well as many accessories to improve the game experience, such as steering wheels and pedals;

- Fantasy and anime games: it comprehends video games related to both Japanese anime and fantasy world;

- Kid games: it consists of video games usually played by kids, for example, games inspired to LEGO.

Since these clusters were the only ones grouping games by genre, while the other clusters grouped games by console and included various video game categories, it is possible to conclude that, except for the aforementioned genres, people do not have particular preferences, but they generally play games of several categories.
In addition, during the analysis of these clustering results, we managed to extract two other interesting insights that we had not taken into account before starting this study.
More in detail, the first insight concerns the consoles owned by users. It emerged that people tend to own only a console because the clusters labeled as *Nintendo*, *PlayStation 4* and *Xbox One + PC* are related to just a console. The last cluster contains products related to both PC and Xbox One because Xbox One accessories are compatible with Windows PCs, since both are products developed by Microsoft.
Instead, the second insight is related to the reason why people buy products of different consoles, as observed in the cluster labeled as *Xbox One + PC*, which contains a lot of PlayStation 4 accessories, such as headphones and controllers. By analyzing the reviews of the users who bought both Xbox One and PlayStation 4 accessories, we discovered that many of them bought these accessories mainly to play video games on computers and only in rare cases because they owned both the consoles.
In conclusion, a further insight can be extracted by analyzing the degree distribution shown in Figure 5. In fact, we noticed that the three nodes with the highest node degree were “Crash Bandicoot N. Sane Trilogy - PlayStation 4”, “FIFA 19 - PlayStation 4” and “Grand Theft Auto V (GTA V) - PlayStation 4”, which were connected with more than a third of the products of the entire video game network. Practically, this means that these products are bought together with several other products. Based on our domain knowledge, we can affirm that they are video games bought by any kind of users, regardless of their preferences. Indeed, “Crash Bandicoot N. Sane Trilogy”

is a video game loved by users of any age, from kids to adults, because it is one of the most known video games in the history. Although there are users who are not regular players of sport games, “FIFA 19” is one of the most played video games with friends, while “Grand Theft Auto V” includes a lot of modalities that make it suitable for any kind of player.

5 Aspect-based sentiment analysis on Amazon reviews

5.1 RQ4: What do people think about “FIFA 19”?

5.1.1 Document corpus and motivations

This second part of the project aimed to answer *RQ4*, that is, understanding what users think about “FIFA 19” with the objective of discovering both the positive and negative aspects of this game. We accomplished this task by using *ASUM* [2], a state-of-the-art model to perform aspect-based sentiment analysis.
To carry out this study, we focused our attention on all the available reviews for this game, considering all the consoles available in the market. Figure 6 shows the distribution of the reviews with respect to the consoles and, as we can see, “FIFA 19” for PlayStation 4 was the most reviewed product. This means that the aspects of “FIFA 19” potentially extracted by *ASUM* [2] would have been mainly related to PlayStation 4. In total, we extracted 1521 reviews, where 94% of them had the *verified* field equal to true, thus meaning that they were written by people who actually bought the game.
The dataset described in Section 2 shows that every review is composed of multiple fields, but, for this part of the project, we considered only the *title* and the *body* fields. Although the *body* field is more relevant than the *title* field, the latter may contain useful words to understand the sentiment of the considered review. For this reason, we considered a review as the concatenation of its title and its body. In this way, we obtained the document

corpus used for our study.

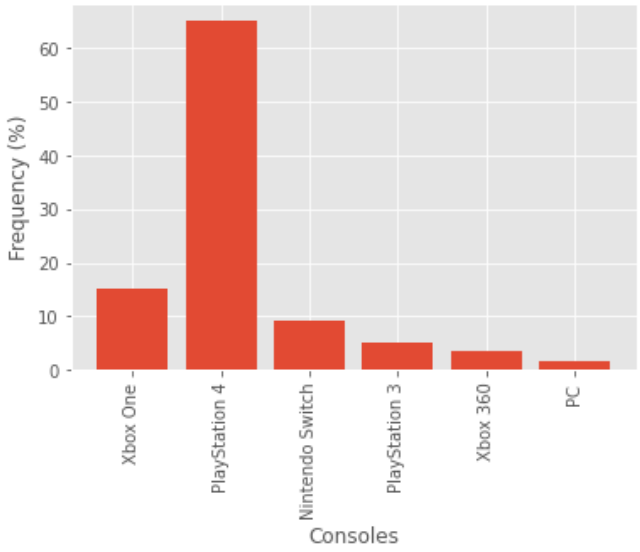


Figure 6: Console distribution with respect to “FIFA 19” reviews

5.1.2 Preprocessing

We preprocessed each review of the document corpus by applying the following steps:

1. *Tokenization*;
2. *Short word removal*;
3. *Negation handling*;
4. *Punctuation removal*;
5. *Number removal*;
6. *Stop word removal*;

7. *Lemmatization*;

8. *Stemming*.

More in detail, with the *short word removal* step, we removed all the tokens with less than 3 characters. Then, we decided to handle the negations in the reviews because they can reverse the sentiment of their sentences. We simply placed the *not_* tag in front of all the words that followed the word "non" up to the first punctuation mark or the first adversative copulative word (e.g., "ma", "però", "invece", "anzi", et cetera).
As regards the *lemmatization* phase, we tried to apply *stemming* without *lemmatization*, but this resulted very inaccurate because Italian verbs have different tenses and, thus, many different tokens related to the same verb were produced. For this reason, we tried to solve this problem by applying *lemmatization* before stemming the tokens, but, also in this case, we observed inaccurate results because some nouns were often confused with verbs (e.g., the noun "prodotto" was confused with the verb "produrre"). Finally, we managed to obtain good results by first lemmatizing just the verbs to standardize them to their infinitive form and then applying stemming to all the tokens. It is worth mentioning that verbs have been recognized using *POS tagging*.

5.1.3 Most common words

Before starting the experimentation with *ASUM* [2], we computed the most common words with respect to positive and negative reviews with the purpose of discovering the most used words to review the game. We divided the reviews in positive and negative considering their stars: we considered a review positive if its number of stars was greater than 3, negative otherwise. We did not take into account the reviews with 3 stars because we considered them neutral. Figure 7 shows the most common words used in the positive reviews, while Figure 8 shows the most common words used in the negative ones. As we can see, positive reviewers generally talk about

the shipment and the graphics of the game, while negative reviewers negatively comment the soccer players and the game modalities, such as Fifa Ultimate Time (FUT), also mentioning the rival game “Pro Evolution Soccer (PES)”.

Figure 7: The most common words used in the positive reviews

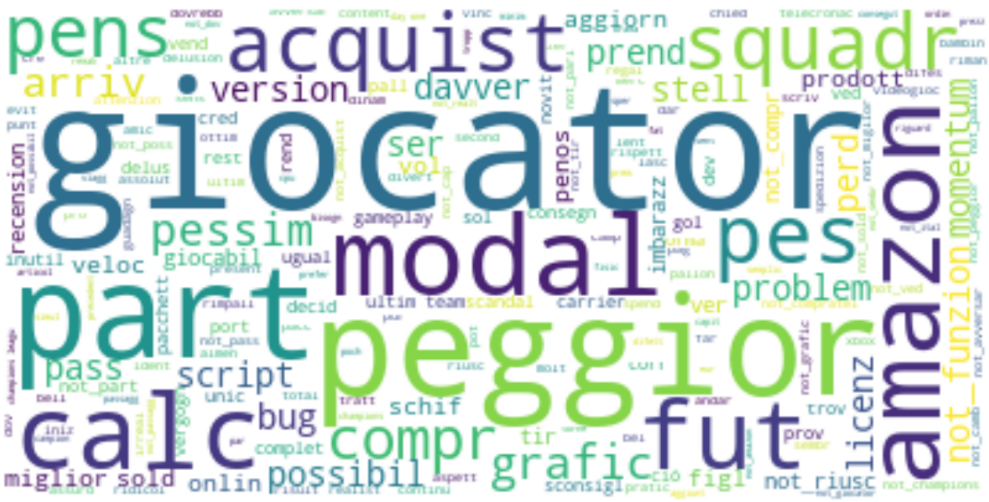
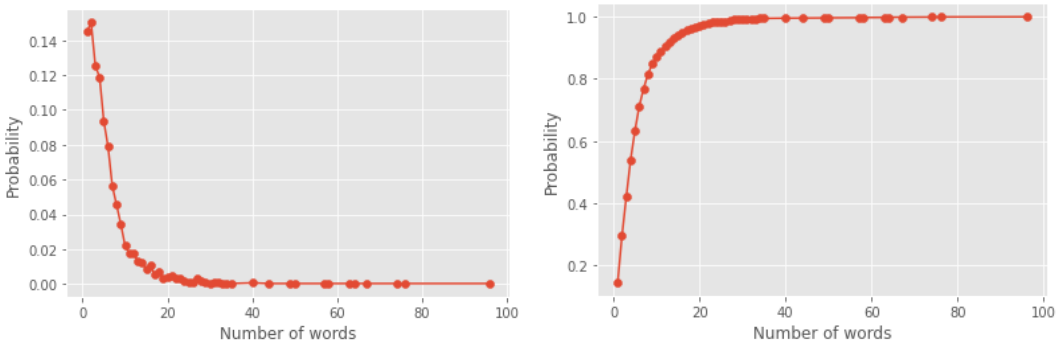


Figure 8: The most common words used in the negative reviews

5.1.4 Dealing with ASUM limitations

According to the authors of *ASUM* [2], the model works better if the sentences within a review are composed of many words because short sentences lack of evidence for both sentiments and aspects. Hence, we investigated the number of words within every sentence and the number of sentences composing the reviews. Figure 9 shows the percentage of sentences having a specific number of words after the preprocessing phase. As we can see, the majority of the sentences composing the reviews were short and this means that *ASUM* [2] could have had problems in discovering what people say about "FIFA 19". By analyzing the reviews at a finer grain level, it emerged that reviews had short sentences because people simply said the game was wonderful or terrible without specifying further details. Even if this fact might have potentially penalized *ASUM* [2], we managed to extract interesting insights anyway, exceeding our expectations.



(a) Relative frequency of the sentences having a certain number of words

(b) Cumulative relative frequency of the sentences having a certain number of words

Figure 9: Relative frequency of the number of words within the sentences

5.1.5 Sentiword seeds

In order to help *ASUM* [2] better distinguish positive words from negative ones, we provided it with two lexicons, the one created by Valerio Basile [3]

and one generated by us.
The lexicon by Valerio Basile [3] contains 74609 words, each of which has a

Positive seed words	**Negative seed words**
“appassionante”, “raccomandare”, “realistico”, “reattivo”, “economico”, “magnifico”, “meraviglioso”, “fantastico”, “eccellente”, “raccomandato”, “bellissimo”, “divertente”, “super”, “figo”, “wow”, “top”, “ottimo”, “miglior”, “buono”, “spettacolare”, “perfetto”, “contento”, “felice”, “felicissimo” “bello”, “bellissimo”, “not_brutto”, “stupendo”, “buonissimo”, “soddisfatto”, “consigliare”	“brutto”, “not_raccomandare”, “not_raccomandato”, “schifo”, “bruttissimo”, “not_soddisfatto”, “restituire”, “meccanico”, “macchinoso”, “amareggiato”, “bug”, “errore”, “ripetitivo”, “costoso”, “not_economico”, “not_realistico”, “caro”, “costosissimo”, “stufa”, “not_consigliare”, “not_bello”, “orribile”, “terribile”, “not_divertente”, “schifoso”, “merda”, “schifezza”, “disgustoso”, “not_contento”, “vomitevole”, “penoso”, “osceno”, “inguardabile”, “pessimo”, “not_buono”, “peggiore”, “imbarazzante”, “deludente”, “vergognoso”

Table 10: The lexicon used for *ASUM*

polarity varying from -1 to 1. Since it has been automatically generated, it is expected to find that the polarity associated with each word is not always accurate. In fact, by manually inspecting some of the words, we noticed some inconsistencies. To mitigate this problem, we decided to filter only the words with polarity equal to -1 or 1 for negative and positive words, respectively. As a result, we obtained a lexicon of about 15000 stemmed words. Even in this case, we noticed some inconsistencies about the polarity associated with the words, but we made the decision to run *ASUM* [2] with this lexicon anyway. However, we obtained poor results because of two problems. The first one regarded the interpretability of the results, because we did not manage to extract the aspects captured by *ASUM* [2],

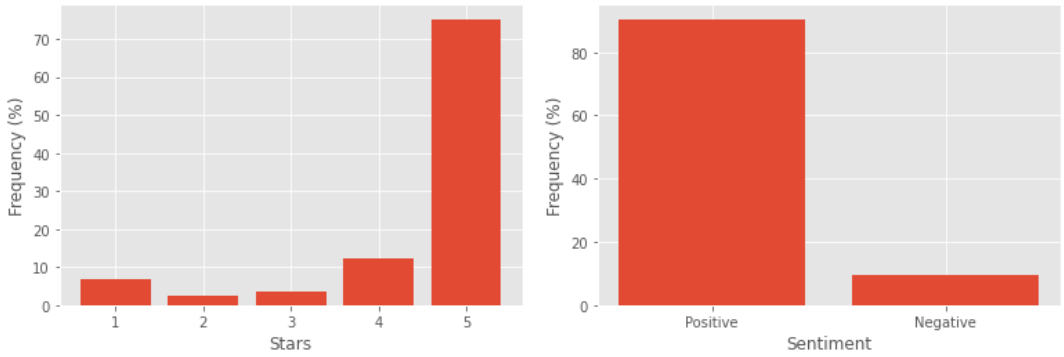
while the second one concerned the fact that positive words were found in topics associated with the negative sentiment and vice versa. For this reason, we decided not to use this lexicon but the one generated by us, which comprehended fewer but more accurate words. Our lexicon is reported in Table 10. As we can see, we used domain independent words that assume the same positive or negative value in every domain. In addition, since we handled the negations using the *not_* tag, as explained in Subsubsection 5.1.2, we also added positive words tagged with *not_* to the negative words and vice versa.

5.1.6 Experimentation with ASUM

ASUM [2] required the following parameters to discover aspects and sentiments within the reviews:

- α: it defines the Dirichlet prior on the document-topic distribution. It is independent from the sentiment;
- β: it defines the Dirichlet prior on the topic-word distribution. It is composed of three values, where the first one refers to non seed words, the second one refers to seed words in the correspondent sentiment and the third one refers to seed words in other sentiments;
- γ: it defines the Dirichlet prior on the sentiment distribution;
- t: the number of topics for each sentiment;
- s: the number of sentiments.

First of all, we decided to use $s = 2$ because we were interested in finding just positive and negative aspects. In order to decide the best value for γ, we estimated the sentiment distribution by analyzing the stars associated with the reviews. Figure 10(a) shows the star distribution of the reviews, while Figure 10(b) shows the percentage of positive and negative reviews without considering the ones with 3 stars.



(a) Star distribution

(b) Sentiment distribution

Figure 10: Sentiment estimation using the stars of the reviews

Since about 90% of the reviews was positive, we set γ equal to 0.9/0.1, where 0.9 is associated with the positive sentiment and 0.1 with the negative sentiment.
Finally, we executed *ASUM* [2] with different combinations of the remaining parameters α, β and t. For each execution, we manually inspected and interpreted the results. We first tried with $t = 5$ and, even if the number of topics was quite small, we managed to extract interesting aspects anyway. Then, we tried to increase the number of topics up to 15, but, in this case, we found a lot of duplicated aspects. We noticed that a good setting was $t = 10$, because it allowed to discover different aspects without many duplicates. In addition, we did not notice important differences in the results to the varying of the values for α and β, as also confirmed by the authors of *ASUM* [2]. It is worth mentioning that we tried running *ASUM* [2] without providing any prior sentiment distribution, but this setting led to obtain a number of negative topics containing positive words.
In conclusion, we reported the most relevant positive and negative aspects in Table 11. As we can see, we isolated 5 negative and 4 positive aspects. More in detail, concerning the true aspects of the game, it emerged that people are generally not satisfied about the servers, the commentary and Fifa Ultimate Team (FUT), which is a game modality. On the contrary, they

Aspect	Sentiment	Related words
server	neg	“server”, “peggior”, “bug”, “penos”, “fut”, “pessim”, “ingioc”, “online”
fut	neg	“schif”, “ultim”, “team”, “bug”, “modal”, “delusion”, “vergogn”
telecronaca	neg	“ripetit”, “imbarazz”, “telecronac”, “peggior”, “italian”, “not_comment”
poste italiane	neg	“pac”, “corr”, “italian”, “post”, “peggior”, “consegn”, “corrier”, “rot”, “disc”
grafica switch	neg	“nint”, “switc”, “version”, “grafic”, “confront”, “rispett”, “altre”
champions league	pos	“modal”, “ottim”, “champions”, “licenz”, “pes”, “rispett”, “leagu”
grafica	pos	“grafic”, “ottim”, “bellissim”, “miglior”, “real”, “realist”
regalo	pos	“regal”, “ottim”, “figl”, “natal”, “nipot”, “compleann”, “entusiast”, “apprezz”
spedizione	pos	“veloc”, “spedizion”, “arriv”, “puntual”, “consegn”, “amazon”, “rapid”

Table 11: The most relevant aspects found with *ASUM*

seem satisfied about the graphics and the introduction of the Champions League modality, as *FIFA 18* did not have it because of licenses.
Instead, focusing the attention on the other aspects, people are very happy about Amazon shipments and very unsatisfied about the delivery service offered by Poste Italiane. Finally, it emerged that this game has been purchased a lot of times to make gifts especially for Christmas and for birthdays.
In conclusion, we have verified the results obtained by *ASUM* [2] by filtering the reviews using the aspects as keywords and by manually inspecting them. By quickly reading the reviews, we can confirm that *ASUM* [2] correctly assigned the sentiment to the aspects in the majority of the cases. Table 12 shows some examples of opinions for some aspects. Unfortunately, we have verified that it was not possible to compare the results using the stars associated with the reviews, because a negative opinion about a specific aspect may appear in a positive review with more than 3 stars. For example, a reviewer can say that the game is wonderful with respect to many aspects and provide 4 stars, but very terrible considering just an aspect (e.g., the Italian commentary).

Aspect	Examples
server	“I server di Fifa sono davvero penosi!”, “I server sono troppo pieni”, “Potrebbero migliorare i server EA”
fut	“Il FUT è scandaloso tra giocatori speciali e non”, “In FUT il 90% delle partite sono pilotate dal gioco”, “la modalità FUT con momentum e bug imbarazzanti”, “nel FUT Champions lagga tantissimo”
telecronaca	“la telecronaca si riferisce all’anno prima...”, “Telecronaca scandalosa”, “Telecronaca italiana imbarazzante e ripetitiva allo stremo”
grafica switch	“La grafica ahimè, mi ha fatto storcere un po’ il naso”, “Su switch la grafica non è delle migliori”, “La grafica ne risente un po’”
grafica	“Solita grafica molto realistica”, “grafica spettacolare”, “grafica bellissima sembra di vedere una partita vera”
champions league	“tra cui la nuovissima licenza della Champions-League”, “La Champions League è senz’altro un grande acquisto”, “Champions league regala a questo titolo il 10 in pagella”

Table 12: Example of opinions related to the aspects found

6 Conclusions

This report presented the work we carried out to extract valuable insights from the video game world using a dataset containing products and reviews from the Italian Amazon market. We answered four research questions by using network analytics and aspect-based sentiment analysis techniques. More in detail, the first answer aimed to discover the most recommended video games in the Amazon network; the second question set the goal of finding groups of products frequently sold together in order to create new retail opportunities; the third question had the purpose of understanding the user preferences with respect to video game categories and, finally, the fourth question aimed to figure out what people think about the football

game “FIFA 19”.
Concerning the first question, we discovered that the most recommended video games on Amazon are very well known and appreciated products. As regards the second question, we discovered that most of the video game groups found were mainly related to the same video game series, genre or console. With the third question, we found out that people usually play at multiple video game genres and they generally own only a console. In addition, there are players who like playing at games belonging to specific genres, such as sport cars. Finally, concerning the last question, we discovered that people are very satisfied about “FIFA 19” with respect to graphics, gameplay and the Champions League modality, but they are not happy about the Italian commentary, Fifa Ultimate Team and the online multiplayer playability because of the low availability and robustness of the EA servers.

Bibliography

[1] Ilker Koksal. Video gaming industry & its revenue shift. https://www.forbes.com/sites/ilkerkoksal/2019/11/08/video-gaming-industry--its-revenue-shift/#73491e07663e, November 2019.

[2] Yohan Jo and Alice Oh. Aspect and sentiment unification model for online review analysis. pages 815–824, 02 2011.

[3] Valerio Basile. Twita. http://valeriobasile.github.io/twita/about.html, 2018.