Politecnico di Torino

Master's degree Course in Computer Engineering

Master's degree Thesis

How have streaming platforms changed the way people consume entertainment and what roles Data science and Word-of-mouth play in making them successful?

An analysis of the Streaming Video-on-demand sector through the market leader Netflix

Student's name: Stefania Angelastri

Thesis Supervisor: Tania Cerquitelli

Thesis Co-supervisor: Luca Cagliero

S252785

Academic year 2019/2020

Abstract

English Version

Streaming platforms have revolutionized the way to consume entertainment giving to the customer freedom in where, when and how to watch and also changing his watching habits, introducing the concept of Binge-watching. This prolonged viewing experience has effects on how we engage with fictional stories. It is not anymore sufficient to sole watch; the majority of the audience extends their experience on social networks, engaging in discussions, commenting and linking related content for several and different reasons: to be part of a community, to have fun or either to be useful in suggesting others.

Therefore, companies can exploit this behavior in two significative ways. They can collect data regarding their customers to create better products and more personalized experiences and, at the same time, they can leverage on the Word-of-mouth phenomenon to increase visibility and awareness about their service, their products and create long term customer relationships.

Considered as experience good, entertainment is affected by word-of-mouth: social ties, reviews and social network discussions have an impact on customer's content perception and also in his decision-making process to what to watch. It is companies' job and more specific marketing job, to be able to arouse customers' interest but above create customer engagement through a continuous interaction since it is proved that engaged customers are loyal customers.

This type of industry is currently booming and sees more and more players competing to attract customers and increase their subscriber base. An analysis of the Streaming Video-on-Demand industry is provided to the reader by observing closely to the leading company Netflix and its strengths and weaknesses.

Index

Table of Figures

This page intentionally left blank.

1. Introduction

The entertainment industry includes different forms of arts and media such as cinema, television, theatre, radio and music.

In the last decade, we have assisted in a shift from traditional publishing to a new business model: innovations and the advancement of information communication technologies have allowed us to shape a new way of delivering content and, as a consequence, also of consuming entertainment.

In April 2020, 58% of the global population result to be active internet users (Statista), this means that more than half of the world can access and consume online forms of entertainment.

Barriers to entry in the entertainment sector have always been very high due to the immense production and distribution costs. However, the advent of the internet has given way to new players who have joined the market, becoming the greatest threat of well-established cinema companies, such as Disney, Warner Bros, 20th Century Fox, as well as the first source of consumption today thanks to the great accessibility that distinguishes them. People have access to any kind of content, any time and from all sorts of locations, without need to drive to theatres.

This disruption brought by the streaming platforms, social media and more in general by technology has reshaped how people consume entertainment, sharpening competition and changing the rules of the game.

If the Video on-the-demand market that has revolutionized the entertainment sector is growing very fast, as we will see in the following sections, it is also true that more and more new players are competing to obtain a greater number of subscribers. Researches will show that in the next few years, the number of users is destined to saturate with an expected increase in average spending for those already registered because each customer will be subscribed to more than one service at the same time.

As streaming has reshaped the entertainment industry, social media has changed the way people approach and watch entertainment, adding a second screen in addition to the one where the show is played.

The influence that social media has on people and even more the influence that other people have on their peers becomes interesting to observe as it is an integral part of the decision-making process of selecting the content to watch.

Media effects studies, indeed, assert that since media changes over time due to innovations and social developments, how people communicate keeps changing and with it also the degree of influence of each medium (Maravelakis, Laroche and Cleveland 2006).

Social media have been disruptive and shaped a total new approach of communicating for both consumers and companies with a high degree of influence with positive and negative effects; deleting physical boundaries and giving the possibility to be always connected with the entire world, the social network has given consumers more power, they express their opinions freely, and they are every day more demanding. Customers are not anymore passive; they are active participants, even co-creators, they talk, share and "engage" with the content and with other viewers: they can comment storylines or character developments, exchange videos and pictures, speculate on what will happen next, discuss on something that just happened, they share their opinions with others. This behavior obliges companies to develop, maintain and enhance long-term relationships through interactions (Harwood, Garry, & Broderick, 2008).

Both streaming platforms and social media services, in addition to being complementary elements of the same customer experience, as we will explain later, are both the result of a technological revolution that finds its foundation in the data and in the immense amount of knowledge that is available today. It is possible to create increasingly personalized services customized for each customer. In fact, we are witnessing a shift from product-centric to customer-centric companies of which streaming platforms are the representative for the entertainment sector.

With these premises, as a case study, we decided to analyze Netflix, the industry leader in most countries where this type of service is provided, to closely observe its strategy and how it can develop in the future.

If Netflix is recognized as a producer of hit TV series such as "Stranger Things" or "La casa de Papel", data shows that, first, most subscribers watch more licensed content, and two, how original content is not sufficient to cover production costs. Besides, many production houses, which sold the reproduction rights to Netflix, but who now are launching their streaming platforms thus claiming their films and series back, drastically reducing the selection in Netflix's catalog is a serious threat; The originals, therefore, are a product that is useful in the acquisition phase, because they are content, not available elsewhere, however they are expensive and not sufficient to keep customers subscribed to the service.

To understand the next steps to take to remain the industry leader, an analysis of Netflix's strategy regarding content was then conducted through data analysis and natural language processing techniques. In addition, a survey was conducted in order to analyze the consumer's opinion regarding this type of service and understand if external agents can influence his decision-making process.

The document is therefore structured as follows: the first section provides a simple explanation of the data analysis process, chapter four, and text mining techniques, chapter five; In chapter six, some application examples with a focus on business and marketing are introduced. In chapter seven, a review of researches conducted on the word-of-mouth phenomenon and the impact this has had on marketing is proposed to the reader; To conclude the review of the literature with chapter eight with an analysis of how the entertainment has changed with the rise of streaming and binge-watching, the role of marketing in this industry and the antecedents of the WOM related to entertainment. Chapters nine and ten introduce the problem this research will try to answer, and the methodology followed by findings illustrated in chapter eleven. Chapter twelve provides recommendations and a conclusion, proposing further developments for future research.

2. Literature review

2.1. Data science: an introduction

Nowadays, companies collect millions of data every day, which is transformed into knowledge: this process is called “data analysis”. When we talk about data analysis, we refer to the collection, organization, and structuring of data to be analyzed and used in decision-making processes.

Several domains such as medicine, genetics, informatics, education, business, and many others (Petre, 2013) now use data science techniques to find solutions and discover unexpected results that previously were simply not imaginable: they can be structured or unstructured, behavioral, attitudinal, descriptive, sound or graphic data, and all of these would be merely unmanageable and useless without the adoption of techniques that allow their organization and manipulation (K. Mishra, Hazra et al. 2016).

KDD - Knowledge Discovery in Database is the process that describes the different stages from the collection of information to the extraction of a more in-depth knowledge extracted through

data analysis (Maimon, Rokach 2010). It is a sequence of operations to be followed when deciding to start exporting data.

As described by Maimon and Rokach in their book "*Data Mining and Knowledge Discovery Handbook*" (2010), this process consists of nine phases (Figure 1): 1. Understanding the domain of application and defining its objectives; 2. Collection of available data; 3. Data pre-processing: this phase is fundamental to perform a correct analysis and consists of removing noise and outliers, handling missing values , and cleaning; 4. Data transformation: this stage varies from project to project but is a necessary step in making the data manageable and includes techniques for reducing dimensionality, aggregation, or sampling; 5 & 6. Data mining task & algorithm's choice: this depends on the objective that we want to achieve and on the type of data available; 7. Data processing: application of the algorithm to the data; 8. Evaluation: obtained results are evaluated and interpreted with those expected, taking into account the techniques adopted. 9. Use of the extracted knowledge for the set objectives. (Maimon, Rokach 2010).

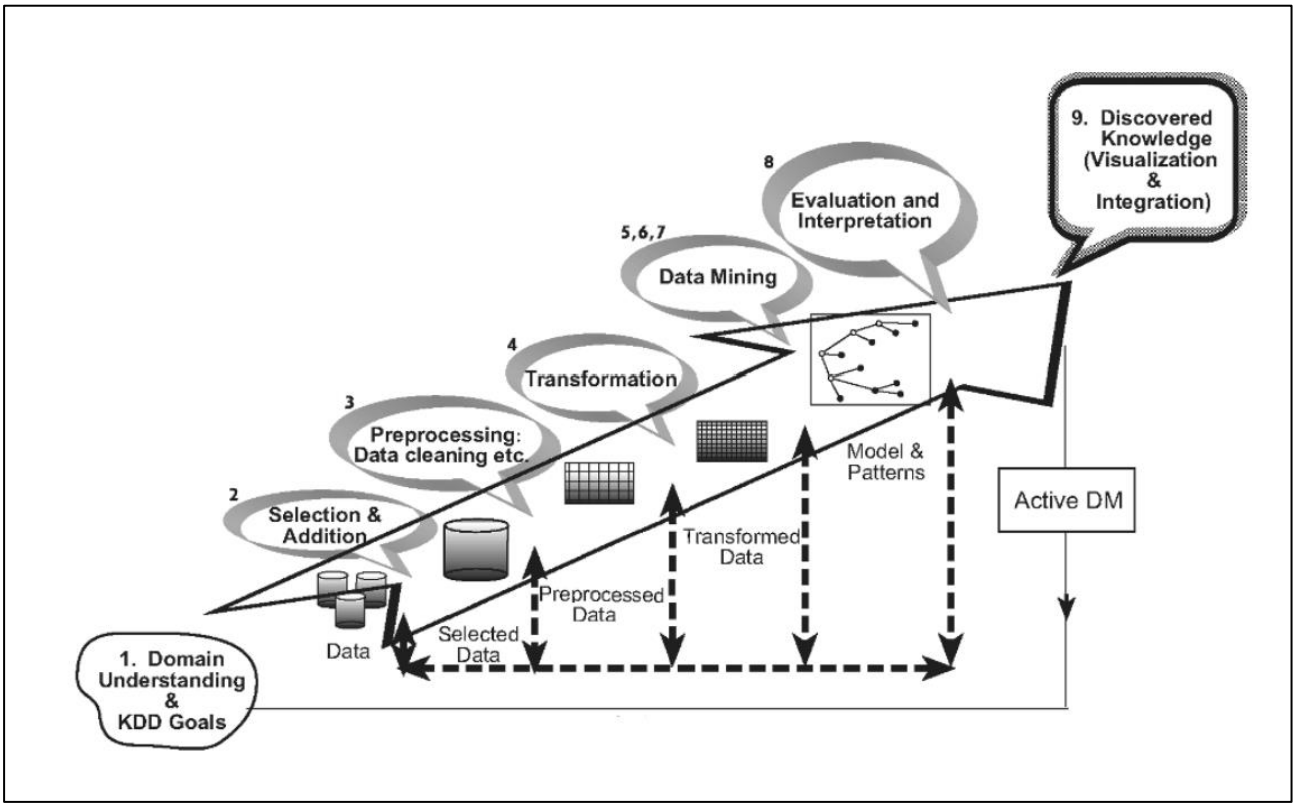


Figure 1: Knowledge Discovery in Database
Source: Oded Maimon, Lior Rokach Data Mining and Knowledge Discovery Handbook, 2nd ed, 2010

2.1.1. Pre-processing techniques

In the first four phases, pre-processing techniques are adopted to make the data suitable for the analysis; ignoring these preliminary stages can bring to unsatisfying results. Advanced algorithms would not be able to retrieve significant results, or it would require very long times at high computational costs without using pre-processing techniques.

The size of the dataset, a collection of data, has a significant impact on the analysis and on the choice of the techniques to be adopted. In the early stages of exploration and analysis, it is always better to work on a small set of data to verify the results obtained. In the second stage, the accuracy of the techniques is evaluated to make them scalable on a more extensive set of data. There are several data reduction techniques, including feature selection or sampling. In the first, we only consider those attributes, columns of the datasets, which are relevant for the analysis, with sampling, instead, we try to create a smaller dataset that maintains the same distribution of the original dataset and therefore, significantly representative even smaller. There are several types of sampling: simply random, systematic, clustered and stratified sampling. These methods, the so-called probability sampling, take into consideration the entire group recreating a smaller version with the same characteristics, without deleting features so that it will make more straightforward conduct more generalized research.

In simple random sampling, each individual in the sample can be selected with the same probability: the method is quick and reduces the hypothetical bias that can be included using other methods; however, exists the risk of not obtaining a sample big enough with the characteristics that we want to analyze. Systematic sampling in which an individual is extracted at regular intervals is even simpler than simple random sampling. The structure of the sample conditions this method, and it can introduce bias as well as lead to groups of samples that are not relevant for research. Clustered sampling consists of dividing the entire dataset into smaller clusters, groups of elements with the same characteristics, and then select whole clusters; it is up to the analyst to then decide whether to proceed to take all the elements inside that cluster or apply furtherly a random sampling. The stratified sampling, instead, the dataset is divided into sub-groups called “strata”, internally homogenous; From each of these, then individuals are picked so that the general sample is representative of the entire dataset.

When the dataset has been reduced to make it easy to handle and analyze at reduced costs, the next stage includes data cleaning techniques.

The scope of this is to reduce noise and remove outliers from the dataset. An outlier is defined as an extreme value that differs from all the others. It can represent either an error that was made in measuring or collecting or the accurate representation of an event, which is, however, considered rare and, therefore, not representative of the average. The removal of the outliers depends on the knowledge domain and the type of analysis to be carried out.

The last step of the KDD process before starting with data mining is the transformation of data from one format to another in case the original shape is challenging to analyze, in order to use a specific model or for visualization and representation purposes.

Stages from one to four are common to all data analysis processes, data mining techniques have been adopted in order to obtain the expected result.

2.1.2. Data mining methods

Mishra, Hazra et al. have defined data mining as "the process of knowledge discovery in a database which can be used in decision making" (Mishra, Hazra et al. 2016) whereas Petre as "a dynamic and fast-expanding field, that applies advanced data analysis techniques, from statistics, machine learning, database systems or artificial intelligence, in order to discover relevant patterns, trends and relations contained within the data, information impossible to observe using other techniques" (Petre, 2013).

As mentioned above, data mining is a complex process that unites different disciplines in order to reveal and show the knowledge that is hidden under thousands of lines of data. Two main categories group different methods of approach: predictive methods and descriptive methods. The former aims to build behavioral models to predict future values of variables, whereas the descriptive ones have the aim of interpretation, the discovery of hidden correlations, and manipulation in order to make them more understandable.

Regression, Classification and Clustering are among the most used techniques by researchers or companies.

Regression is a technique used primarily in the context of prediction or to describe causal relationships between dependent and independent variables. In the first case given a set of data characterized by a target attribute and a numeric target attribute, through a mathematical function, it can predict the target value of new objects; In the second it explains the effect of the dependent variable on the independent variables (Allison, 1999). The method assumes that

what has occurred in the past will repeat itself in the future and therefore exploits collections of historical data to understand the relationships between variables and reproduce them on new data sets (Nugus, 2009). This technique finds several applications: it can be used to analyze relationships between the point of sales and purchases, number of complaints and wait times of callers; it can improve performances identifying areas with maximum impact, estimating costs of houses, forecast stock prices and many others.

Classification is one of the most used data mining techniques. Through this method, it is possible to predict class labels or create interpretable models of a given phenomenon. Given a collection of labeled data, which has already been previously classified with the correct label, the method can create a description of the class to be applied to unlabelled data. The first set used to train the method is called the Training set; A Test set is a set of unlabelled data that is used to test the accuracy of the model and classification rules. Classification techniques include decision trees, association rules, Neural networks, SVM[1], Naïve Bayesian networks or K-NN[2].

Clustering aims to create a set of clusters. This technique finds similar objects and groups them together, distancing them from those that are different and, therefore, will be part of another group. Clustering finds applications where we want to understand the similarities between entities in different application fields, or it is also used as a size reduction technique. The algorithms used for clustering are K-means and its variants, Hierarchical clustering and density-based clustering.

2.2. Text mining: text processing and techniques

Text mining tries to explore techniques that allow computers to understand and manipulate texts.

In the KDT (Knowledge Discovery in Textual databases – Figure 2), data mining techniques are applied to unstructured or semi-structured data such as texts (Vijayarani, Ilamathi et al., 2015). Structured data resides in row-column databases; instead, unstructured data has not a proper organization and can contain different types of information such as messages, dates, tags…

[1] Support Vector Machine
[2] K-Nearest Neighbours

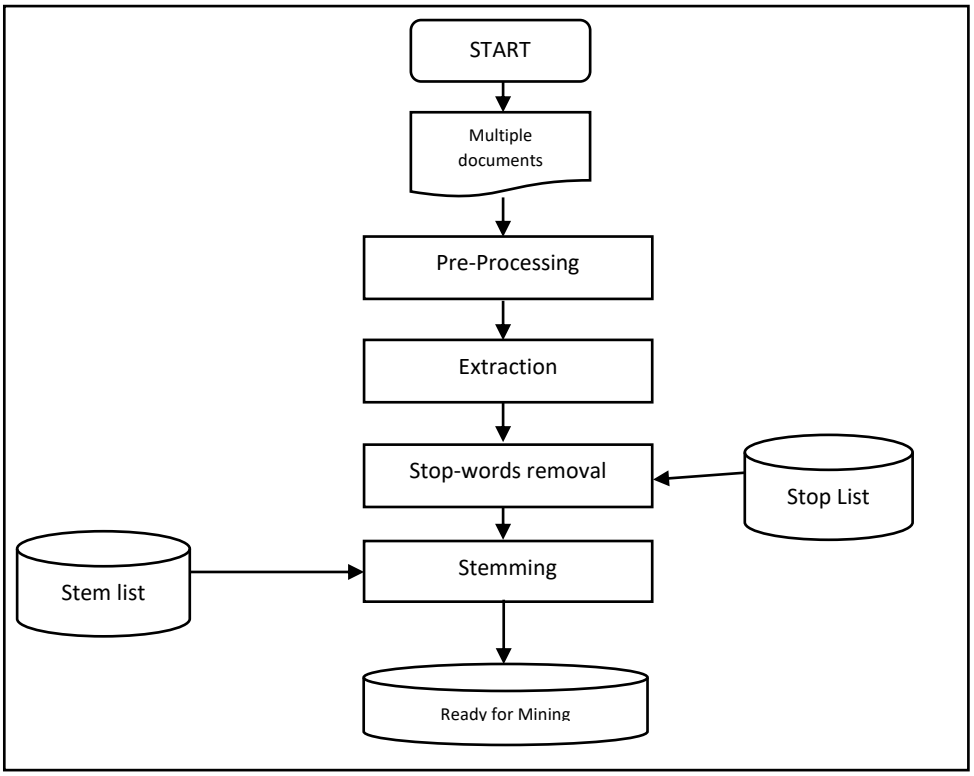


Figure 2: KDT - Knowledge Discovery in Textual Databases

2.2.1. Text pre-processing

In this section, the most adopted techniques to pre-process textual data are reviewed as an introduction to the procedure that will be performed.

Tokenize

Tokenization is a technique used to split texts into smaller parts, called tokens. Long texts can be divided into sentences and sentences into words. Different strategies to tokenize have been created during these past years, and they can be applied depending on the purpose, which kind of structure we want to achieve and the type of information we want to keep. The most common tokenizer splits sentences into words where it finds punctuation or blank spaces; it is possible to split also only considering blank spaces to keep words bound by apostrophe or score altogether. Other versions of the tokenizer divide texts reading a provided list of sub sentences. All these methodologies are available into the NLKT library of python created on purpose for text analysis.

Tags, punctuation and special characters

The first step into the pre-processing is to consider which elements are meaningful for the scope of the research. In an approach all-inclusive, all text would be processed; however, it would require much memory, being costly in time and resources. For this reason, it is quite usual to start by removing mark-ups such as HTML or tags, special characters and extra blank-space characters. According to the type of textual data can be useful to keep specific characters for further analysis (e.g., in case of tweets, we can consider keeping hashtags). Punctuation is removed in most of the cases because uninformative in many applications.

Numbers

Numbers' utility is related to the studied domain: in many applications, they are not meaningful, but for example, if the studied field is Law, then numbers become very important because they refer to codes and laws, in the same way in the health sector they can represent drugs or medicines' names.

Lower case

As for numbers, most of the time, capital letters do not add information to researches, and for this reason, lowering cases is a widely accepted practice for several words. Transforming capital letter into the corresponding lower case does not change the meaning: the word "Car" has the same meaning of "car", however, if we take the word "rose", the flower and its version starting with the capital letter "Rose" this can refer to a proper name, having then a different meaning. This issue has also been identified with some brands that take generic names such as "General Motors", "Apple", "Sky," and many others.

This can be relevant also in the context in which we do a sentiment analysis. Words written in capital letters can assume double value, bringing emphasis to the meaning: it can strengthen the tone and enforce a different level of expression. As stated by Pak and Lee (2018), in their research, if the nature of the review is positive, capitalization makes it more positive. Similarly, capitalization tends to increase negativity in negative reviews (Pak, Lee 2018).

Stop-words

Stop-words are words that appear with high frequency inside texts and do not have a discriminative power. They are articles, pronouns, prepositions and conjunctions.

Removing them has become a good practice before analyzing a text because, in this way, we reduce the dimension of the associated vectors and the processing time needed significantly.

These words have a negative impact on weighting functions: indeed, all the remaining words would result in less impactful because of the high frequency of stop words inside the text.

Other words can be added to the stopword lists accordingly to the study field and purpose; those words that represent the subject and that are repeated several times do not contribute to the research (e.g., word “episode” or “watch” when we talk about a show).

Stemming

Stemming is a technique to reduce words into their corresponding stems: for example, words “amuse”, “amused”, “amusement”,” amusements” and “amusing” are reduced all to the single stem “amus”. Elimination of suffixes and prefixes are the most common form of stemming. Porter introduced this technique in 1980, and since then, other variations more and less aggressive have been proposed, such as Lovins, Lancaster, Paice and others.

The most significant disadvantage of applying the stemming method is that algorithms are not able to understand the meaning of words or context so that errors arise, for example, the word “party” will be cut into its stem “parti” when used in both contexts of “political party” or “birthday party”.

Lemmatization

Lemmatization is similar to stemming, but it brings back words into their canonical form called lemma. This technique tags each word understanding the part of the speech it belongs to and its meaning inside the sentence; it considers the neighbor sentences as well as the whole document. It takes a longer time to perform in comparison with a stemmer that is faster.

In some cases, Stemming and Lemmatisation can match; for example, taking the word “walking” in both cases, it would transform into “walk”. In contrast, the word “meeting” that has a double meaning will be stemmed in to “meet.” However, according to the context, it means “meeting” if in the context is used as a noun and not as the present continuous form of the verb. Words such as “better” are brought into their lemma “good”.

2.2.2. Feature representation

As mentioned above, texts are unstructured data that need to be transformed to be analyzed. Documents are, therefore, represented by weighted feature vectors in order to perform text mining tasks with a mathematical approach (Shashi, Kumari Singh 2019).

A brief introduction of the main methods is presented in the following section.

Bag-of-words

Bag-of-words is a simplified representation that underlies many more complex methods. It assumes that the texts can be considered as bags of terms without considering the context and, therefore, without taking considering the order in which words appear. Documents are transformed into lists of words with their respective occurrence. Taking an example, the following two sentences as documents:

- "We resolve to be brave. We resolve to be good. We resolve to uphold the Law according to our oath."
- To be, or not to be, that is the question.

Their bag-of-words representation will be respectively:

- {"We": 3, "resolve":3, "to":4, "be":2, "brave":1, "good":1, "uphold":1, "the":1, "law":1, "according":1, "oath":1}
- {"to":2, "be":2, "or":1, "not":1, "that":1, "is":1, "the":1, "question":1}

This construction does not pay attention to the order words are presented. The simplicity of this model is the reason why it is used as a preliminary assumption in many other more complex models.

N-Gram

An N-gram model is a probabilistic model supporting word prediction (Tonella et al. 2014). It determines the most probable derivations among multiples based on the idea that what has been previously found it is likely to happen again, by the following rule:

(1) $$P(e|e_1, \dots e_n) = P(e|e_1, \dots e_{N-1})$$

It consists of turning into conditioned probabilities the number of times that e precedes e_1 so that the next chosen word will be proportional to the frequency of occurrence of the tuples.

It can be used as a feature selector where N-grams that occur several times above a given threshold are kept, whereas others are discarded.

TF-IDF

Term Frequency-Inverse Document Frequency is the most common method used to convert corpus of documents into matrix representation of vectors. It is used to give relevance to each term in a text, assigning them a weight.

We define as *t*, a term inside a document *d* in a collection of *m* documents *D* called "corpus". TF indicates the number of times a term occurs inside a document; high values mean that the word is often quoted in the text, so it assumes importance. In this context, we understand why it is essential to remove stop-words before proceeding with analysis due to their high frequency inside texts, and they would show high values of term-frequency invalidating the entire research.

IDF, on the contrary, compute the number of times the same term *t* occurs inside the whole collection of documents *D* in the following way:

(2) $$idf(t) = \log\left(\frac{m}{freq(t,D)}\right)$$

This weighting score eq. (5) is useful when we have a collection of heterogeneous documents because it allows us to filter those terms that are common and discover those keywords that are peculiar of a few documents and not in others.

(3) $$tf - idf = frq(t,d) * \log\left(\frac{m}{frq(t,D)}\right)$$

Online social network analysis literature often relies on TF-IDF matrix representations (Curiskis, Drake et al., 2019).

2.2.3. Text mining techniques

2.2.3.1. Association rules

Association rules are used to extract frequent patterns, associations and interesting correlations among a set of data. This approach relies on two main metrics Support and Confidence.

The first one is an indicator of the frequency of an itemset in a database. In contrast, the second one indicates the number of times a rule is present in the dataset.

Let us define I as an itemset and *D* as a set of transactions T, defined as a set of items (X, Y) present in the itemset, $T \subseteq I$. Given the rule $X \Rightarrow Y$:

(4) $$Support(\{X\} \rightarrow \{Y\} = \frac{T\ containing\ both\ X\ and\ Y}{Total\ number\ of\ T}$$

(5) $$Confidence\ \left(\{X\} \rightarrow \{Y\}\right) = \frac{T\ containing\ both\ X\ and\ Y}{T\ containg\ X}$$

The first task is to find all those transactions in which support is above a given threshold to identify the most recurrent ones and then generate association rules from those item sets with minimal confidence.

A third metric, Lift, becomes relevant in determining which are the most significant rules. It expresses the degree of dependency between items and the strength of their correlations. Defined as

(6) $$Lift\ (\{X\} \rightarrow \{Y\} = \frac{(T\ containing\ both\ X\ and\ Y)}{(T\ containing\ X) * (T\ containing\ Y)}$$

- If its value is greater than one, then X and Y are dependent on one another, it means that if one is present, it is very probable that also the other one will be.
- If its value is lesser than one, then the two items are substitutes; it means that one is present, then it is probable that the other one will not be.
- If it assumes a value equal to one, then the two items are independent, and so no rules can be drawn.

There are several algorithms to mine frequent itemsets, here a high view of Apriori and FP-growth is introduced.

Apriori algorithm is a brute force approach that allows us to find frequent itemsets; It is based on the Apriori principle that "all subsets of a frequent itemset must also be frequent".

In this method, itemsets of increasing length are generated, cutting all those combinations where the Support value does not reach the minimum level. Once obtained the most frequent itemsets, the task is to select those rules that have a minimum confidence level. As last, to take decisions, we can prune those rules that do not achieve a minimum lift level.

This method requires a lot of time and memory since it scans the entire database several times.

The FP-growth, Frequent Pattern Growth Algorithm, can be considered as an improvement of the Apriori method because it does not require candidate generation. However, it fragments the database into a tree, starting with the longest transactions and proceeding in descending order.

Once identified the most frequent items (Figure 3) and a Frequent pattern set is created, it scans all the transactions, adding them into an Ordered item set (Figure 4) if these contain one or

more of the frequent items. This has a tree structure, and when an itemset belongs to two transactions, it will share the same root (Figure 5,6).

Completed the FP tree (Figure 7), a conditional pattern base is computed grouping the path labels for each node and for those paths that are common, Support is computed as the sum of all support of all paths in the conditional pattern base. Rules are, eventually, created by joining items with their corresponding conditional frequent pattern (Figure 8).

Item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

Figure 3: Frequent Items in FP-growth Algorithm

Transaction ID	Items	Ordered-Item Set
T1	{E,K,M,N,O,Y}	{K,E,M,O,Y}
T2	{D,E,K,N,O,Y}	{K,E,O,Y}
T3	{A,E,K,M}	{K,E,M}
T4	{C,K,M,U,Y}	{K,M,Y}
T5	{C,E,I,K,O,O}	{K,E,O}

Figure 4: Ordered Item Set in FP-Growth Algorithm

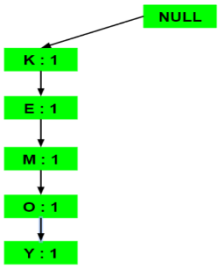


Figure 5: FP Tree in FP-Growth Algorithm

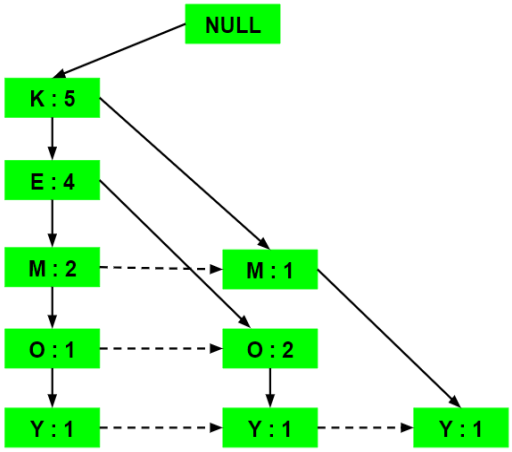


Figure 6:FP Tree in FP-Growth Algorithm

Items	Conditional Pattern Base	Conditional Frequent Pattern Tree
Y	{{K,E,M,O : 1}, {K,E,O : 1}, {K,M : 1}}	{K : 3}
O	{{K,E,M : 1}, {K,E : 2}}	{K,E : 3}
M	{{K,E : 2}, {K : 1}}	{K : 3}
E	{K : 4}	{K : 4}
K		

Figure 7: Conditional Pattern Base in FP-Growth Algorithm

Items	Frequent Pattern Generated
Y	{<K,Y : 3>}
O	{<K,O : 3>, <E,O : 3>, <E,K,O : 3>}
M	{<K,M : 3>}
E	{<E,K : 3>}
K	

Figure 8: Frequent Pattern Generated in FP-Growth Algorithm

Other versions of this algorithm have been proposed during these past years to improve its performances. However, it is still now one of the most used techniques in finding correlations and associations rules.

In-text mining, association rules are used to find relationships among topics, implications between concepts that characterized corpora. The purpose of this technique is to find those topics that, if present, implicitly implies the presence of another topic.

2.2.3.2. Clustering methods

K-means algorithm

This technique is one of the most used clustering methods thanks to its simplicity (Kodinariya, Makwana, 2013). It creates a certain number K of clusters, with K given a priori. Each cluster is shaped around a "centroid", K centroids in this case. Those points should be as different as possible from each other so that clusters will be distant. Each point of the dataset is assigned to a cluster where the distance between the point and the centroid is the shortest. The first set of clusters is created, and from each of them, new centroids are chosen. The assignment process is repeated until there are no available moves left.

In order to evaluate the result, it is common to calculate the Sum of Squared Error (SSE), where y is a point in a cluster and the c the centroid; The objective is to minimize function (7).

(7) $$SSE = \sum_{k=1}^{K} \sum_{i \in C} \|y_i - c_k\|^2$$

Multiple approaches have been proposed to identify a way to decide the right K, Kodinariyaand Makwana (2013) gives an overview of the most common ways, among them we can find the Elbow method and Silhouette.

The Elbow method is a graphical way to understand which is the maximum number of clusters that is a trade-off between being able to explain the variation and overfitting; This number is identified as the elbow in the curve drawn by the graph WCSS (Within-Cluster-Sum-of-Squared-error) and No. of clusters, after that point generated clusters will not be very different from each other, risking to create an overfitted model (Tripathi, Bhardwaj, Poovammal E, 2018).

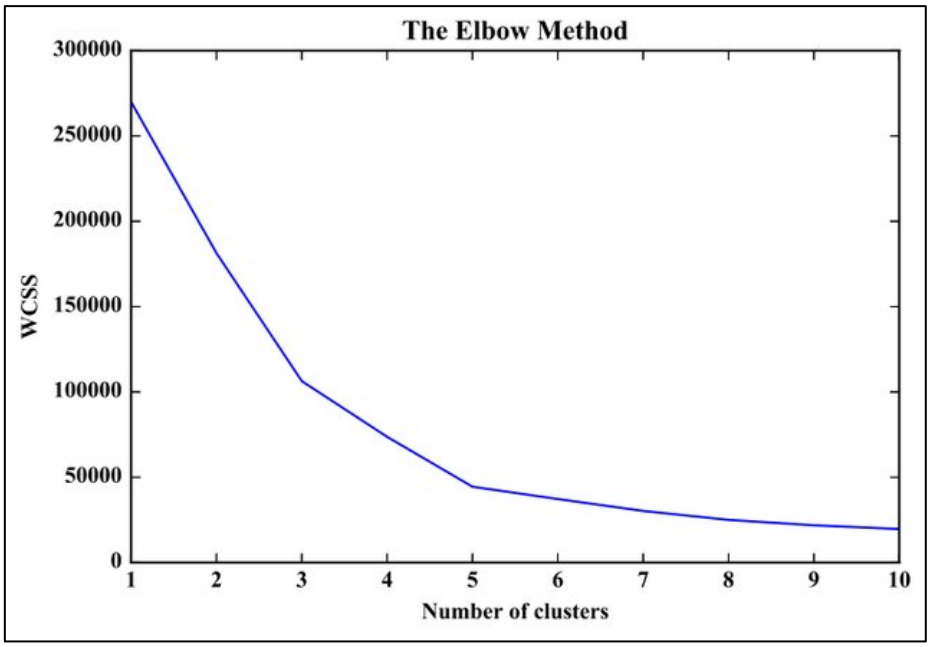


Figure 9: The Elbow Method
Source: Shreya Tripathi, Aditya Bhardwaj, Poovammal E, "Approaches to Clustering in Customer Segmentation", International Journal of Engineering & Technology 7(3.12):802, 2018

Number K of clusters can also be decided by computing the silhouette width that considers the distances within a cluster and the separation from the others. In this method introduced by Kaufman and Rousseeuw(1990), the element $a(i)$ represents the average distance between the element i and the other points belonging to the same cluster whereas $b(i)$ is the minimum distance between i and all other points in other clusters. It ranges between [-1,1] where values below zero mean that the point has been wrongly assigned to that cluster, values close to 1 mean that the point belongs to the right cluster and value around zero say that it can be in that cluster or another without any particular consequence.

(8) $$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$$

A good number K of clusters is the one that maximizes the average silhouette for different values of K.

A drawback of this method is its dependency on the initial choice of clusters' number. When this is chosen improperly, we will not obtain good results. This instability makes K-means a valid choice when the research is carried at a local level and not if we look for optimal global results (Xue and Wang, 2017).

Clustering methods are usually evaluated through two categories of measures, intrinsic such as cluster cohesion and separation, and extrinsic like precision, recall and F1. The first category

describes the cluster and its variations compared with others; cohesion indicates how much points included in the same cluster are similar, and separation is the indicator of the difference with other clusters. The extrinsic measures, instead, require a labeled dataset to compare with.

True positive (TP) indicates that two similar documents have been assigned to the same cluster and True negative (TN) that two dissimilar have been assigned to different clusters. The error is when we have False Positive (FP), meaning that two dissimilar documents are assigned to the same cluster and False Negative (FN), that two similar documents are in two different clusters. These values are combined to compute precision, recall and F-measure.

(9) $$Precision = \frac{\#TP}{\#TP+\#FP}$$

(10) $$Recall = \frac{\#TP}{\#TP+\#FN}$$

F-measure is used to evaluate the accuracy, and it takes into consideration both precision and recall giving us a single score that is a balance between the two.

(11) $$F - Measure = 2 * \frac{P*R}{P+R}$$

This technique is widely used in text-mining because this unsupervised process allows us to analyze large group volumes of text data, finding patterns, and it results to be suitable for unstructured data such as texts as well. Alnajran, Crockett et al. (2017), state two main reasons why it is a good way to perform analysis on texts: first, the amount of data that manually would need to be labeled is too vast and second that "the existence of unforeseen groups may carry important nuggets of information which can only be revealed by unsupervised learning" (Alnajran, Crockett, McLean and Latham, 2017).

2.2.3.3. Topic modelling

Latent Dirichlet Allocation

LDA, Latent Dirichlet Allocation, is a probabilistic model that is used to extract topic information from texts. It relies on the Bayesian model with the assumption that words are independent of each other. Once it has been trained properly, it will generate a word distribution per topic and a topic distribution per corpus. The basic idea is that documents are representations of latent topics, and each one is characterized by a distribution over words (Blei DM, Ng AY, Jordan MI, 2003).

A simplified explanation is here provided.

The first step is to set a *K* number of topics representative of the documents' collection; this number is decided a priori empirically. LDA, then, assigns each word in all documents to a topic randomly. We obtain in this way a topic representation by words and a document representation by topics.

In the following step, LDA will analyze each word to check if the assignment to a topic was correct. For each *w* (word) into *d* (document) the following probabilities are computed:

I. $P(t|d)$: the proportion of words in document *d* that have been assigned to topic *t*.

This because if the percentage of words belonging to topic *t* for document *d* is high, then it is probable that also *w* will belong to *t*.

II. $P(w|t)$: The proportion of documents that have been assigned to topic *t* because of the word *w*.

As we have previously mentioned, if a word has a high probability of belonging to a topic, then a document with that word will have a higher probability of being associated with that topic as well. Using this, LDA will move a word from a topic to another if the probability that a word belongs to a topic *t* is equal to the product of the two above mentioned probability, as shown by Eq(x).

III. $$P(w\ in\ t) = P(t|d) * P(w|t)$$

LDA works better when texts are rich, long, but the number of treated topics is not so large to make it hard to identify them.

2.3. A Natural Language Processing application: Social network Analysis

Social networks have been defined as "a social structure made up of people, or entities, connected by some type of relationship or common interest" (Camacho, Panizo-Lledot et al., 2020).

The Global Social media research summary in 2019 shows an increase of 9% of connected users every year, meaning that more and more information is exchanged by users directly and indirectly.

The reason why social networks are considered a resource to be exploited is the speed of diffusion and the influence they exercise that has no equal. Guille, Hacid et al. define as social influence: “a social phenomenon that individuals can undergo or exert, also called imitation, translating the fact that actions of a user can induce his connections to behave similarly. The influence appears explicitly when someone “retweets” someone else, for example,” (Guille, Hacid et al. 2013).

In a more technical language, this influence can be defined as linkage data, meaning the graph structure that connects the entities that communicate within social networks (Aggarwal C.C., 2011). These entities exchange text, images, video and audio with each other. It is a considerable amount of data that can be analyzed with data mining techniques and exploited for a set purpose.

Research conducted by Jansen, Zhang et al. in 2009 claims that this type of communication has a powerful effect in WOM branding and consequently impacts brand image, brand awareness and customer relationships. Understanding this, marketers today act directly on social media by creating pages, posts, tweets or others to create and maintain long-term relationships with their customers.

This becomes possible thanks to the ability of the companies to understand customers directly and in a certain sense without filters, observing closely their behavior and the idea they have about brands.

User profiling, topic detection and sentiment analysis are types of analysis most commonly carried out by companies because they allow understanding 1. who are the customers of a specific product and his/her behavior, User profiling; 2. what customers say about the product, Topic detection; 3. what is the sentiment that emerges from the opinions shared by users, Sentiment analysis (Ahmed Elragal, Nada Elgendy 2014); Data collected by social networks, therefore, allows to answer to these three questions, thus supporting businesses in their decision-making processes. In the following sub-paragraphs, a high view of these three study fields is given to the reader.

2.3.1. User profiling

User profiles are defined as “behavioral patterns, correlations and activities of the user analyzed from the aggregated data using techniques like clustering, behavioral analysis, content analysis and face detection” (Camacho, Panizo-Lledot et al., 2020). User profiling is the technique used

to process data to understand users' interests and use this knowledge to provide a better experience to the customer and create satisfaction (Kanoje et al., 2014).

Personalization is the key. Companies need to address their customers according to their preferences if they want to be successful nowadays; the abundance of information available overwhelms the user who must actively seek what may be interesting for him; however, companies that manage to communicate in a personalized way are winning because they can distinguish themselves.

There are many areas of application: e-commerce, banking, social media and thanks to the latter, user profiling applications have also been developed in marketing and advertising, to name a few (Farnadi, Tang et al., 2018).

The main objective is to collect data related to users' interests in a short period of time and to create a base of knowledge to measure their satisfaction (Kanoje, Girase, Mukhopadhyay, 2014). User profiling is used to segment customers in groups based on shared characteristics and create recommendation systems.

M. Gao, K. Liu, and Z. Wu (2010), in their research, identify three classes of user modeling: interest, behavioral and intention modeling. In the first class, the degree of interest is assessed concerning a brand, product or service; the data can be obtained explicitly, requesting it directly to all or implicitly observing historical data of purchases or browsing data. Behavioral modeling observes the interactions between users and platforms to then estimate future interactions, while intention modeling groups users based on their final actions (e.g., purchase) to observe their movements and replicate them. This last class was born from the union of the previous two.

Although some methods use data collected explicitly, through surveys, questionnaires or interviews, this is an inefficient and limited way because customers are often not honest or are not available to respond as it is a time-consuming activity. (Raghu, Kannan, Rao, and Whinston, 2001)

Implicit methods through the support of machine learning techniques (Kelly and Teevan 2003) are preferable as they can collect more information as well as update automatically: as Webb, Pazzani and Billsus (2001) show, sites like Yahoo or Google receive millions of views daily, and through automated systems, it is possible to able to manage and analyze this abundance of information.

Social networks result to be a valid source of data for this kind of analysis. It has been studied in many different works: G. Farnadi, Tang, De Cock, and Moens (2018) present a framework that combines textual data, visual data and relational data to infer personality traits as well as age and gender of social media users. Their model, called UDMF, "user profiling through deep multimodal fusion", has the purpose of predicting multiple attributes of social media users given their activities, their generated content and social-relational content. It is based on the assumption that the integration of different sources of data can provide a more accurate description than an individual one; Liang, Zhang, Ren, Kanoulas (2018) use Tweets to create a probabilistic model DUWE, "dynamic user and word embedding", to measure similarities between users and words in constructing user profiles over time. Their objective is to retrieve the most relevant keywords related to users' interests dynamically over time.

The listed above are just a few examples of a significant branch of study that involve social networks, machine learning, texts and user profiling.

2.3.2. Topic detection

Topic detection, the discipline that aims to discover topics in texts. It is usually carried out with clustering techniques and topic modeling.

The topic model has become a key research field because it extracts valuable and useful information from many texts. The traditional representations commonly used, for example, SVM, do not consider the underlying semantics and the relationships between implicit topics. Topic modeling, instead, is a statistical probability model able to mine the semantic information so that to discover hidden topics (XU, Meng, Chen et al., 2019). The topic model is applied to different sources, from face recognition to NPL. In the context of text mining and machine learning, it has become a real hot issue (Qiu and Yu, 2018).

Here below a series of researches conducted, either using clustering methods or topic modeling, such as LDA, is presented to the reader.

Liu, Li et al. (2015) proposed a Single-Pass[1] Clustering in LDA method to extract semantic information from various sources regarding food safety problems: their approach is divided into two stages, in the first part uses LDA to create a topic distribution of documents, and in the second one they used K-means to cluster documents and highlight topics. This method aims to solve the problem of sparse data, increasing precision and recall. Gropp, Herzog et al., 2019,

[1] Used in VSM, value stream mapping, it analyses the current state to design a future state.

introduce CLDA, “Clustered Latent Dirichlet Allocation”. They apply this method to research papers and journals segmenting information by time and type of journal. They decompose data into segments running the LDA in parallel to output topics for each one that then is the new input for a parallelized K-means whose output will be a list of representative topics. Each of the original topics is a part of a global topic cluster. Wang and Zhang, 2014, use the LDA model to find centroids in the clustering K-means algorithm so that to solve the problem related to the right number of K; In 2014, Godfrey et al. analyze a twitter data set applying K-means clustering to the tf-idf representation of tweets to create clusters of the most mentioned topics. Steinskog and Therkelsen, 2017, find in clustering methods the best way to perform topic detection of social media data such as tweets where LDA results to be inadequate due to the sparsity and variety of short texts. Alnajran, Crockett et al.,2107, carried a comparative analysis of different clustering methods on tweet analysis, stating its importance in pattern recognition for this kind of unstructured data as well as its weaknesses.

The ability to extract topics from web discussions or other sources becomes relevant in business because, in this way, companies can detect emerging topics as soon as possible. Addressing customer complaints or remedy to negative impressions becomes easier adopting techniques such as those above mentioned as well as finding new opportunities in emerging consumer-generated trends to transform into a strategic advantage (Colbaugh, Glass 2011).

2.3.3. Sentiment analysis

Sentiment analysis, or opinion mining, is “the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes” (Liu, 2012). Its impact is relevant both in NLP[1] studies and management sciences, economics, social sciences or politics (Liu, 2012).

With the enormous growth that social media had during these past years, this practice has become increasingly important. It is used more and more in decision-making processes. Web scraping, or web content mining, “describes the use of a program to extract data from HTML files on the internet” (Haddaway Neal R., 2015); it is the activity of collecting and analyzing data from the web to understand the opinion of a product or brand. It does not limit to social

[1] Natural Language Processing

media but includes the whole web. Due to the voluminous data, automatic systems are required and more and more implemented in the context of sentimental analysis.

As we see, sentiment analysis is often used to give voice to customers and try to collect that information that is not directly reachable. In order to do that, complex research as to carry on "tackling many NLP tasks, including personality detection, domain adaptation, and multitask learning (Camacho, Panizo-LLedot, et al. 2020).

Today, most of the techniques adopted are supervised techniques that require training data and labeled samples where the sentiment is explicitly indicated (Cambria et al. 2015).
Sentiment polarity usually is divided into "positive" and "negative"; Alsaeedi and Zubair Khan (2019) made a further distinction: they add "neutral" in case the classification is a document-level or "unbiased" in case the research is a sentence-level.

However, understand the sentiment of a sentence is a very difficult task due to the complexity and variability of the expressions that can be used. Annett M., Kondrak G., in 2008, identifies one of the major obstacles in thwarted and negated expressions; a thwarted expression is a sentence that contains words with a polarity that is opposite to the meaning of the expression itself, the following example is used to explain it better: "Johnny Depp was all right. The previous two pirate movies were unrealistic and boring. The plot was awful. However, the special effects made the third pirate movie excellent" (Annett M., Kondrak G., 2008), words like "unrealistic"," boring," and "awful" are negative words, but the sentence itself is positive.

Other factors that affect the sentiment's understanding are sarcasm, "neutral reporting of balanced information", "success or failure of one side concerning another", "rhetorical questions" and also "quoting someone else or re-tweeting" because we cannot be sure if who share has the same opinion of the person who quoted (Das, Cambria et al., 2017).

There are many studies that we can find regarding this topic as well as experiments conducted, here few examples: in 2007, Liu, Xiangji et al. adopted opinion mining techniques to extract the sentiment from blogs and then forecasts sales performances of movies, similarly in 2010, Asur and Huberman used Twitter chats to forecast box-office revenues and how the sentiment extracted can be further exploited to improve predictions regarding the success rather than using the ordinary market-based elements.

Yano, Tae et al. used LDA[1] to capture reactions and opinions from posts, blogs and documents regarding Political topics to anticipate users' comments, whereas McGlohon, Glan, Reiter in 2010 compared reviews of products and merchant to assess true quality.

2.4. From WOM to E-WOM: how social networks have changed the way to do marketing

The theory of the principle of sufficiency or minimum effort assumes that a person always wants to reduce the thinking process by trying to obtain information that simplifies the decision-making process and the amount of memory to use rather than considering all the available options (Wood et al., 1985). One of these ways is collecting information coming from other people.

Word-of-mouth marketing has been defined in several ways, Arndt (1967) defines it as "oral, person-to-person communication between a receiver and a communicator whom the receiver perceives as non-commercial, concerning a brand, a product or a service", for Nyilasy it has three main characteristics: it is interpersonal communication, its content is commercial but despite that people are not commercially motivated, or at least this is the perception the most of the time (Nyilasy, G., 2006). It is an interaction that becomes free advertising triggered by customer experience.

Over the past 15 years, word-of-mouth communication has developed in different forms, and what previously was a face-to-face contact is now projected on many online and multimedia channels. Hennig-Thurau et al. (2004) describe the phenomenon of the online word-of-mouth as "any positive or negative statement made by a former, actual, or potential customer about a product or an organization to more than one person or institution via the internet" (Hennig-Thurau, Gwinner, et al., 2004)

Consumers tend to rely, at first sight, to their social ties to validate the WOM usefulness: family and friends (Brown and Reingen,1987) but also remote ties such as celebrities (Duhan et al., 1997) have an impact on how it is perceived. Those ties are missing in a virtual environment, where who writes reviews are most of the time strangers that share their opinions for different reasons, and this usually forces readers to evaluate the message's usefulness almost solely based on the content of the communicated message (Walther 1996).

[1] LDA: Latent Dirichlet Allocation

It has become an everyday habit to visit online review sites to be informed about the positive and negative aspects of a product through the experiences of others because it is perceived as less biased and easily understandable by the majority of readers (Mostafa, 2013).

There are significant differences when going through the virtual world.
First of all, the scalability and the speed with which the opinions are spread have no precedents: if before it was possible to share within a small group of friends and relatives synchronously, we talk about a topic while we are living it. The world of the web allows us to communicate asynchronously with anyone (K.H. Hung, S.Y. Li 2007). The information remains available and accessible over time and is usually traceable and organized into categories, threads, which allow the user to find it quickly and, therefore, less chaotic or sparse than a face-to-face conversation (C.M.K. Cheung, D. R. Thadani 2012).

Compared to traditional WOM that is usually from known sources, small in number, and either positive or negative, eWOM is anonymous in the source, voluminous in quantity, and variable in valence (Yue Pan and Q. Zhangb, 2011). The credibility of the traditional WOM is more natural to evaluate for those who receive it: get to know the person who commented helps to evaluate the value of the content more easily. It allows, then, to adapt the comment accordingly to both which is the content and who is the sender (Cheung, Thadani 2012).

Credibility is, indeed, one of the main factors studied by Brown, Broderick, and Lee (2007) to understand the power of WOM to influence others' decisions, its persuasive strength.

The first to be identified is the tie between sender and receiver, the stronger the bond and the higher is the influence of this in the decision-making process (Brown & Reingen, 1987). Second is homophily, or the similarity between the members of a group in terms of attributes and characteristics such as age, gender, education, culture and lifestyle (Rogers, 1983). However, this leads to different results: on the one hand, it limits the type of information that is received and the type of interactions that are experienced because they are all related to the same type of belonging (McPherson & Smith-Lovin, 1987) but on the other hand, towards the similarity between individuals, it creates bonds of trust and understanding that are possible only between people who share the same interests and the same situations for which the influence they generate is powerful. (Schacter, 1959)

The lack of context's knowledge where an online comment comes from has an impact on the credibility of the source, which is identified as the third element that allows understanding the influence of the WOM. Without any certainty regarding sources, in the online world, proxies

are sought to assess validity. Among these, expertise and lack of self-interest are the most recognized: someone who is recognized as an expert in the sector and has knowledge about a field has a more considerable influence, but more, in general, there is a tendency to believe to online reviews because they are not guided by a personal interest in advertising a product. (e.g., Arndt, 1967; G. Silverman, 1997)

With these premises, Cheung and Thadani in their research "*The impact of electronic word-of-mouth communication: A literature analysis and integrative model"* (2012) tried to understand if this type of communication influenced the customer's intentions, claiming that this process plays a significant role in consumer buying decision making.

From a business point of view, eWOM and strategy of online buzz marketing, the interaction between users that amplifies or alters the original message, is less expensive and easy to distribute widely and, in some instances, even more measurable.

As said above, consumers easily accept fellow's recommendations compared to traditional marketing forms. If a social network user posts a comment about a product, then his friends will read, add messages turning in a discussion, when its volume is significant then it may become a buzz. In their study, Pauwels, Bucklin and Trusov (2009) have compared the effects of traditional marketing and social network WOM, pointing out that the elasticity of the second one is 8.5 times higher than ordinary actions the first day. This difference grows even more in time, "it is approximately 20 times higher than voice spread by marketing events and 30 times than one by media appearances" (Pauwels, Bucklin, Trusov, 2009), making WOM one of the most powerful communication tool. Mahajan, Muller, & Kerin state: "WOM communication overcomes shortcomings of seller-centric marketing communication messages in that it provides useful information by peer consumers, who have purchased and experienced products/services" (Mahajan, Muller, & Kerin, 1984).

2.5. Application in the Entertainment industry: how the advent of streaming platforms and social media marketing transformed the entertainment sector

2.5.1. How streaming platforms lead to a change in consumer behavior

It is called “the Netflix effect”: the disruption created by the streaming platform that proposes a more convenient and customized way to consume video content in one sitting for hours, which sometimes brings to lose the concept of time (Matrix 2014).

This mode of consumption has been defined as “Binge-watching”. There is not yet a standard definition for this phenomenon. However, Netflix itself refers to it as “the activity of watching two to six episodes of the same show in one sitting” (West, Kelly, 2014).

According to Netflix’s 2019 annual report, over 167.1 million people in over 190 countries subscribe to Netflix with an average of 71 minutes per day and a cumulative 165 million hours of watched daily across the globe (www.netflixinvestor.com).

The freedom given by the streaming service can be one of the motivations of its success since it allows customers to manage in a natural way where, when, what to watch in addition to the choice of the device. However, this new approach of watching for hours, ‘blends culture and technology’ (Steiner and Xu, 2018), means prolonging a viewing experience and engagement with a fictional world, as well as emphasizing the story world over the lived experience (Perks, 2015).

The narrative transportation, defined as the phenomenological experience of escaping into the world of a narrative, written or audio-visual, (Green and Brock 2000) is one of the foundations on which the research of Erickson, Dal Cin and Byl in 2019 are basing their theory. They studied, indeed, how this new way of watching multiple sequential episodes of a TV series in a compressed time increases audience engagement.

Transportation theory suggests that the levels of enjoyment and engagement of a viewer increase when it is repeatedly carried into the reality of media narratives (Green et al. 2008), and binge-watching maximizes this in a short time, also minimizing interruptions.

Moreover, parasocial relationships, that are the perception of an intimate bond with a character (Boon and Lomore 2001) are deeply developed in a context such as the one resulting from the

binge-watching activity because of the speed of disclosure and increased degree of familiarity with media figures that are functionally similar to interpersonal relationships (Giles 2002).

These two aspects are directly correlated to media engagement, a known predictor of increased media effects (Green et al. 2004). Observing a relation between binge-watching and narrative engagement would suggest that the changing ways in which audiences are engaging with media content may have significant implications for the strength of media effects on these audiences.

2.5.2. Marketing communication in the movie industry: from advertising to E-WOM

Movies and TV shows can be considered as experience goods; their quality and utility indeed can be evaluated only after the consumption (Wallentin, 2016).

This kind of product is related to enormous production costs that make the uncertainty related to its success even more important. After the release, the content cannot anymore be modified neither withdrawn from the market because all the investments would be lost; for this reason, distribution and marketing become the two pillars to make it successful.

Movies' lifecycle is very short as a consequence interest becomes the key: people should be interested and intrigued by the movie in order to watch it, and it is marketing's job to create these feelings in consumers' minds.

Print and advertising for many years have been the two main channels used to communicate the release of new products. However, the possibilities have been multiplied, and the channel map contains many different tools through which companies can advertise their content. During the pre-release period, those tools are the most used because they inform the public about the product as well as investors about the potential profit (Hanssens and Joshi, 2009).

Today, those channels, usually manage by third-party agencies, that a company pays in order to promote itself to a broader audience, are called Paid media. Even if in the movie industry they still represent a significant portion of the promotion, as argued by Rennhoff and Wilbur (2011): "frequent new product introductions and short product life cycles lead to unusually high levels of advertising in the movie industry" , thanks to innovation developments, channels, that before had a minor role, have gained importance giving new opportunities to companies: Earned channels result to be a potent communication instrument, they assume different forms such as

conversations and comments on social media but also face-to-face communication (Fill, Turnbull, 2019), reviews and, more in general, we can refer to them as Word-of-mouth.

These new tools are not directly generated by brands (Mattke, Müller & Maier 2019). For this reason, they represent both an opportunity and a threat. They are perceived as more credible and trustworthy by consumers that on the contrary, see advertising as a complimentary of the product. Reviews and word-of-mouth are more probable to be unflattered according to the audience; if well leveraged, then they can increase for free the value and the global awareness of upcoming movies or Tv series.

Film marketers publish trailers on television, create official websites and social media pages so that people can gain awareness, receive online information, to arouse interest and let moviegoers create buzz and spread messages with friends, family and the entire web.

To reduce uncertainty related to an experience good such as a movie or a TV series, indeed, it is common to look to different sources of information to gain knowledge about the content.

Marketers have learnt how to engage with fans and customers to arouse their interest, here some examples: in 2010 an online test on Facebook have been published before the release of the movie “Percy Jackson and the Olympians ” to allow users to determine which Greek god they were or similar was the app, related to “The walking dead” to turn people’s pictures into zombies or the avatar maker to create the 1960s style portrayed inspired to the TV series “Mad Man”. All these examples have contributed to increasing awareness and visibility.

Online WOM takes different forms in the entertainment sector: reviews are considered as a source of product information because they should reflect user experience and consumer satisfaction, whereas discussion boards, chat rooms, blogs and community sites involve consumer expectations that can be influenced by social structure. (Viswanathan, Malthouse, Maslowska, Hoornaert, Van den Poel 2018).

Oha, Roumanib, Nwakpac and Hu (2016) identify three phases before the advent of the web in which communication between peers took place. In the beginning, people used to gather into clubs, face to face meetings, to share comments, opinions or anticipations on new releases; secondly, with Web 2.0, there was a migration to film review sites such as IMDB or Rotten Tomatoes. Finally, social media allowed the development of a multitude of activities such as

participating and interacting in discussions, creating viral content and as well as searching for information.

In this latter phase, we also distinguish two consumers' models (Heinonen, 2011) we have the passive consumer "lurkers" or who consume the content and the active one that publishes content defined as "posters" (Shang et al., 2006; Shao,2009). The latter is what makes the social network different from the usual means because it is the consumer who acts to generate user-generated content (Krishnamurthy and Dou, 2008).

Oha, Roumanib et al. (2016) instead associate the type of consumption to the type of consumer input, defining three Consumer engagement behaviors: the behavior related to liking a Facebook profile or following a Twitter account as "participation CEB[1]", "production CEB" is when content is created and "CEB of consumption " of who watches videos or reads posts and articles (Heinonen, 2011). However, there is never a single distinct profile, but combinations of them (G. Shao, 2009).

Interaction and involvement are actions attributable to interest and attention (Brodie, Iliet al. 2013) shown by the user towards a brand or product (E. Abdul-Ghani et al., 2011): Rui, Liu, and Whinston (2013) have studied how directly this and the E-WOM influenced the box -office and rankings, highlighting that when there is significant participation, then the consumption is higher (R. Rishika, A. Kumar 2013; Wu, Huang, Zhao, Hua 2015). Using the number of likes and retweets as a metric (D.L. Hoffman, M. Fodor 2010), it is found that a high number leads to high levels of CE: "users who "like" a product or brand spend up to five times as much money on their liked products compared with those users who do not "like" these products" (N. Hollis, 2011, from http://www.millwardbrown.com/global-navigation/blogs/post/mb-blog/2011/04/04/Thevalue-of-a-social-media-fan.aspx). Oha, Roumanib et al. argue that "movies with high personal CEB will positively correlate with increasing movie box-office revenue" (Chong, Yaman et al., 2016).

Among social media available today, Twitter plays one of the most relevant roles in this context. With 8,951 tweets per second (Internet live stats), it reports significant volumes of web traffic that are used by companies to spread their messages to their target audience, allowing the audience to communicate back. Users use this in two main ways: the first one, with a utilitarian purpose; they follow movie channels, read information about the plot or new releases and watch trailers (Oha, Roumanib, Nwakpac and Hu, 2016), or they "live tweet".

[1] CEB: Consumer Engagement Behaviour

Live-tweeting has been defined as the activity of share experiences while watching a show, just to feel connected with a broader audience; it is part of the so-called "Social TV", that is, the use of social networks as Twitter or Facebook stimulated by TV programs. (Buschow, Schneider and Ueberheide 2014).

Fans create online forums and real-time social experiences by using tags and hashtags, commenting with other connected viewers on a second screen, usually a laptop or a mobile.

Users have to follow a real etiquette when they practice this online activity: for example, avoiding spoilers, live tweet only when the show is airing on television, avoid to over-tweet but also to signal own intentions before the show begins result to be good norms.

2.5.3. Antecedents of E-WOM

If from one side, it seems that the customer engagement is the result of the binge-watching activity as a natural consequence, on the other side we can identify other motivations that can explain why people decide to spend time and energies in showing their involvement with entertainment content, creating their versions of it and developing WOM or buzz.

Yu and Ramaprasad defined this kind of engagement as: "a user's degree of voluntary allocation of personal cognitive, emotional, and behavioral resources (e.g., time and energy) to a platform related interaction, which can involve a product/content, other users, or the platform itself" (Yinan, Jui, 2019); Viswanathan and Malthouse (2018) use the acronym CEB, "Customer Engagement Behaviour", as the customers' behavioral manifestation toward a brand or firm, beyond purchase, resulting from motivational drivers.

Previous research (Braojos-Gomez, Benitez-Amado, Llorens-Montes 2015) distinguishes two types of consumer engagement, defining conventional engagement towards the official website of a brand or product and social engagement that one related to the use of social media websites. The latter takes on a vital role in the B2C relationship.

Engagement's antecedents are related to several areas, such as the need for information and entertainment, but also the social need to feel a sense of belonging with someone sharing our same interests.

In 2008 Dou and Krishnamurthy created two groups into which they divide the main motivations that can explain this engagement: rational and emotional. The former includes the search for information, knowledge sharing and advocacy; the latter includes self-realization, self-expression and connection between peers. Also, Shao, Courtois et al. in 2009 had identified

information and utilitarianism, social interaction, community engagement, self-realization and entertainment as drivers.

Social utility or altruism can be identified between the reason why people decide to comment: individuals, who engage in eWOM with altruistic goals, share their experiences for the benefit of others without expecting anything in return (Parikh et al., 2015).
Cheung & Lee (2012) consider enjoyment in helping others by leaving reviews as an essential motivation for consumers to spread eWOM, saving other users from bad experiences, especially when they are their fellows.
Tong et al. (2013) call this phenomenon as "self-fulfilment" due to the personal satisfaction that users feel in contributing with online content to improve others' consumption experience or in case of a bad result as a sort of vengeance against the product/service and to help others in making a better choice.
The search for belongingness is another critical driver. Social networks are used, by many users, in particular by teenagers, to establish their self-esteem and to create their identity (Yermolayeva, Calvert & Pempek, 2009). Many, indeed, engage on those platforms in order to find people like them, someone that can understand their point-of-view and perspectives, models to recognize themselves.
In their research Schirra, Sun and Bentley (2014) called 'Together Alone', they point out, indeed, how people look for connections with broader audiences, while they watch TV series, to feel part of a phenomenon that extends beyond their group of friends but make them in contact with people across the country that are strangers but with whom they share same interests and the same interactive entertainment experience.
Some of the people they interviewed stated: "the difference between watching a Tv series with and without Twitter is sort of like the difference between watching a movie at home on a DVD and watching the movie in a movie theatre … Like when you go to a movie theatre, and you feel like you are part of an experience because there are other people sharing it with you." (Schirra, Sun and Bentley (2014)).
It is usual for people belonging to the same family or groups of friends to have different tastes, this way of connecting on the web, therefore, gives users the feeling of not being alone when they are alone in their houses watching a series on their laptops or tv; an interviewee says: "If I cannot say it to anyone else around me, I'll just put it up on the Internet. Someone will read it, someone will agree with me at some point. It is just a reassurance that I am not alone." (Schirra, Sun and Bentley, 2014).

The next step is when fans may want to go further and experiment a feeling of “camaraderie”. It is related to group affiliation and togetherness. It can evolve in creating or joining an online community (P. Hedlund 2011).

Group pages on Facebook dedicated to education, political views or entertainment are an example of how members exchange information, even if they do not know each other, gathering to form a virtual fandom or forums.

This can be considered as ‘real’ presence because reviewers engage in talk and dialogues with the other members, they narrate “stories” with appropriate content and structure to their reviews, giving a personality to what they tell, creating as a consequence an impression of real existence (Kumar and Benbasat, 2006). Consumers engage jointly to express mutual sentiments, to commit and accomplish common goals (Kozinets, 2002); they socially reinforce consumption. Consequently, it increases awareness, loyalty and visualizations.

McAlexander et al. define a community as "a specialized, non-geographically bound community based on a structured set of social relationships among admirers of a brand" (McAlexander, Schouten and Koenig, 2002). Ouwersloot and Odekerken-Schröder (2008) analyze how these communities develop and how certain users become leaders by creating debates, discussions or organizing events and gatherings. As already mentioned above, 51 V. Madupu, DO. Cooley also distinguishes two types of personalities within the communities: the so-called lurkers or those who do not interact and those who are active members. Among the lurkers, there is still a possible distinction, namely those who simply read messages and articles and those who forward the information collected to others through other channels; most users seem to belong to this second category (V. Madupu, D.O. Cooley 2010).

Reward and social recognition are some of the identified drivers in the decision to share content online. This type of motivation is divided into two parts by Yang and Lai (2010), the first involves the group's expectations and how accomplishing them is leading to personal satisfaction, and the other as to be part of the group itself and thus gaining own status and reputation.

Therefore, be rewarded and praised for contribution is a reliable driver: in writing their own opinions, many users seek for recognition from others, to know that what they say is approved and shared, they feel satisfied when others agree with their observations.

Another external factor that can influence consumers to leave a review is the economic reward. Financial rewards (Tong et al. 2013), then, are motivational drivers in case they can enhance the self-image of the reviewer or if writing a comment, it is not considered an effort. Indeed,

this driver is relatively insignificant for customers that are not involved by the product or service, or they do not feel any kind of utility.

Shao (2009) connects all these drivers to the type of consumption, stating that the production of posts, videos or images is linked to self-expression as part of a general sense of entertainment, participation as likes is linked to a need for social connection and finally, the passive consumption relates to the field of information and entertainment. In general, Courtois et al. (2009) state that information, social connection and entertainment are the most common motivations.

3. Research gap and questions

Globally, the on-demand entertainment sector represents 29.9% of the Digital Media market, second only to the video games industry.

As mentioned above, the advantage brought by it is the freedom given to the customer in being able to choose how, where and when to watch content. Before on-demand services with analogical systems, it was necessary to wait for the television to broadcast a program or buy a DVD or a videotape; today, we are no longer dependent on any means, time or place. In addition, the possibility of having a variety of programs and a wide selection available allows the exploration and discovery of new genres or productions previously unknown.

On-demand streaming platforms have seen continuous growth: as reported by Statista, the sector sees a continuously growing revenue index (Appendix 1) with around 8.949 $ M of revenue and 146M of users in Europe in 2020. The figure reported in appendix 1, 2020 the revenue growth index had a peak during the Coronavirus health crisis, which prevented people from leaving their home, which led to an increase in the number of subscribers to SVoD services looking for entertainment.

The number of subscribers is key in this new business model, which is different from the traditional systems' one: it is no longer necessary to go to the cinema and pay a ticket to watch a movie or watch advertising on television while waiting for the next program. The revenues that allow these companies to survive come from subscriptions and therefore rely on a completely different approach.

In the traditional television system, a production company creates a show that becomes its product for sale. In order to amortize production costs and earn, usually, the first step is to find a television network that pays the licensing fee to have the right to broadcast the episodes. The channel then will generate revenue through the advertisements that will be placed between one break and another.

However, the Netflix streaming platform is based on a mixed business model; it transmits content produced by third parties to whom it owes broadcasting rights for each country, and at the same time, produces original content. Its revenue comes from the number of subscribers and not from advertisements.

In the first case, when a series is no longer broadcasted, the main reason is usually the recorded low audience so that the costs related to the transmission cannot be justified. In the second case, however, the yardstick on which the cancellation or renewal of a series is based is more complex. In this case, everything is based on the number of subscribers who renew the subscription every month.

To decide whether to renew a series, Netflix relies on efficiency metrics that describe the ability to sign up and retain customers: shows that succeed in both purposes are renewed while those with smaller fans, even if they are keen on it, are often canceled.

The decision is based on the concept of valued hours: it does not merely refer to the time spent watching Netflix, but what percentage compared to the total time spent watching Netflix that particular show represents. If a customer is willing to pay the entire subscription to watch only one show, then it means that this is important for the customer, and this becomes important for the platform as well. If a customer watches many programs within the platform, then his time is less valuable and therefore, consequently, the fact that he has seen a series is not indicative that this will be renewed. The main task for a streaming platform is, therefore, to propose content that is able to retain customers that will pay in order to watch it.

Statista research shows that the growth in the number of users who are willing to pay for a streaming service is almost saturated, but the fact that there is a shift, from the traditional TV/Cinema consumption to streaming video, will lead to an increase in the ARPU[1], i.e., same users will be willing to pay more to have access to multiple services simultaneously spending more time-consuming content (Appendix 2,3).

[1] ARPU: Average Revene Per User

This data, therefore, creates an even stronger competition between the major companies that offer this service because they must share the number of users. Two are the biggest players nowadays, Netflix and Amazon Prime video, followed by other platforms such as Hulu, NowTV and the new entry Disney +. With the entry of this latest service, which boasts a strong brand identity, it becomes necessary to find strategies to differentiate from the competitors.

In an industry that is growing but which risks saturating due to excessive supply, that is very competitive, where customers are bombarded by thousands of inputs and can tap into multiple resources, how can companies still attract and retain their customers?

As we mentioned above, so far, Netflix has led this sector, showing to be able to satisfy its customers; however, a study carried by 7Park Data shows that the 63% of members watch licensed content (40% watch only licensed content) and 37% watch original content (Figure 10).

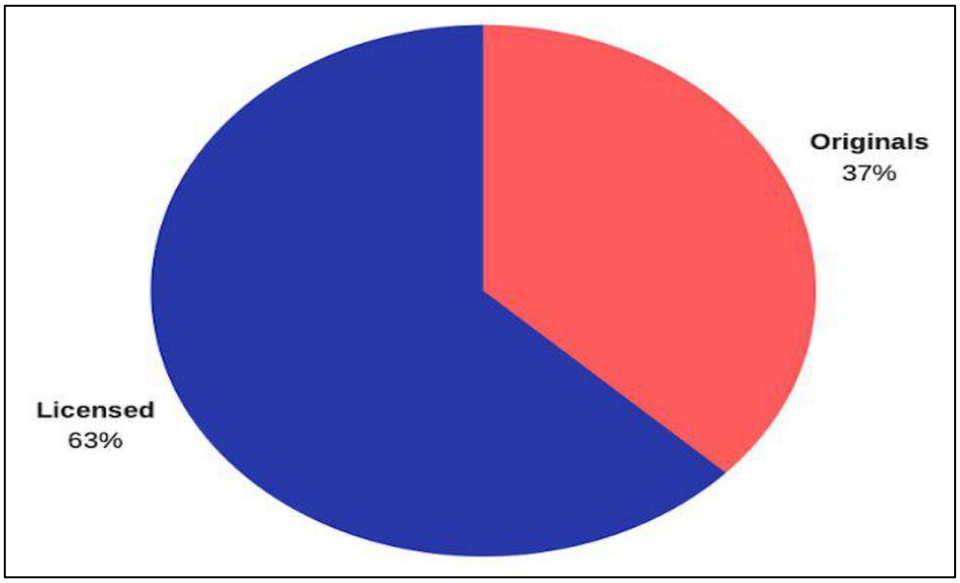


Figure 10 - Netflix Licensed Vs. Original Content Viewership
Source: 7Park Data

Three out of five programs among the most viewed content appear to belong to other services (Appendix 4). In 2019, Netflix paid $100 million to be able to offer in its catalog “Friends”, cult tv show belonging to Warner Media, that positions itself as the second most-streamed program in 2018 in the USA on the platform; Similarly, other iconic shows keep making money even after a long time from the date of release.

Thanks to the big success of this business model, movie companies decided to propose their platforms: Disney just launched Disney+, Warner Media and NBC Universal are planning to

launch their services soon, claiming back their productions such as *Grey's Anatomy*, *The office* and the already mentioned *Friends* that are all pillars of the Netflix offer.

If Netflix's members are watching more third parties' content, how will the company adapt to the new competitive environment? How will it be able to retain customers once the rights to broadcast licensed content will be over?

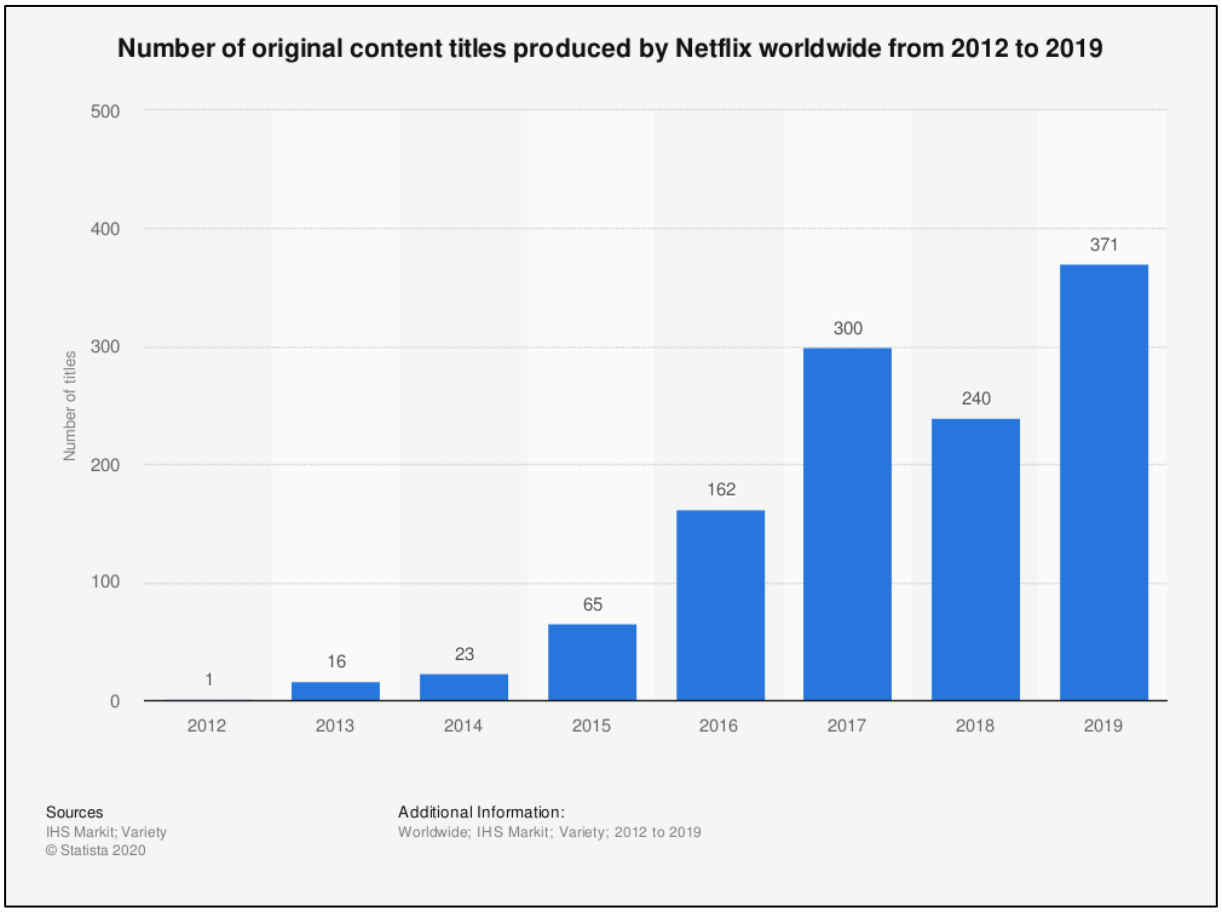


Figure 11 - Number of original content titles produced by Netflix from 2012 to 2019

As we can observe from the chart (Figure 11) published by Statista from 2012, they focused their strategy on original content productions that pass from 1 to 371 movies and TV shows auto produced available in the platform. The production of originals is a good strategy to try to reduce the threat of content loss due to the end of streaming rights. However, as we can see from the chart (Figure 12), the revenue is not matching the costs that the company needs to afford to be able to produce its content. As a result, it will be necessary to raise prices giving way to its competitors such as Amazon Prime Video, Disney+ and others to take advantage of that.

In the following years, therefore, the company will face terrible times losing part of its most valuable offer as well as fighting against companies that can rely on a very strong brand image and a big selection such as Disney or a solid financial base that allows big investments as Amazon.

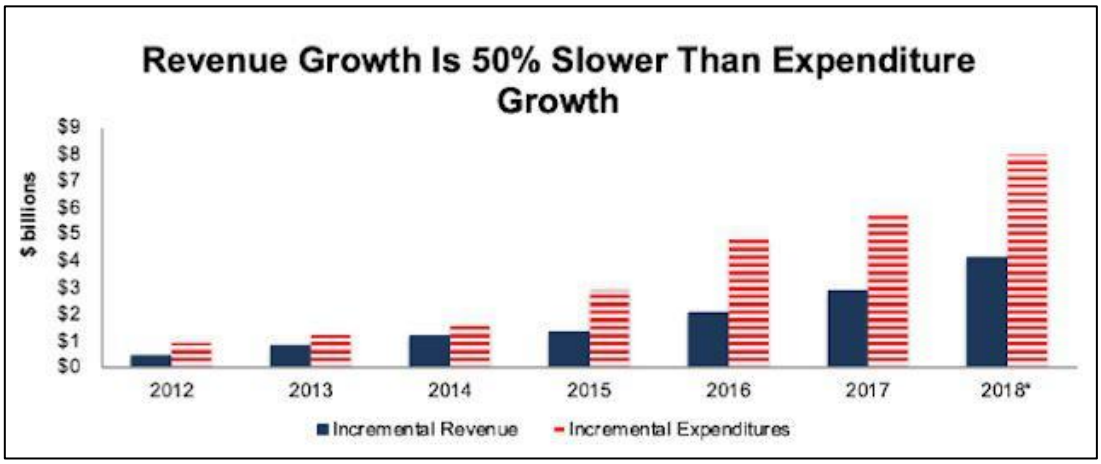


Figure 12 - Netflix Revenue growth VS Expenditure growth
Source: New Constructs LLC – 7Park Data

This research is going to use NPL techniques to identify the different stages in Netflix's strategy in the past six years, looking at its content choices and marketing strategy.

In order to do that, this research has been structured in four parts: the first section will analyze the letters to the shareholders released every quarter from 2014 to 2020 to see what and how Netflix communicate about its content choices; second, it will look into the current Netflix catalog to identify the offer how it has changed in time. In the third part, social media data are used to collect customer's opinions. A deeper study on the social media engagement strategy is considered for further researches. To conclude, we will give voice to customers to understand their choices, how peers' influence affects the entire decision making during the content selection and, as a consequence, the entire industry.

Q1. What strategy has Netflix adopted in its content offering in recent years?

Q2. What is the customer opinion of the company? What are the elements that characterize it?

Q3. What are the elements that influence the customer's decision-making process in choosing the content to watch and how this can consequently affect product success?

The purpose is to have a complete overview of both the business and customer side to identify weak points and possible resolutions.

4. Methodology

For this study, a combined methodology has been used. To answer to questions number one and two, data mining techniques, in particular, text mining techniques, will be applied to retrieve information from Netflix letters to the shareholders, content descriptions, and tweets. A survey has been conducted to understand the role he has in the decision-making process during the content selection and his motivations to influence other people's choices.

4.1. Text Analysis

Netflix releases every quarter a letter to its shareholder where it communicates financial results and forecasts, changes in leadership, marketing strategies and content. In this latter section, it evaluates past release performances, announces new content and the general strategy. The purpose of analyzing the content part of the letters to the shareholders is to dive deep into what Netflix has proposed during the last six years and what it focused its attention on.

Once collected letters into one folder, through the programming language Python, a simple text mining pipeline has been applied to be able to visualize into a network graph the most mentioned bigrams as a representation of the corpus. Pre-processing techniques have been applied to the corpus transforming it into lowercase and removing punctuation, numbers and stop-words of the English dictionary. Lemmatization has been considered less intrusive and sufficient for this purpose than stemming from keeping the correct meaning of words and correct spelling. To lemmatize the free lexical database, Wordnet has been used that results in being one of the most used for the English language.

Python's NLTK library for natural language processing contains a method that returns bigrams, meaning words that are found in often one after the other, e.g., "original_series", "orginal_content", "stranger_things".

The fifty most common bigrams have been selected based on their frequency in the entire corpus and visualized with a network graph where each node is representative of a word, and the edge link this one to its partner.

This procedure has been applied to all the collected letters from 2014 to 2020 as a whole and in a second stage applied on selected periods to see differences in time (2014-2016, 2017-2019, 2020).

On the same documents, we ran a basic LDA to verify if the same topics were detected by a more performing algorithm. In order to do that, in the pre-processing phase, we have divided texts into sentences and then tokenised each sentence by word. Pos tagging has been applied to the lemmatised text to be able to select only those parts of the speech that were contributing to the understanding such as nouns and verbs. A vocabulary has been built, including bigrams and trigrams as well. All those words that were not presents in at least three documents were removed because not significant. This algorithm requires to enter a number of topics that we think it might accurately represent the corpora. In addition, two parameters need to be set up to communicate similarities between documents and words. Alpha controls per document topic distribution and Beta per topic word distribution: the first one is used to say that a document is likely to contain a mixture of most of the topics and the second one that every topic is likely to be described by the same words. We have therefore set a high value of Alpha since documents are likely to talk about the same topic and low value of beta since topics can be described but different words. Number of topics has been chosen accordingly to the number of topics found through the bigrams network.

In a second stage, the actual content of the offer was observed, analyzing a database containing all the TV shows and films on the platform from 2015 to 2019. For this task "netflix_titles.csv" database available on Keggle, an online community of data scientists, containing 6236 titles was used. It is structured as follows: “show_id”, “type”, “title”, “director”, “cast”, “country”, “date_added”, “release_year”, “rating”, “duration”, “listed_in”, “description”. All this data was collected through the website flixable.com. We manually added a further column indicating the corresponding quarter in which the content was entered into the platform.

A feature selection has been made before starting: “type”, “country”, “date added”, “listed_in,” and “description” columns were analyzed. The analysis focused on identifying the selection and variety of the proposed offer. Python's libraries panda, nltk and matplotlib were used: the database was transformed into a dataframe in order to manipulate data and apply functions more effectively.

To identify the genres added each year from 2016 to 2019, the most frequent words in the "listed-in" column have been visualized, after applying tokenization and lemmatization, this because each program can contain several sub-genres, for example, "International Movies, Sci-Fi & Fantasy, Thrillers" or "Dramas, Independent Movies, Romantic Movies ". In order to visualize results, a horizontal bar graph using the matplotlib library has been created. The same procedure has been applied to retrieve data regarding the country of productions.

Excel functions have then been used to create simple charts representing the evolution of the offer regarding the number of movies and tv shows in time.

The open-source library of Python Tweepy has been used to access the official Twitter API and collect tweets. With the free subscription, Twitter allows us to collect a limited number of tweets in the past seven days, searching for a query that is the text we are looking for in tweets. In our case, a query containing the word "Netflix" ran, and around five thousand tweets have been collected in the week from August the 17th to the 23rd, all in English.

Twitter is a text-based microblogging network that allows users to write posts up to 140 characters; By default, those tweets are visible to everyone without the need to give permissions to users to let them read, so that big networks of followers are easily created (Sun, 2019).

The choice to observe data coming from this social network derives from its power to impact word-of-mouth: it allows people to connect and share thoughts from everyplace and all devices; this length can be compared to newspapers' headlines, and for this reason, it is very catchy and makes it easy to consume offering immediate reactions and insights (Jansen, Zhang, Sobel and Chowdurry, 2009).

Tweets have been stored into a text file; Text data like tweets are a very heterogeneous and variable type of data that requires to be cleaned before to be analyzed. On all the entries, a series of cleaning techniques have been applied in order to create a tidy dataset removing what does not contribute to the content.

As a first step, all possible Html tags, links and mentions have been removed. Emoji have been encoded in UTF-8 and then removed. Contractions expanded, and acronyms substituted with their extended form comparing those with words inside two pre-created files containing all substitutes. Before checking if words were well-spelled, stop words, that are parts of the speech that do not contribute to understanding the meaning of the text, have been removed as well as numbers and punctuation and eventually performed a spell checking followed by the lemmatization of words.

The spell checker used in this research is a simple one, based on the Levenshtein Distance algorithm. It is used to compare the similarity between two words, counting how many elementary steps, such as replacement, deletion and insertion, need to be performed to transform a word into another one. The word that needs the least number of steps and therefore has the shortest distance is the one most likely to be the right one. In this case, every word is compared with an internal English dictionary.

Bigrams have been created based on term occurrence on Python, and through the software platform, Rapid Miner possible association rules have been looked inside texts with the FP-growth algorithm.

4.2. Survey

It has been conducted an online survey in order to answer the first research questions through Google form. The purpose is to look to the consumers' behavior closely and try to understand from them, which can be the reason for their actions. A quantitative approach based on the voice of the customer has been considered valid enough. Due to the emotional and motivational aspects of the topic, it has been considered easier to identify answers by listening directly to the voice of the consumer.

One hundred people filled the survey, most of them are students and young workers living in Europe, ranging between 18-35 years old, among them 53% were females, with different habits and tastes regarding entertainment. The scope was not to address neither only declared binge-watcher nor casual viewers but to have a general overview of customer behavior and motivations. This has been considered a valuable segment; indeed, 79% of individuals living in the EU are using the internet daily with a Social media penetration between internet users of 65.2% (Statista 2018).

As we can see from the chart (Appendix 5), everyone consumes at least one form of digital entertainment: TV shows and series are watched by 80% of respondents, followed by movies consumed by 60% of people and finally videogames that are played by only 13% of the sample. In terms of habits, they consume at least one form of entertainment in their daily routine (65%) (Appendix 6). The most used devices are laptop/computer and mobile/tablet chosen by 75% and 41% of the respondents, respectively (Appendix 7). Tv is used only by 27% of the sample, and this is due to the demographic nature of the sample characterized by students and young workers that likely do not have a stable residence yet. The freedom provided by those portable devices is the reason why they are among the most used.

4.3. Limitations

This research relies on methods that have major limitations.

Firstly, related to the conducted survey that has been distributed and filled voluntarily to a sample that does not pretend to be significantly representative of the total population. It can be

considered as a proxy of those parts of the population that watches TV series since 80% of the respondents answer to watch them. Besides, all the answers are subjective and related to the respondents' tastes and feelings. For the reasons above mentioned, the intent of this study is not to propose a generic truth, whereas addressing the problem taking as a sample a restricted group.

Similarly, to be able to conduct a text analysis with the available tool, a smaller number of tweets have been collected through methods that were open sources and free. The retrieved data, then, can be missing in some parts or not be completely representative of the whole web traffic. Due to the heterogeneity of data, a small sample has not been sufficient to conduct a proper analysis that would retrieve interesting patterns, and this has been left for future development.

5. Findings

From the fifty most quoted bigrams (Figure 13), we can identify six macro topics in Netflix's strategy. The first topic starts from the node "original", orange bubble, that has several ramifications creating bigrams with "feature", "film", "content"," language", "programming", "series," and "investment". Original productions received the biggest investments so far. Local content and non-English production can be identified as a second key point. As we can notice in the green bubbles, Netflix is currently investing much money in the production of local movies and TV series outside the United States and the United Kingdom, "La casa de Papel", "Dark," or "Call my agent" are some examples of Spanish, German and French productions.

Many are the most successful shows mentioned in the corpus, blue bubbles. From bigrams in purple bubbles, like "high quality", "Golden globe", "Academy award", we can understand that the strategy pursuits by Netflix are towards the creation of quality content that can position it among the cinema giants, competing with well-established production houses for the most prestigious prizes. The key genres are highlighted by the black bubbles: "action_film"," talk_show", "sci_fi", "kid_family", "comedy_series," or "drama_series" and "diverse_taste". The last identified topic is continuity, where indeed, words such as "second" and "third season" or returning season are mentioned to announce the launch of the following seasons.

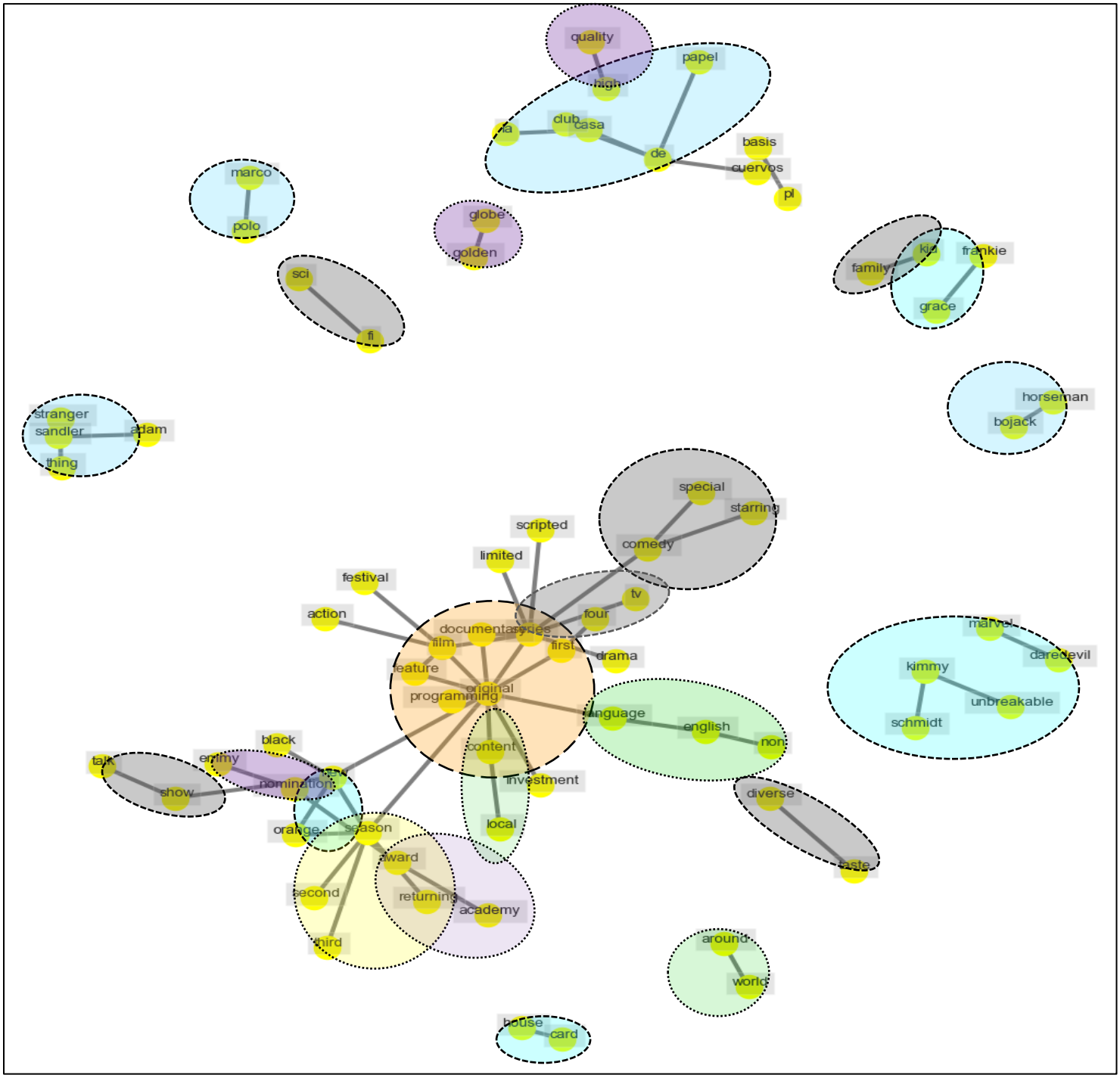


Figure 13 - Bi-grams of Letters to shareholders 2014-2020

Applying the same methodology to documents split into three groups by years, 2014-2016, 2017-2019-2020, two are the main differences from the analysis ran on the entire corpus: in the three-year from 2014-2016 “comedy series” is the most mentioned genre whereas it is no longer

mentioned as frequently in the period 2017-2019 where drama_series, limited_series and kid_series are the most quoted TV series genres. In addition, in the window 2014-2016, there is not yet a strong direction to the production of non-English series. This bigram is not visible anymore; instead, in the period 17-19, "spanish_language_content" is frequently mentioned as a demonstration of the development of more local content outside the US and the UK.

In the analysis of the two letters belonging to 2020, first and second quarter (Figure 14), there is a complete disruption of the previous pattern: indeed, we can notice a radical change in the network; there are no longer focuses on key genres, non-English productions or original content. This network represents the two letters published during the Coronavirus period; it is not surprising then that words such as "people", "shelter", "home", "hope," and "help" are among the most mentioned.

In the last six years, Netflix has always more centered its strategy on the production of originals, high-quality content and renewal of the most successful products; there has been a shift from comedies to dramas and content for children as well as a more local production in a non-English language.

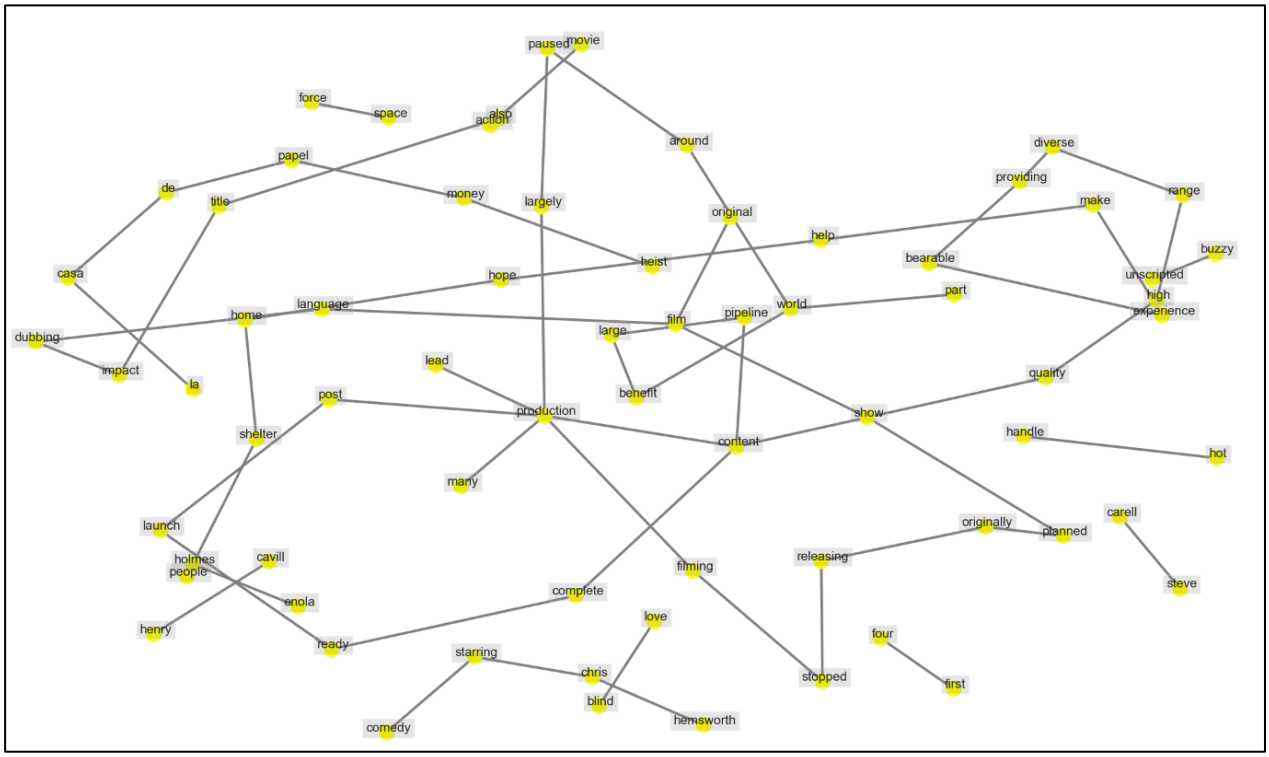


Figure 14 - Bigrams Letters to shareholder 2020

Applying a different methodology to the same set of documents we found the following results. The LDA algorithm, given a number of topics, shows the words that most represent that topic and then what are the most representative topics in the corpora.

By giving as number of topics five, we can observe results on Figure 15. Most of the words appear in more topics and it is difficult to identify a clear distinction between them. For example, in the first topic words such as “globally”, “international”, “market”, “increasingly” make us think about expansion, topic five seems to talk about key genres since we see words like “comedy”,”drama”, “talk_show”. However, topics two, three and four seem to be almost the same: “original content related to local productions in India or Mexico”.

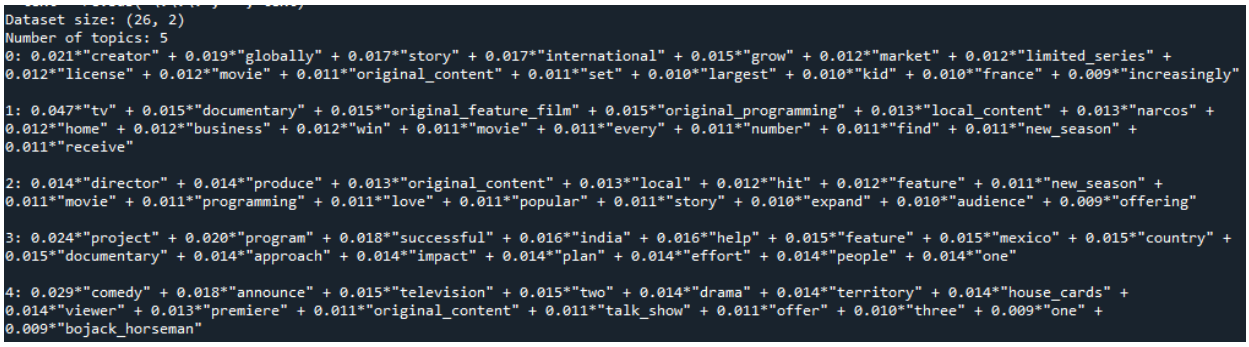


Figure 15- LDA algorithm result - N. Topics =5

We tried, therefore, to reduce number of topics to three to see if a more precise result was given.

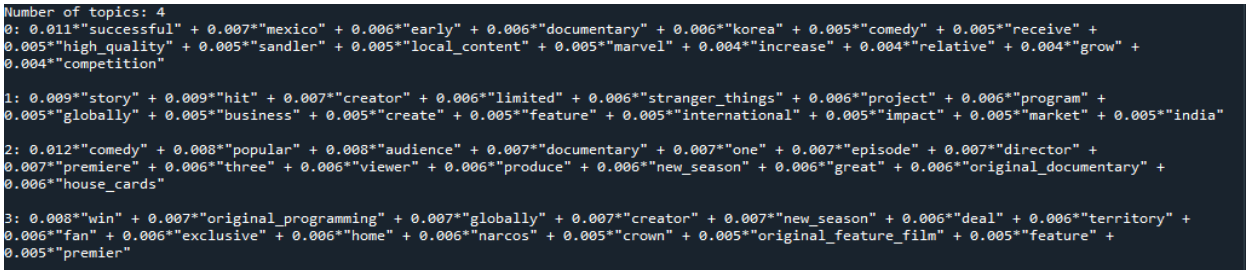


Figure 16- LDA algorithm result - N. Topics =4

With four topics, it seems that Topic 1, talks about “increasing successful high quality and local content against the competition”; Topic 2 “having an impact on international market such as India”; Topic 3 includes key genres and production of new seasons and finally Topic 4 seems to be a mixture of those mentioned above.

Last attempt was made with number of topics equal to three, as we can see in figure 17. In this case, the five macro themes discovered in the bi-grams networks seem to be identified however they not clearly distinct. In Topic one, investments in High - quality content and original content have been identified. Topic two, instead, is talking about expansion in new markets as well as

creation of new hits. In Topic three, we can find key genres and continuity with new seasons.

```
Number of topics: 3
0: 0.008*"original_programming" + 0.006*"nomination" + 0.006*"home" + 0.006*"business" + 0.005*"narcos" + 0.005*"crown" +
0.005*"language_original" + 0.005*"early" + 0.005*"comedy_special" + 0.005*"fan" + 0.005*"receive" + 0.005*"investment" + 0.005*"third" +
0.005*"total" + 0.005*"win"

1: 0.009*"hit" + 0.009*"story" + 0.007*"creator" + 0.006*"project" + 0.006*"program" + 0.006*"successful" + 0.006*"limited" +
0.005*"stranger_things" + 0.005*"base" + 0.005*"india" + 0.005*"globally" + 0.005*"country" + 0.005*"feature" + 0.005*"two" + 0.005*"france"

2: 0.011*"comedy" + 0.008*"audience" + 0.007*"popular" + 0.007*"documentary" + 0.006*"one" + 0.006*"three" + 0.006*"produce" + 0.006*"episode" +
0.006*"globally" + 0.006*"premiere" + 0.006*"territory" + 0.006*"director" + 0.006*"original_documentary" + 0.006*"new_season" + 0.006*"viewer"
```

Figure 17- LDA algorithm result - N. Topics=3

Playing on parameters value can help to create more distinct topics, however those texts are really homogeneous and most of the words belongs to more than one topic so that it becomes hard to distinguish them clearly.

If so far, we focused on what Netflix communicated, now we pass to analyze what has effectively changed in its catalog.

Looking at its evolution from the first quarter of 2016 to the fourth of 2019 (Figure 18), we see that the number of films on the platform is always greater than the number of TV shows with 4063 movie titles against 1860 TV show titles worldwide. In general, a pattern is identified in the distribution; there is a growth in the number of TV shows added every year from Q1 to Q4 and starting from 2018-Q1, there is a growth from 27% to 36% of the added TV-type content.

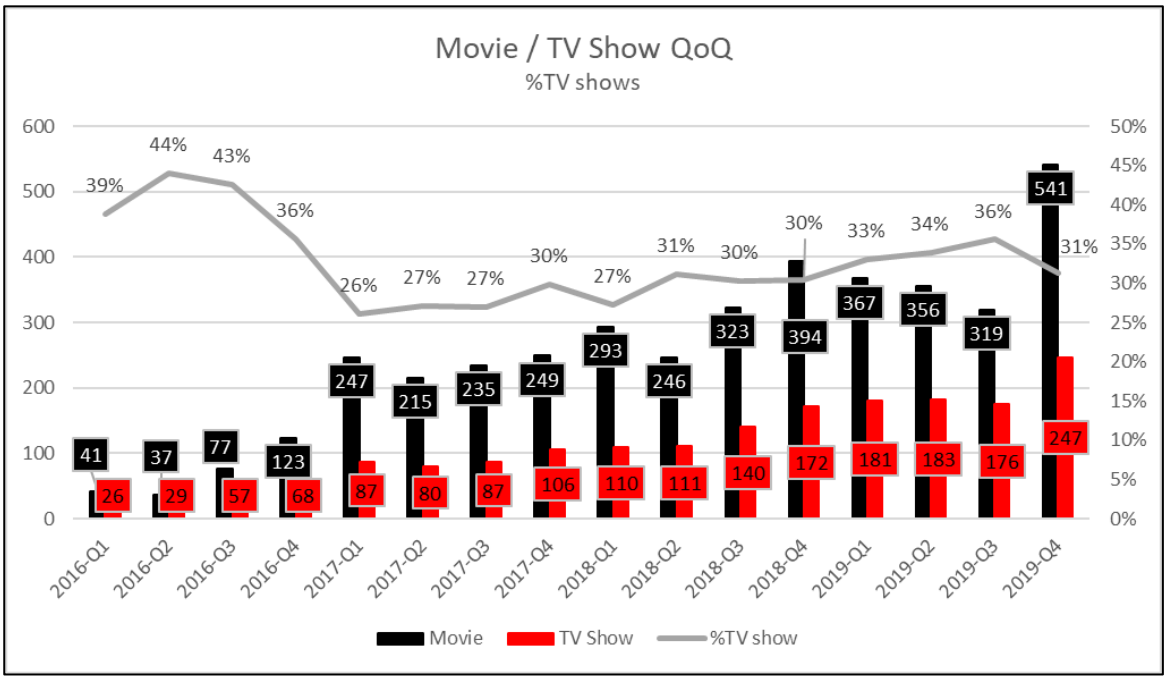


Figure 18 - Evolution of added titles from 2016 to 2019 by type

In the letters to the shareholders issued by Netflix, we have identified some macro-themes, originals, locals, non-English productions and a focus on genres such as comedy, drama and children.

Each of us has our favorite genres, and it often happens that the search for a program to watch is done through keywords such as genre; This is an important variable to consider when evaluating the offer. Each title in the database is characterized by two or three different genres; for this reason, it was necessary to apply some pre-processing techniques in order to capture the single genres. Titles without genre descriptions have not been considered.

Charts below (Figure 19 - 20) represent the fifteen most frequent genres. "International" is the most frequent genre in the title's description for all the four years considered.

As we previously discovered, Netflix announced a shift in the desire to produce drama rather than comedy content that has also been confirmed by its offer. Indeed, in 2018 we have an increase in the number of dramas added to the platform (64% more than the previous year) that is higher than the number of added comedies that results to be increased only by 45% from 2017. However, comedies are still the second most popular genre in the offer, followed by "romantic", "thriller" that alternate from year to year and finally documentaries. It is interesting to note that among the most frequent genres added every year, there are some: in 2016, we find the "LGBT" category, in 2017 "spanish_language", in 2018 "adventure" and in 2019 "horror".

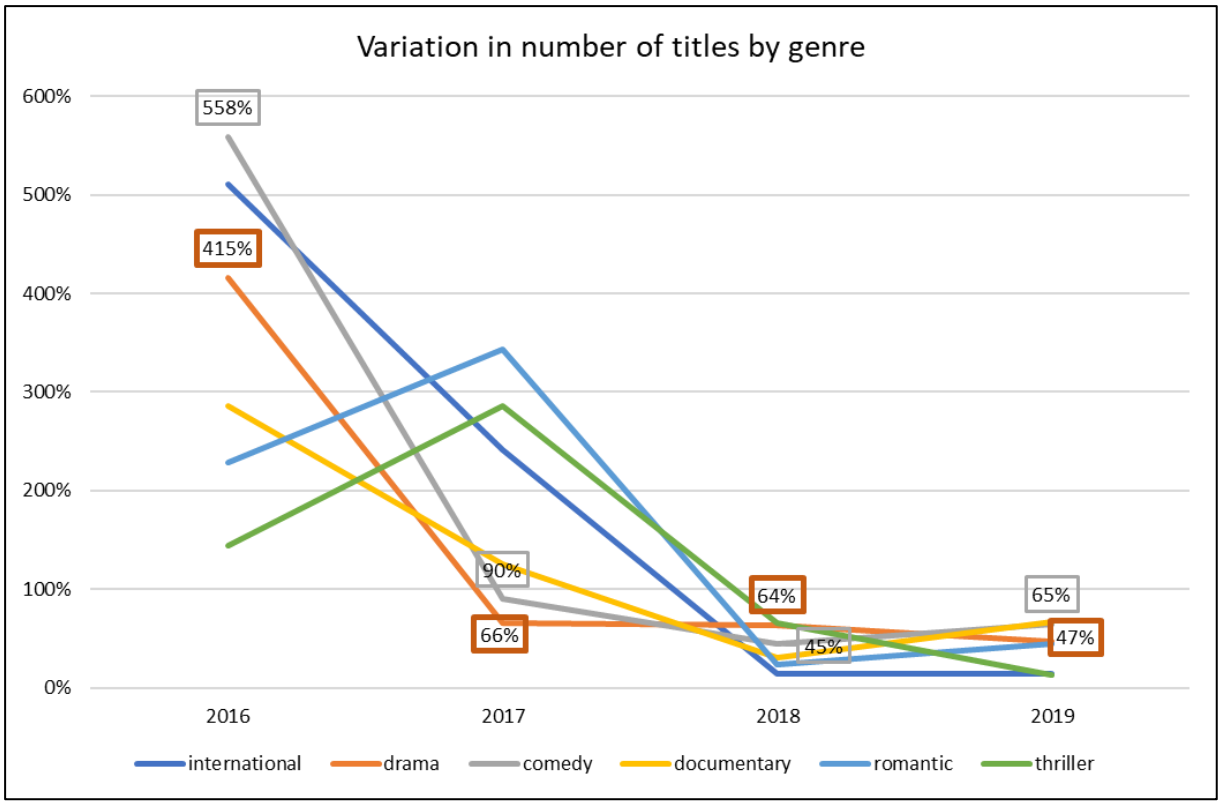


Figure 19 – Variation of top 5 genres added from 2016 to 2019 on Netflix

It is not surprising that the most present categories are drama and comedy: opposed to each other, they can be considered macro genres containing sub-categories such as "romantic", "thriller", "fantasy," etc. While it is true that everyone needs to laugh and to watch a good comedy is certainly a great way to start, it is also true that the production of comedies allows reaching a large audience. Comedies are a genre that in general is good for all age groups, gender and occasions, to see both with family and friends.

Obviously, there are nuances that can make this type of production controversial when they contain jokes that touch-sensitive topics such as race, sexual orientation, or simply use too vulgar language. However, in general, this type of content is usually an excellent investment for a production company because it is less segmented and therefore allows us to reach more people. At the opposite pole, we find the dramatic genre, which is generally much more serious than comedies. People watch this type of content because they want to be moved, they want to immerse themselves in a cathartic experience.

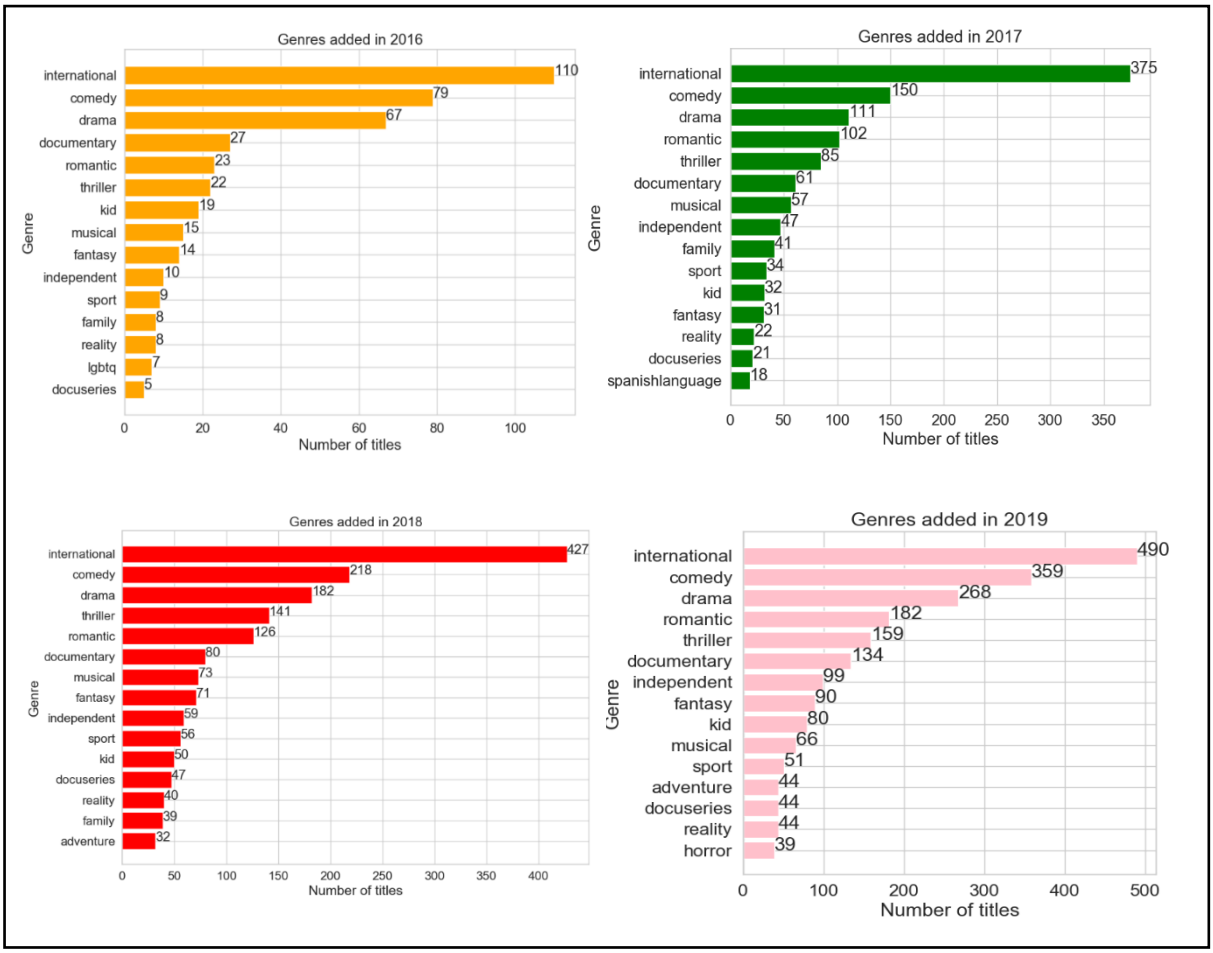


Figure 20 - Fifteen top genre added on Netflix from 2016 to 2019

If comedies are for everyone, we see that on the contrary, the dramatic genre is watched more by women and less by men, as confirmed by the data released by Statista in 2018 (Appendix 8).

As mentioned by Netflix, the production of non-English language content is part of its growth and expansion strategy in Europe and Asia. Investing in local productions not only helps to widen the selection and variety of the catalog but above all, it helps to become competitive and to fight domestic competitors: in fact, these play a dual role by attracting the American / English audience with contents that can be considered "exotic" and at the same time capturing the attention of the local public, looking for more familiar things different from the usual American productions.

From figure 21, we see that number of English productions is still much bigger than other locals, but not-English content is increasing its presence in the catalog despite all this.

Looking more precisely, we see that the biggest productions outside the United States and the United Kingdom are coming from India, Japan, Spain, Mexico and China (Figure 22).

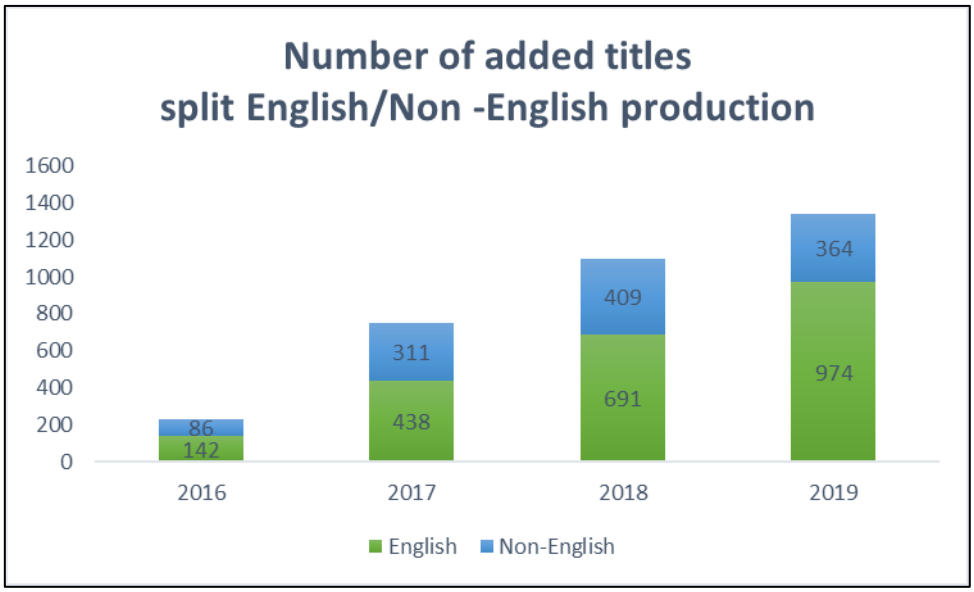
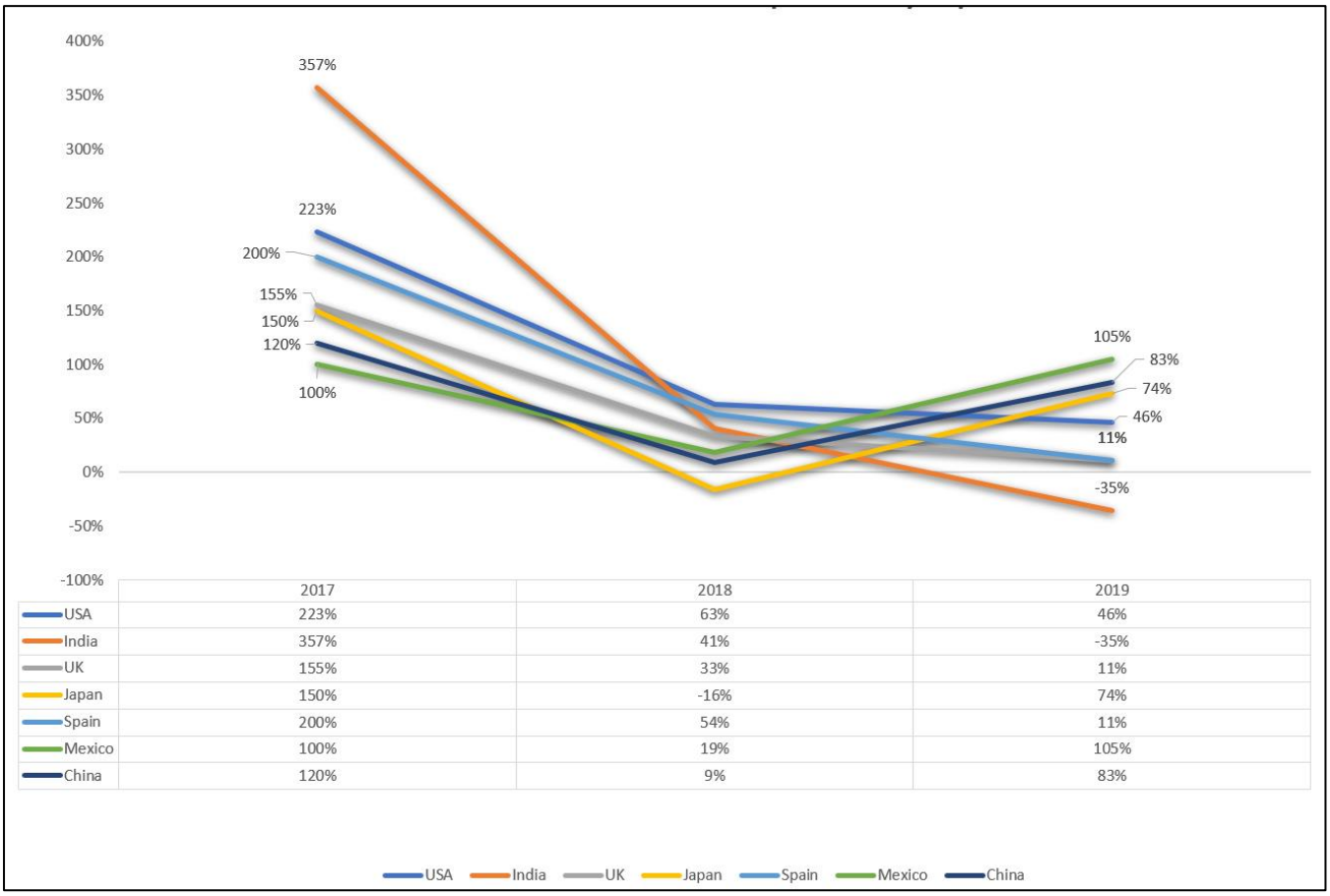


Figure 21 - English/ Non-English added titles from 2016 to 2019

If the United States and Europe show to have a flat growth in the number of users in the following five years, India and Mexico show a steeper line (Appendix 13) that can justify the bigger number of titles added in the last four years. India is one of the few countries where Netflix is not the leader of the sector due to the big regional content released by Amazon Prime Video that allows the tech giant to reach a bigger base of customers. In Japan, there is a big gap between Netflix that owes 35% of the market and its competitors; therefore, offering local content is a good strategy to keep its position in the Japanese market (Appendix 14). China is considered an isolated case since there the American platform is not available. However, as reported by the USA annual Flow report, the Chinese population is the second most present ethnic group in the United States after the Mexicans and has a strong presence also in Southeast Asia such as Indonesia, Singapore, Malaysia and Thailand which record a potential strong growth in the sector as well as a strong Netflix presence.



	2017	2018	2019
USA	223%	63%	46%
India	357%	41%	-35%
UK	155%	33%	11%
Japan	150%	-16%	74%
Spain	200%	54%	11%
Mexico	100%	19%	105%
China	120%	9%	83%

Figure 22- Variation in number of added titles per country of production from 2017 to 2019

As shown by numerous researches mentioned in the literature review, the role of the consumer and also of social networks is becoming increasingly significant in this industry, and it is even influencing its success. In this research, we would have liked to measure the level of engagement that Netflix produces with its users to evaluate its marketing strategy and its performances in the retention process. However, this type of analysis requires more knowledge and suitable tools, for this reason, it has been left for further studies, and an overview of social media content has been performed instead.

Tweets containing the text "Netflix" have been collected to carry out an analysis regarding the opinion on the web. Having collected around five thousand tweets in one week, we applied techniques of text mining to retrieve results. Using RapidMiner, association rules have been looked for. However, due to the heterogeneity of the text and the small sample, no rules have been obtained. This has been then considered a possible development for future research. However, to observe something, bigrams based on frequency have been built. Tweets have been divided according to the posting date in two giving the results shown in figure 23 (a, b). What

we can notice is that the most recurrent bigrams are talking about the current themes related to the last news. Indeed, at the time of the collection of these tweets, two main events have been identified: Netflix has been accused to “disgusting as it sexualizes an eleven-year-old for the viewing pleasure of paedophiles and also negatively influences our children” as reported by Forbes for the marketing campaign of its new release, the French production “Cuties”.

	biGram	count
0	('frenchsenegalese', 'filmmaker')	574
1	('destroyed', 'career')	510
2	('brndnstrssng', 'might')	509
3	('might', 'destroyed')	509
4	('career', 'frenchsenegalese')	509
5	('filmmaker', 'marketed')	509
6	('marketed', 'film')	509
7	('film', 'commentary')	509
8	('video', 'game')	341
9	('commentary', 'onrt')	324
10	('doubleduzit', 'black')	237
11	('black', 'dude')	237
12	('dude', 'created')	237
13	('created', 'individual')	237
14	('individual', 'video')	237
15	('cuties', 'movie')	226
16	('movie', 'cuties')	210
17	('black', 'woman')	208
18	('french', 'senegalese')	202
19	('senegalese', 'black')	201
20	('miggsboson', 'cuties')	191
21	('movie', 'research')	191
22	('research', 'director')	191
23	('director', 'french')	191
24	('woman', 'pull')	191
25	('year', 'old')	163
26	('sexualizing', 'child')	161
27	('eleven', 'year')	157
28	('new', 'movie')	131
29	('child', 'pornography')	127
30	('signed', 'petition')	121

	biGram	count
0	('million', 'dollar')	2082
1	('michelle', 'obama')	1039
2	('twelve', 'million')	1039
3	('book', 'deal')	1039
4	('deal', 'deal')	1039
5	('obama', 'sitting')	1038
6	('sitting', 'twelve')	1038
7	('dollar', 'mansion')	1038
8	('mansion', 'sixty')	1038
9	('sixty', 'million')	1038
10	('dollar', 'book')	1038
11	('jevonwilliamsvi', 'michelle')	1037
12	('deal', 'esti')	1037
13	('avatar', 'last')	344
14	('bounty', 'hunter')	253
15	('teenage', 'bounty')	250
16	('umbrella', 'academy')	228
17	('season', 'two')	202
18	('alone', 'loshme')	192
19	('loshme', 'reeebtpms')	192
20	('reeebtpms', 'wnettyrhitd')	192
21	('last', 'airbender')	190
22	('tomhopperhops', 'justinhmin')	178
23	('project', 'power')	168
24	('han', 'zimmer')	167
25	('lightskinpainx2', 'cmon')	163
26	('cmon', 'happen')	163
27	('happen', 'tcoji2sjfmxg4')	163
28	('agni', 'kai')	156
29	('rocketboiart', 'avatar')	155
30	('last', 'airbenber')	155

Figure 23 - Bigrams based on Tweets posted on August 18-20 (a) and August 21-23 (b)

In Figure 23.b, instead, we can see that the most mentioned bigrams are talking about Michelle Obama since Netflix just announced the release of a documentary about her.

We are not able to observe something significant from tweets due to the small number: it is a too dynamic service always in evolution, it is influenced and affected by events, therefore, it is difficult to find a general opinion as we can see for products since Netflix’s product is content that changes over time.

Despite the missing clear results but convinced about the power of social media, a three-fold survey has been conducted to listen closely to the customer's voice in order to have a direct proof that can help us understand which strategies to adopt to meet his needs.

In the first introductive part we gather insights with questions regarding what kind of content they want to watch to compare with the current Netflix content strategy; In the second and third section, the purpose was to analyze two main behaviors, namely the influence exercised by others that affects the decision-making process and the willingness to share opinions with peers on social media as well as with our closer social ties.

The 100 participants of the survey had to answer some questions related to what they watch and their preferences, how they make a choice when they decide to watch new content and if they are willing to share their opinions or not when they are satisfied or disappointed.

TV series have a particular structure completely different from movies: they are made by several episodes of shorter duration compared with the length of a movie but globally they last longer, the plot is usually more complex, and they are characterized by the extreme use of cliff-hangers to keep spectator’s attention and assure that he will watch the following episode. We can record different approaches in how people consume TV series because they require a more considerable amount of time, but also the information to process is more significant in number and complexity.

73% of respondents state they watch more Tv shows in comparison with movies that are selected only by the remaining 27%. Among the reasons that lead to this choice, the duration of the content seems to be the favorite motivation followed by the type of plot and story that is proposed. Being able to comment on social media does not seem to be good motivation in choosing to watch a film, instead, 10% of the sample, who previously stated to watch more TV shows, seems to choose it because it is easier to share and comment on social media. Finally, not having to search for something new every time seems to be another reason to watch the TV series.

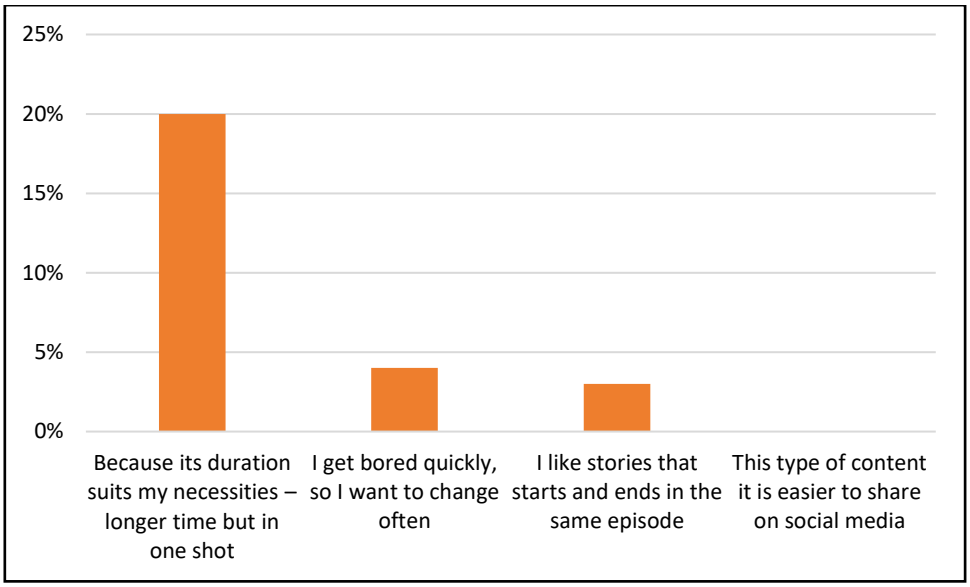


Figure 24 – Survey: "Why do you prefer to watch movies?"

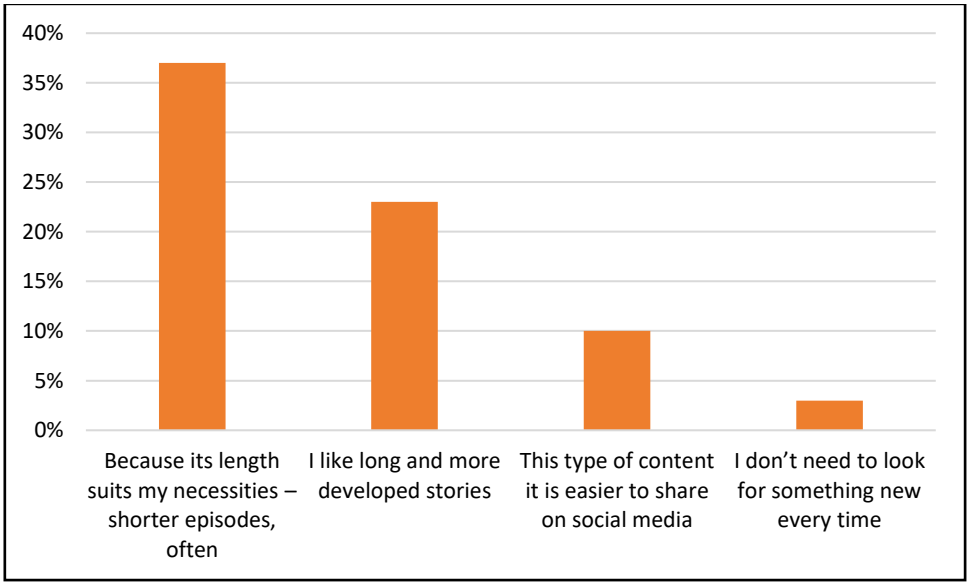


Figure 25 - Survey: "Why do you prefer to watch TV Show?"

Given the results obtained in analyzing Netflix's strategy, we wanted to investigate furtherly whether our sample agrees on the choices made. The entertainment sector is heavily dominated by American or more generally English-speaking productions, with little space left for local productions on the international scene. Streaming platforms have very little national content; thus, it is less frequently watched. It is therefore not surprising that 87% of respondents said that most of the content they watch is of English production. Among the answers regarding the motivations for choosing content in English rather than one in their language we find both quality and quantity: in fact, about 60% say that either they do not find anything interesting or

indeed the content produced in the United States is of superior quality, while 35% simply do not find available local content sufficient.

To the question "would you watch more local content if available?" we see a positive trend towards local productions, which makes us think that it could be a good strategy for expansion and acquisition. (Appendix 17)

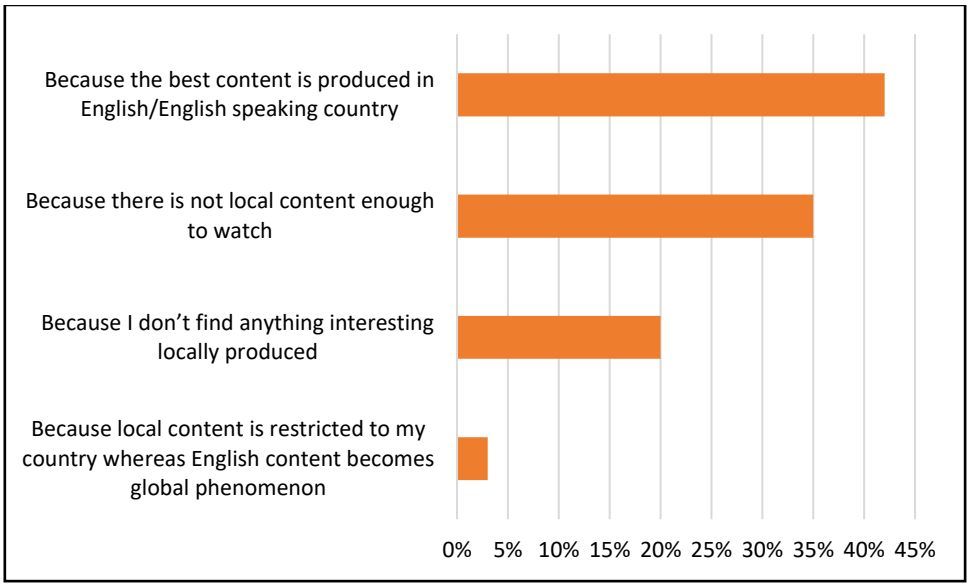


Figure 26 - Survey: "Why do you watch English productions?"

With the variety and multiplicity of contents to which we are exposed, the choice of what to watch becomes more and more complicated. We often spend more time in search of inspiration than in the actual time spent in consuming the product. It is therefore reasonable that filters are used to skim the list of possibilities and make a choice that conforms to our tastes. There are many sources from which we can pull information to make a choice: advertising, web pages and among the most common as mentioned above, there is undoubtedly word-of-mouth both among friends and via social networks.

The first element we considered is popularity. The fact that a series is on everyone's lips and that it could be at the center of conversations between friends or on the web, in fact, seems to be important for 40% of the sample but among them, only 5% consider it the discriminant to watch a show; 28% do not consider it an essential factor and the remaining part of the sample assigns to popularity importance of 3 on a scale of 1 to 5, a sign that is not a decisive criterion but that is not entirely irrelevant. (Appendix 20)

Answers to the question, "*Which of the following factors may influence your choice*?" (Figure 27) show that family and friends' experiences are primary drivers in the decision-making process followed by official critics and social media discussions. These can be associated to filters that allow the user both to save time and to have feedback before deciding if to watch the content of a product which, as defined in previous chapters, is considered to belong to the category of experience-based products and therefore they can only be tested after being consumed. Only 1% of the respondents say that his/her choices are not affected by any external factors.

Nevertheless, how effectively the decision to watch something can change due to the opinion of others? Is there a distinction between positive or negative opinions? Or, in case the instinct is contrary to the common belief, how does the sample behave and what decides?

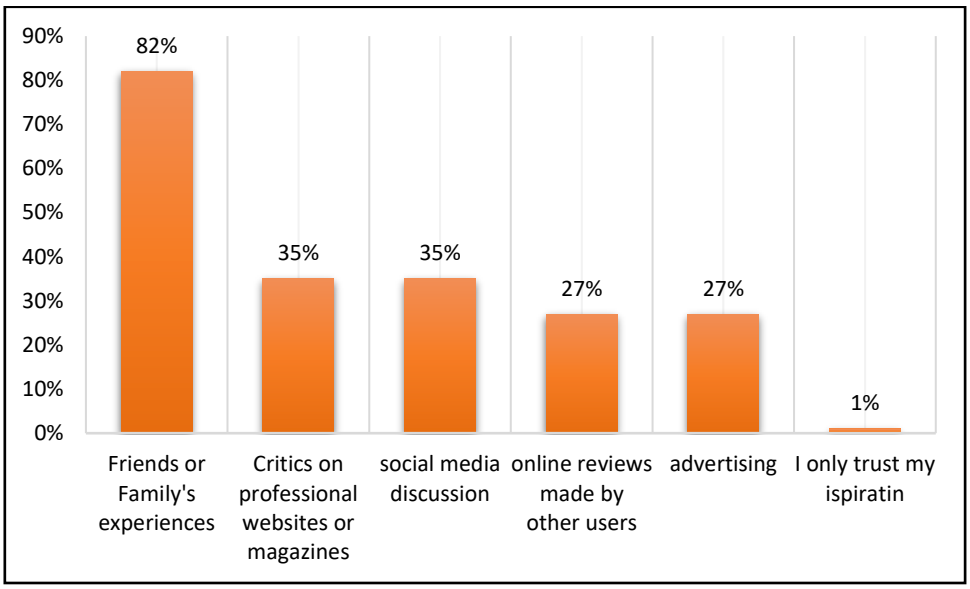


Figure 27: Survey - "Which of the following factors may influence your choice?"

The following questions have been asked, "*Can your choices be influenced by a negative review or a positive review?*" (Figure 28) and "*Do you think that your choices about what to watch may change more after having read a review/opinion?"* (Figure 29) recording these results:

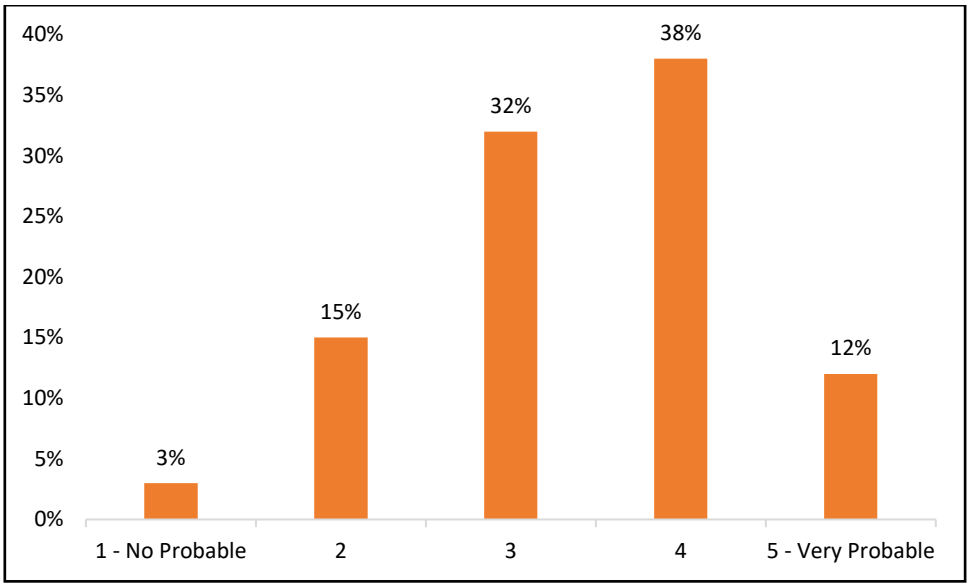


Figure 28: Survey - "Can your choices be influenced by a negative review or a positive review?"

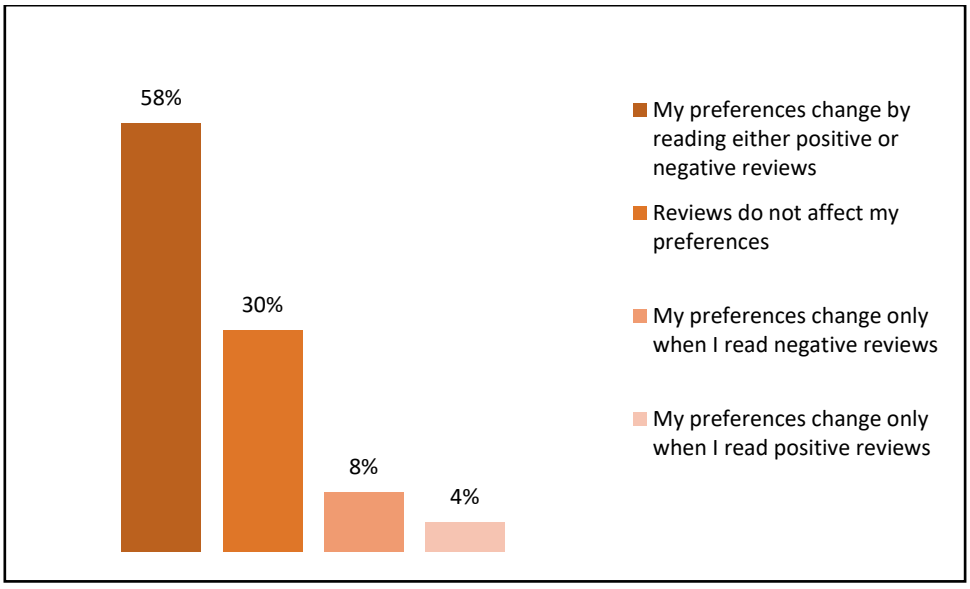


Figure 29: Survey - "Do you think that your choices about what to watch may change more after having read a review/opinion?"

As charts show for 50% of respondents is very probable that their choice can be affected by reading a review/opinion. For 58%, this can be either positive or negative. A 30% state "reviews do not affect my preferences," but still 47% is considering that if not sure at least probable. Only 12% of them give a precise answer saying that their choices are more influenced by positive (4%) or negative (8%) reviews.

Putting the respondents at a crossroads, choosing between their personal instinct and sensations and the advice of others, we can notice that the answer changes accordingly the already existing

intention they have. The personal feeling towards a TV series or movie, after watching its trailer or advertising, is, in general, the first driver. It seems that a bad opinion can hardly change the customer's mind if he/she is already convinced that he/she wants to test the content personally. However, on the contrary, if they have a negative idea regarding the series but the reviews and comments are positive, 44% reply that they would probably watch it and 10% are convinced that they would.

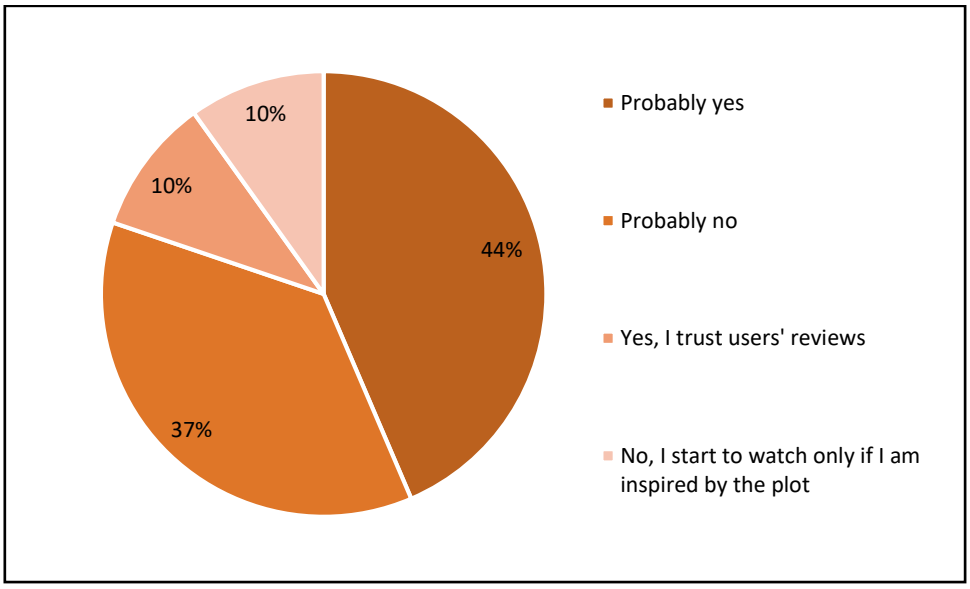


Figure 30: Survey - "Would you watch a series with excellent reviews/comments but with trailers/advertisements that do not inspire you?"

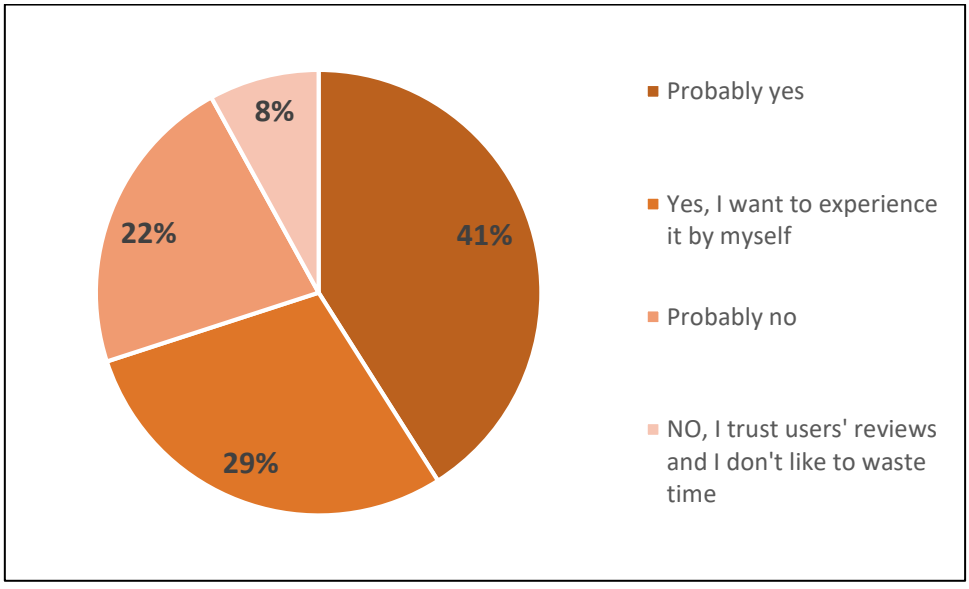


Figure 31: Survey - "Would you watch a series with bad reviews/comments but a catchy advertisement/trailer?"

Having obtained these answers from the selected sample, we see that WOM has an effect when it comes to deciding if to start new content or not and, therefore, in the process of acquiring and converting customers.

"*How do you choose what to watch*?" gives us an idea about the inclination of people to seek advice. Results (Figure 32) show that friends' advice is the most considered tool during the selection process followed by advertising 47% and personalized suggestions made by streaming platforms. Social network discussions are considered by 34% of the viewers that seem to be keener to rely on close social tie or official communication tools.

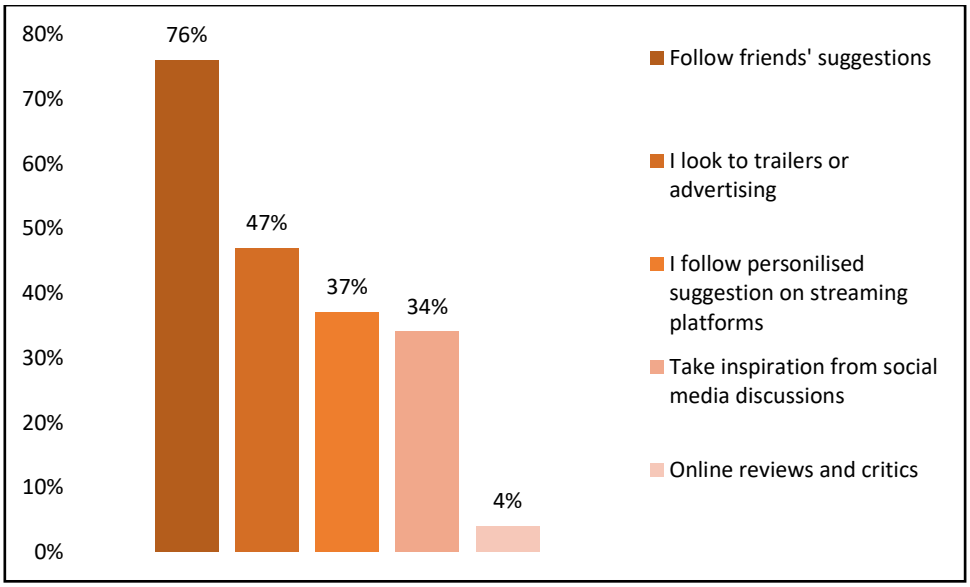


Figure 32: Survey - "How do you choose what to watch?"

If on the one hand, we have found that the opinion of other users, close or unknown, has a positive influence, the next step is to understand if the consumer is inclined to share his experience, how and if he decides to do it what are the reasons that guide this choice. The sample then answered the following questions. "*When you watch a TV series, who would you like to share your opinions with?* (Figure 33) "*Do you contribute to the social network or other channels for what concern entertainment?*" (Figure 34) and "*For which reason would you share opinions?*" (Figure 35).

From the analysis of the previous answers, it is therefore not surprising that 97% of the sample confirmed that they share their experience with family and friends. In comparison, only 25% is also open to communication on social networks.

However, there are many ways in which it is possible to communicate on social networks both in a passive way, by merely "like" pages, posts or other content or in a more active way by writing reviews, creating memes, sharing videos and images or others.

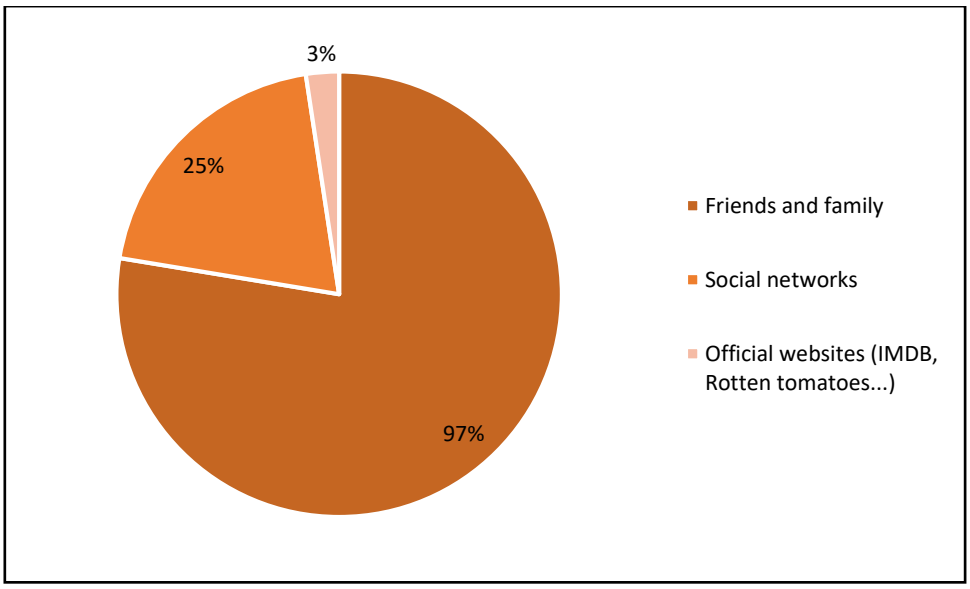


Figure 33: "When you watch a TV series, who would you like to share your opinions with?"

As we can see from the histogram in Figure 34, likes are the way users communicate their opinion more often: given the multiplicity of feelings that can be expressed through a single click (fun, love, anger, sadness) it is easy to understand that the simplicity which content can be promoted on social networks without the need to be active is the most used way. However, a simple like is a potent and handy tool in creating word of mouth on the web and widening the visibility of the content thanks to the continuous development of the recommendation algorithms that underlie the network. So, it is not necessary to openly communicate opinions in words because to affect the popularity and visibility of content. The 18% say they write comments, share posts or even create their content such as memes, images which meaning is used differently from its original context.

38% follow official pages and 11% cinema magazines. Original contents from the production companies such as trailers, interviews and advertising are considered in the informational process as well as critics reviews because perceived as being more reliable sources because based on recognized expertise.

Following fan pages (16%) is a new way of gathering with people with the same interests and creating online communities where it is easy to find fellows and freely express yourself.

In the literary review, some motivations that have been identified as drives in consumers' choice of spreading word-of-mouth have been considered. Respondents of the survey state that having confrontation talking about plot, characters and future development is one of the reasons why they start to talk about TV series and inform others about own experiences, feeling a sentiment of usefulness is also a valuable driver when they think which can be their motivations to contribute to the decision-making process of others. However, entertainment has been selected as a driver by the most of people that still think that watching a TV series is only a portion of the experience that is complete when they engage on social media and read comments, watch fan-made videos and memes and can forward them with friends, creating real word-of-mouth that spread from person to person.

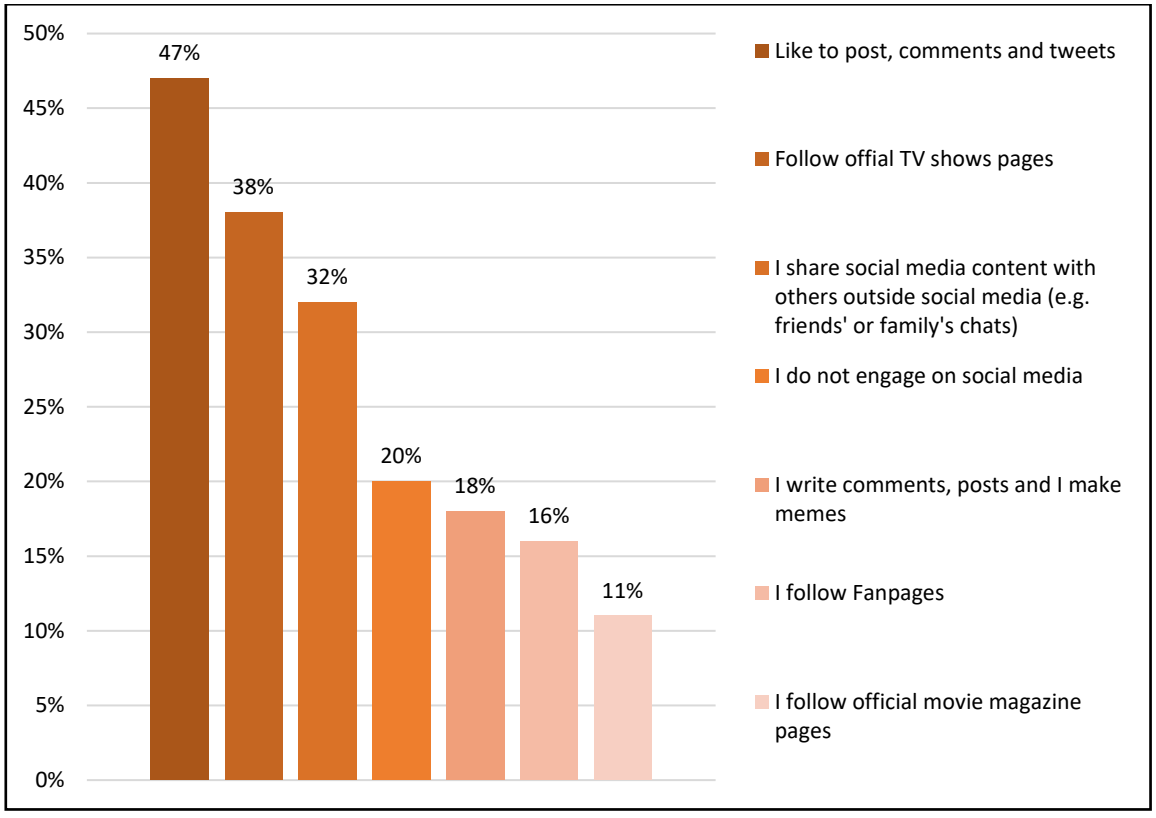


Figure 34: Survey: "Do you contribute to the social network or other channels for what concern entertainment?"

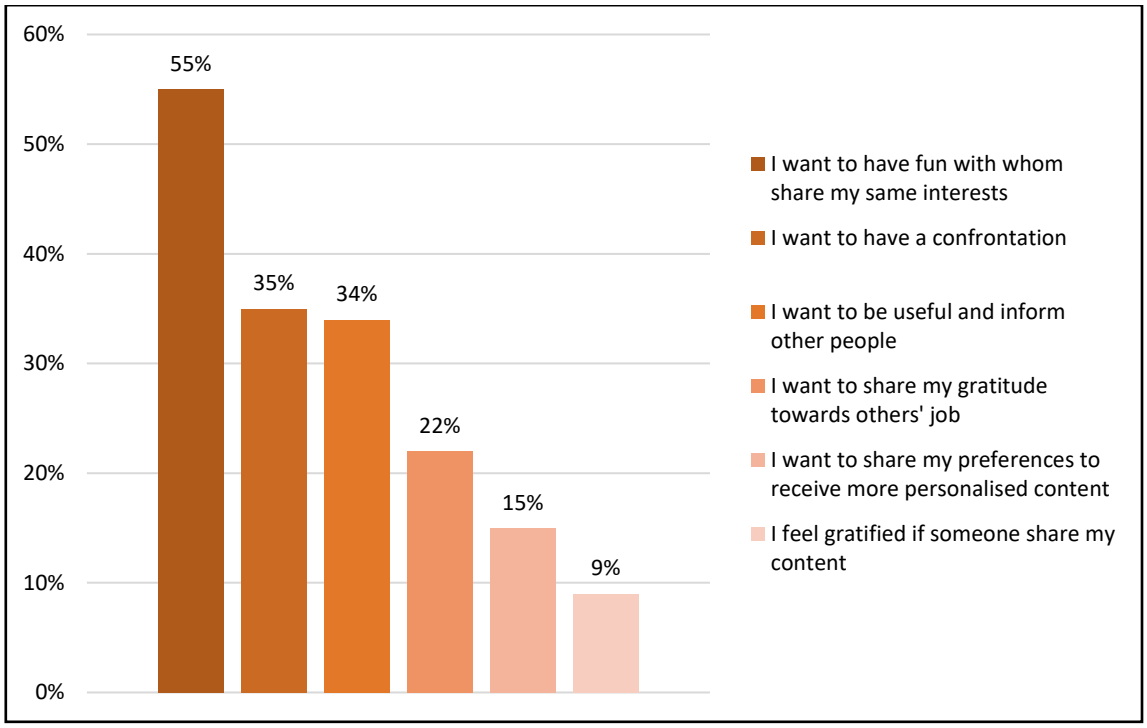


Figure 35: Survey - "For which reason would you share opinions?"

Looking at the type of comments that people are willing to leave, there is no difference between positive or negative for the sample: indeed, 57%, more than half of them state that they do not make any difference, and they write both positive and negative reviews indistinctly. However, 24% prefer to share positive experiences rather than negative; only 7% write when he/she is not satisfied with his/her experience. (Appendix 19).

With the results obtained, in the next section, a general comment will be provided to lead to recommendations and conclusions on how to acquire and, above all, keep customers.

6. Recommendations and conclusion

Data showed that the on-the-demand streaming market that revolutionized the entertainment industry is growing rapidly and sees more and more new players enter competing to acquire more subscribers. Researches, mentioned above, show that in the coming years the number of users is destined to saturate with a consequent increase in average spending for those already subscribed to this service since each customer will be subscribed to more than one service at the same time, even though the streaming war has started.

Due to this new concept of spending much time, more precisely binge-watching, consumers are immersed in a fictional world that had the power to increase their engagement with stories as well as increasing their will to keep doing this on another screen and with a broader audience of fans. Social networks become the perfect means that allow users to comment, create their user-generated content and also influence other people, creating phenomena of Word-of-mouth. This can be exploited by SVoD companies to acquire and retain customers and increase their presence.

To define a strategy, we analyzed the current leader of this sector, Netflix, and we tried to answer questions related to its strategy.

Results showed that Netflix is actively acting aware of the threat of its competitors. The bigrams network has allowed us to discover in more detail main key points such as investments in the production of own content against the loss of licensed content, in addition to the renewal of the hit series to continue to satisfy customers already registered. Non-English-language films and TV series to broaden the membership base outside the United States, in countries where there is a great opportunity to grow. India, Japan, Mexico, China and Spain seem to be the countries where most of the titles added derive, not in Anglo-Saxon language. To position itself among the greatest of the show business, it creates high-quality content to obtain the most prestigious awards. The catalog shows that there is a wide selection both in type, film or TV series and in genres with a preponderance of comedy and drama genres. These two genres are generalist and contain other sub-genres but able to address the tastes of every type of audience.

From the tweets' analysis, we have not been able to understand what is the public opinion on its service, however by looking for the word "Netflix", we saw that it is continually quoted in comments regarding its strategic choices.

Entertainment is a dynamic sector and always at the center of news and criticism: with the advent of social networks, this phenomenon is furtherly emphasized. Therefore, it becomes useful as a means to test audience reaction and to understand the trends from which to draw inspiration.

Researches conducted and mentioned in the literature review have been confirmed by the sample of respondents to the survey, telling us that nowadays, the opinion of other people matters more and more, also influencing entertainment choices.

With these premises, three areas of potential development that can be used to acquire but above all to retain customers have been identified. These also lend themselves as starting points for future research and further analysis.

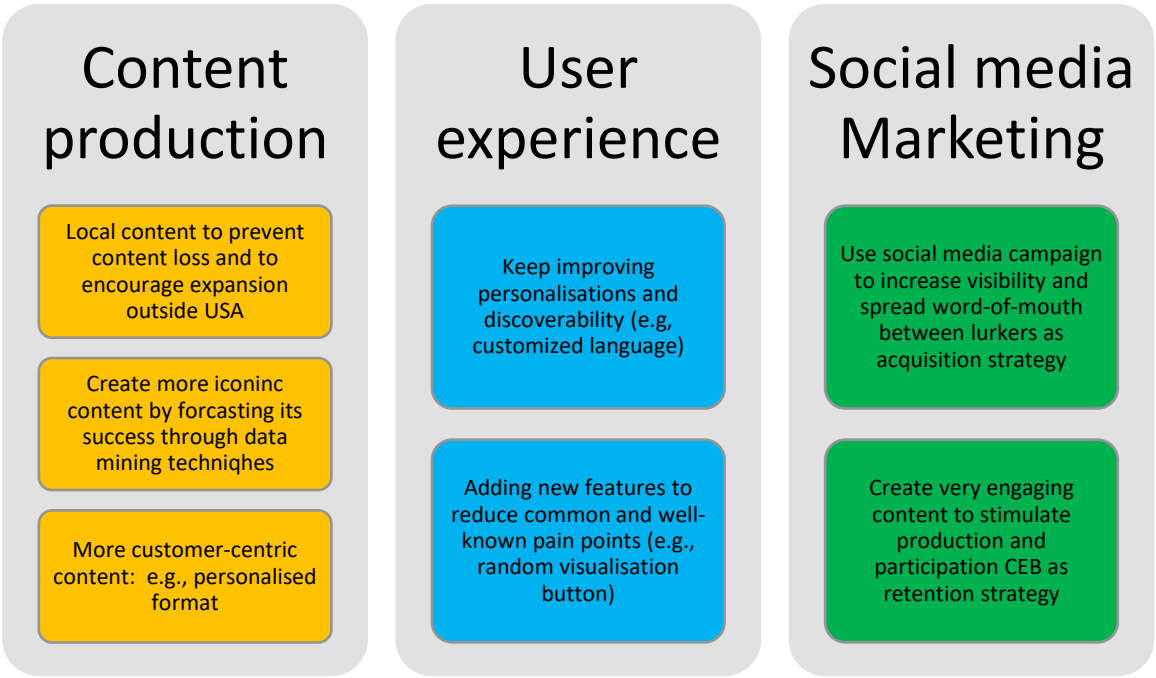


Figure 36 - Areas of potential development

6.1. Content production

Netflix is considered as a content "aggregator" by offering both original and third-party content. The latter can be found on other platforms as well, the originals are exclusive and available only on their platform. For this reason, they are a valid element in the strategy of acquiring new

customers who, curious about a new product, will then decide to sign up, at least, for the free trial of the streaming service.

This seems to take part in the expansion strategy in other markets that are not yet saturated and therefore allow to expand the subscriber base. Local content has a positive impact not only for expansion purposes, but it is also appealing for already subscribed customers that appreciate the variety in the offer selection.

If this, then, seems to answer to two fundamental questions, how to maintain a diverse catalog after the loss of third party products and how to acquire new customers in a saturating market thus helping expansion, it nevertheless proves to be insufficient for the following reasons: the first, as mentioned above, revenue coming from the original contents alone is currently not sufficient to cover production costs (Figure 12), and above all, success is always an unknown factor for this product, thus making investments always risky. Second, a single successful product is not enough to acquire customers who can find alternative ways of accessing it such as piracy which is still an extremely widespread phenomenon and third, often users are not even aware of the existence of certain films or series, thus ending up in look at third-party products that are more popular.

Creating successful series is, therefore, necessary but complex, in this data mining can be exploited to understand what customers are looking for in their visual experience, what factors can be of impact to make the content a successful content and therefore predict the final result.

By analyzing historical data from review websites, social media but also internal data on the number of views, prediction models can be created to determine success: there are numerous variables that can be used, among these, we have actors, locations, language, genre, type, duration that can be exploited in regression models or clustering techniques to find the most impactful combinations and predict success based on the criteria to be met, notably the taste of the public and financial constraints.

By exploiting internal data, we can also try to create formats that meet the audience’s demand: for example, from the survey, we understood that duration seems to be very important. Therefore, analyzing which are the most viewed formats (e.g., if they are long films or if the same long film is watched broken into several sessions, if instead, we prefer short episodes to fill the gaps in the breaks of the day) can help to create products that are truly customer-centric.

Social media and people's propensity to share help immensely in collecting data to carry analysis on successful trends and particular tastes of the audience. This can be used to create

compelling storylines of particularly genres with actors who are currently in vogue and therefore attract audiences. It seems essential to create products that can generate WOM: not just TV series and films that reflect the audience, but they need to become icons.

6.2. User Experience

When it comes to choosing between multiple equivalent services, the service and the experience offered to the customer become essential to retain customers' loyalty. The web services market revolves around the creation of interfaces that are user friendly and, therefore, easily adoptable and understandable by everyone: we try to reduce the friction that can be encountered when the user is not a tech-savvy to allow him to meet his expectations.

Simple, clear graphics that allow us to discover the content easily is necessary: one of the pain points complained by Amazon Prime Video users, when the service was launched, was precise that it was difficult to navigate the website and find the content.

Netflix has proven to be the best in this, being able to create simple, eye-catching and personalized graphics. Personalization has become the key to any successful online service. Famous became the Prize challenge launched by Netflix in 2009, where it awarded a million dollars for those who were able to write the best recommendation algorithm.

Being able to find the content that suits our tastes, moods, and available time is always very difficult; we often spend more time looking for inspiration rather than the actual time spent in front of the screen. For this reason, an analysis of our watching history is used to suggest what we might want to watch or what we would like to watch. Netflix offers a particular feature giving the customer the possibility to express his opinion by clicking on the thumb up to help the algorithm to customize our profile. This collaborative algorithm is based on the general idea that if a user A has watched film 1 and series 2 then user B, if he watches series 2 with a certain threshold of probability, he will love to watch film 1. Similarly, he proposes a ranking of the ten most viewed programs in the locality where we live to allow us to be updated on trending content.

Netflix already uses data analysis techniques extensively to improve the experience of its customers; however, there are still many ways in which data can be exploited, for example, by introducing new features.

Although numerous improvements have already been made on the discoverability of content, a pain point still seems to be the time spent deciding what to watch. A possible solution identified

would therefore be the creation of a button that would allow the random playback of content as we see in music streaming services; By analyzing the users' historical data, it would be possible to offer them two options whether to watch something similar to their tastes or allow them to discover something new that therefore does not match what they have watched up to that moment.

Another innovation that would be interesting to analyze further is linked to the way we consume content: the phenomenon of watching series in the original language is increasingly common even in those countries where the tradition of dubbing is extremely strong. We could then analyze the habits of the user and looking to some variables like genre, type and language try to understand in which language he might like to watch a movie and then to show him directly the program in the language. This brings to eliminate the passage through the settings to change the language or less invasive, ask directly to the customer at the beginning if he wishes to continue viewing in the customized-selected language or in another one.

To keep customers subscribed to the service, it is therefore important to be able to offer them a frictionless, pleasant experience that is customized to their needs.

6.3. Social media marketing

Social media marketing is a type of marketing that has developed in recent years and has become increasingly essential to be able to communicate and interact directly with customers.

The type of entertainment offered by Netflix encourages people to engage even furtherly, beyond the platform, on social media. Social media marketing needs to be exploited in two main ways: social media campaigns can improve visibility and attract the attention of those customers that have been defined as “lurkers” that do not explicitly engage but simply use likes. In this way, we increase awareness reaching a wider audience. A second and more powerful way passes through direct interaction with customers: the purpose is to increase “participation” and “production” CEB that results from having a direct impact on the number of visualization. If, in previous researched the strength of the social media engagement has been studied to see the impact on the box-office, it would be interesting to analyze the same factor applied to the number of subscriptions of a streaming platform.

Netflix currently records more than 57 million followers on Facebook, 15 million on Instagram and 6 million on Twitter, ranking as the most followed among the SVoD platforms on social

networks. The key to its success is the ability to engage with its customers by creating posts, videos and images that allow interaction and, at the same time, focus on the content.

Entire marketing campaigns take place on social networks and are hugely successful. Jokes, quizzes and small games help to maintain a high level of interaction with the public, who therefore enjoy answering. The social pages of the platform publish frequently and appear to have high levels of response, which therefore help to maintain a solid foundation in the relationship with the customer.

While exploiting social media helps to develop word-of-mouth, create iconic content that stimulates people's interest, on the other hand, they also become unstoppable sources of data.

We can collect audience data that we want to use as a target, evaluate the popularity of actors and directors to try to involve the most suitable for the role and then carry out sentiment analysis to evaluate the actual response.

Natural language processing techniques suits well for the analysis of unstructured data such as those coming from social networks but require much work to be able to have a solid dataset as well as the need to be able to collect large amounts of data to have results that are statistically relevant.

Both in the acquisition and in the retention phase, social media marketing plays an important role because it serves to reach all the targets and create buzz phenomena that stimulate interest and at the same time allow continuous interaction with users who create groups and communities around the program.

6.4. Conclusion

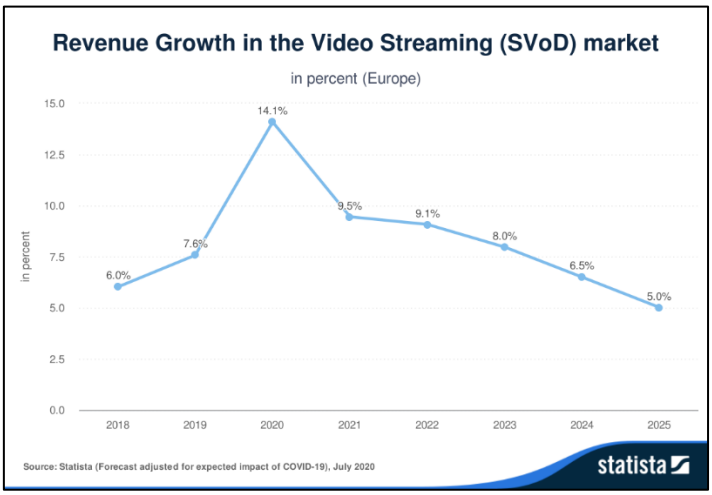
In this document, an introduction to the Streaming Video On-Demand sector was given to the reader to show how this has revolutionized the entertainment industry by giving the consumer new opportunities for consumption. In the same way, the way the consumer approach to entertainment has also changed: the more possibilities and inputs available make it difficult to choose among the numerous programs and therefore seek advice for inspiration.

In this, technology and above all techniques of data mining and text mining play a fundamental role since this type of product is not only content but also experience. Customization becomes increasingly key to distinguish from the others.

After analyzing the Netflix industry leader, three macro-areas of further development have been identified that allow us to further advance in giving the customer a better product: content production, user experience and social media marketing. These show opportunities for future research such as competitor analysis, adoption of more specific data mining techniques for predicting program success or failure, A / B testing of particular features to improve the user experience and comparative analysis of two or more streaming services and their method of interaction on social networks to measure the impact on the consumer.

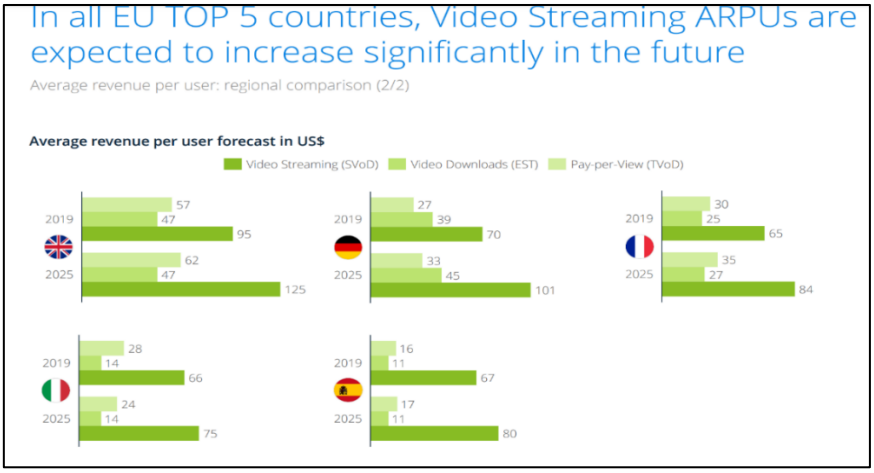
7. Appendix

– *Appendix 1- Revenue Growth in the Video Streaming Market Europe*



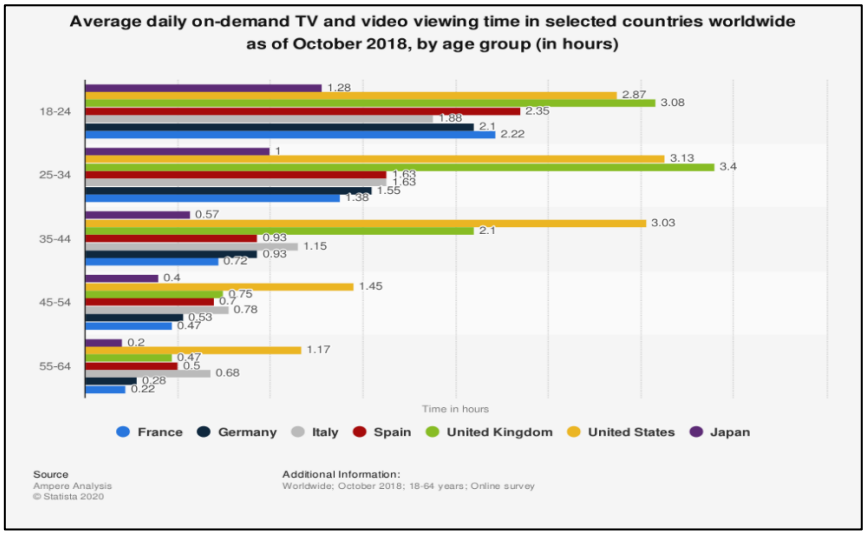
Source: Statista

– *Appendix 2 - Average revenue per user forecast in EU5*



Source: Statista

– *Appendix 3 - Average daily viewing time by age group in hours*



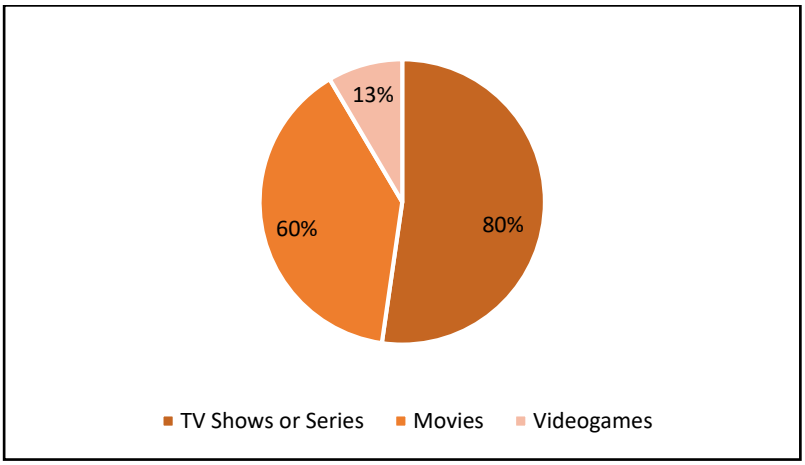
Source: Statista

– *Appendix 4 - Ranking the most viewed TV shows in the US in 2018*

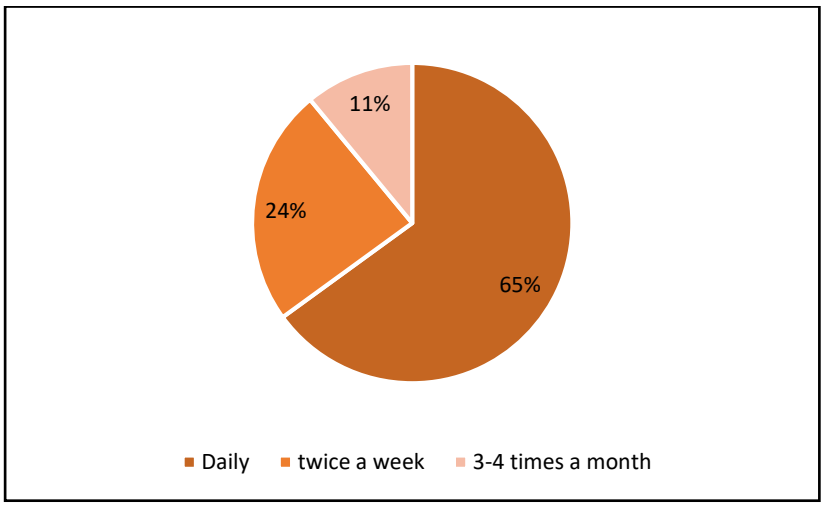
RANK	TITLE	STUDIO
1	THE OFFICE (U.S.)	NBC
2	CHILLING ADVENTURES OF SABRINA	NETFLIX
3	FRIENDS	Warner
4	GREY'S ANATOMY	ABC
5	HOUSE OF CARDS	NETFLIX
6	THE GREAT BRITISH BAKING SHOW	NETFLIX
7	MARVEL'S DAREDEVIL	NETFLIX
8	NARCOS: MEXICO	NETFLIX
9	THE HAUNTING OF HILL HOUSE	NETFLIX
10	CRIMINAL MINDS	CBS

Source: 7ParkData

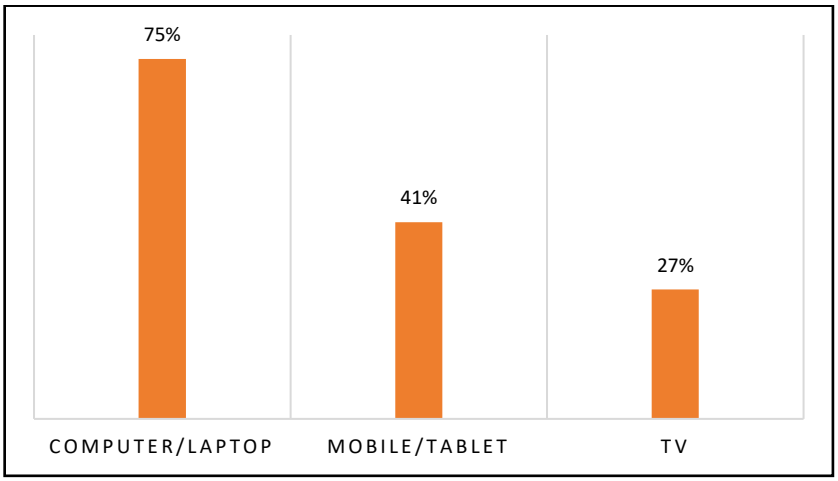
– *Appendix 5 - Survey. "Which is the form of entertainment do you consume the most?"*



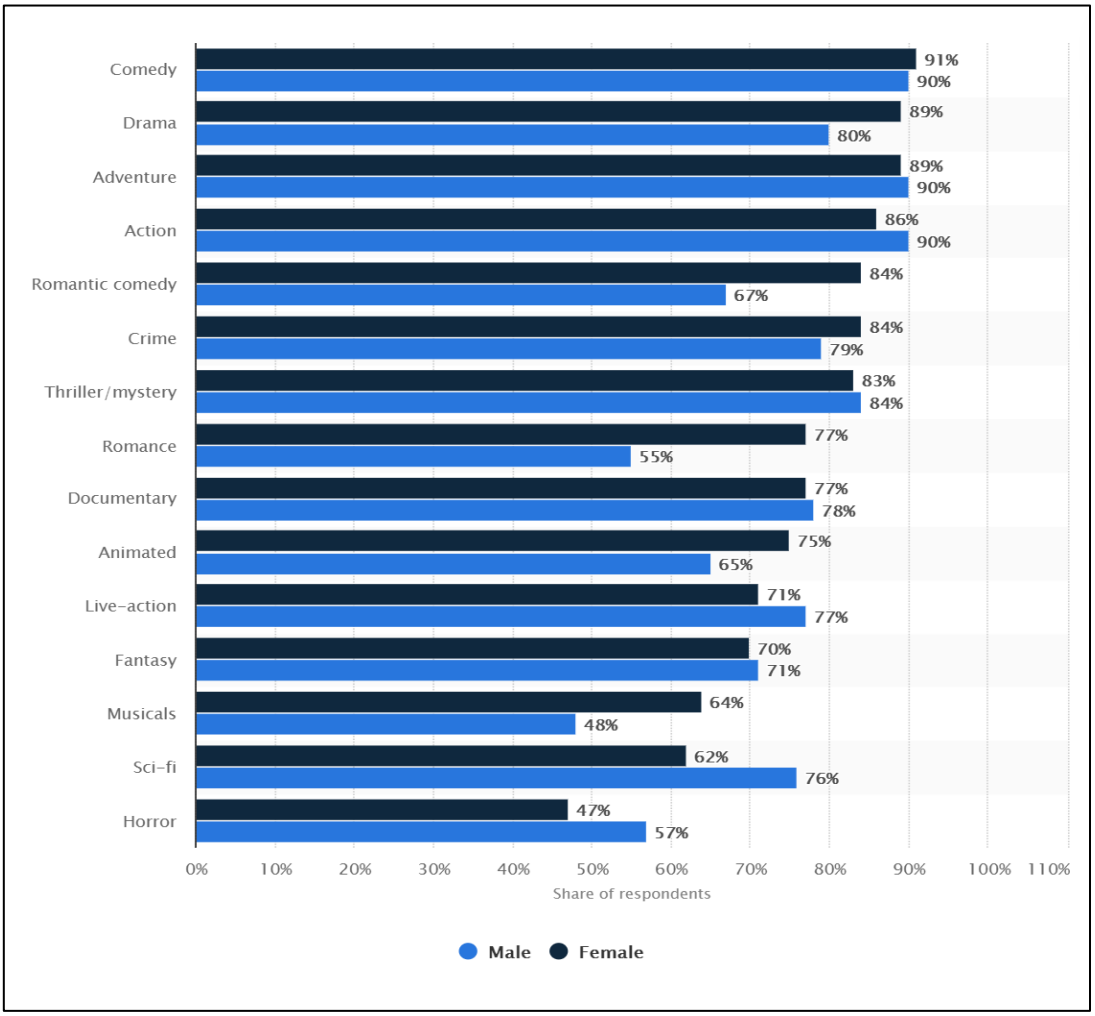
– *Appendix 6 - Survey. "How often do you consume the chosen form of entertainment?"*



– *Appendix 7- Survey. "Which device do you use to consume the selected form of entertainment?"*

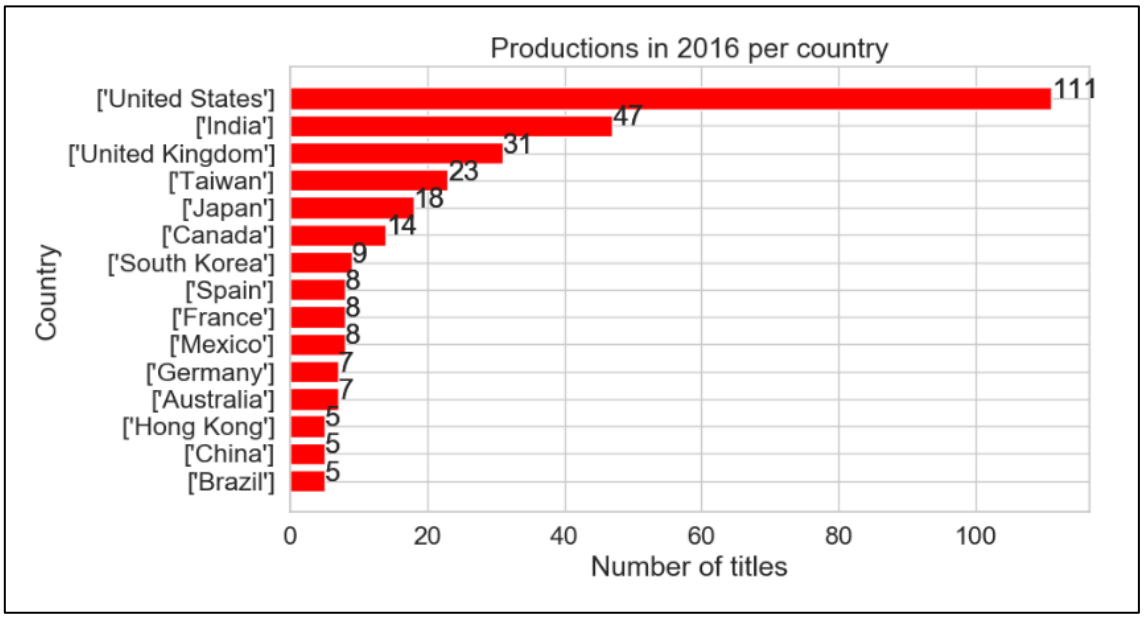


– *Appendix 8 - Most popular movie genres among adults in the USA, 2018 by gender*

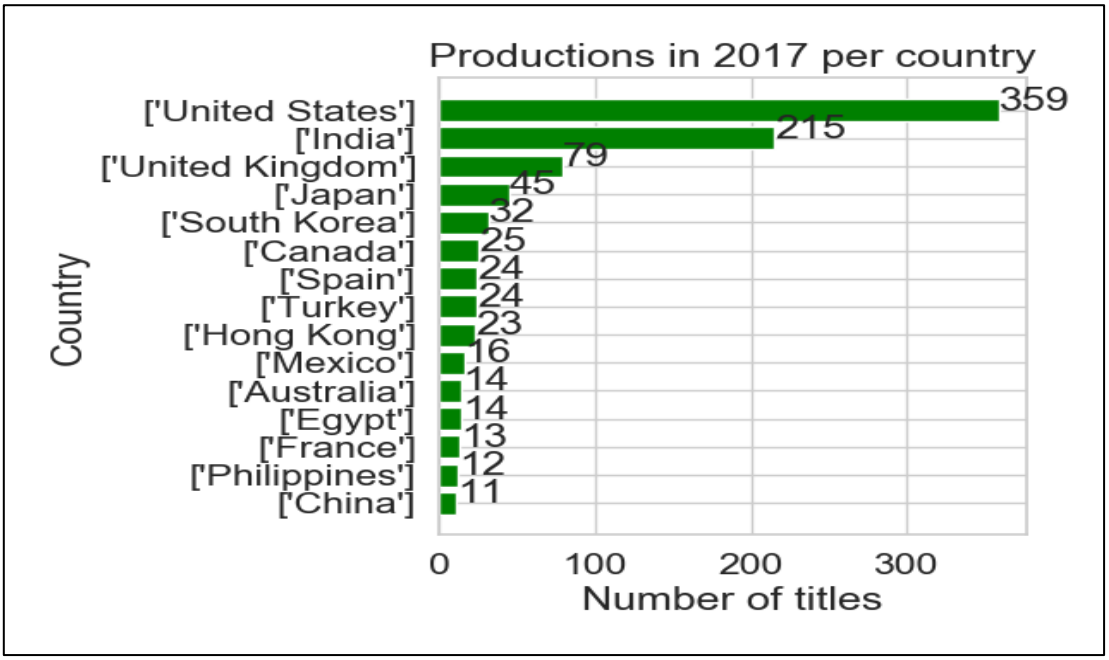


Source: Statista

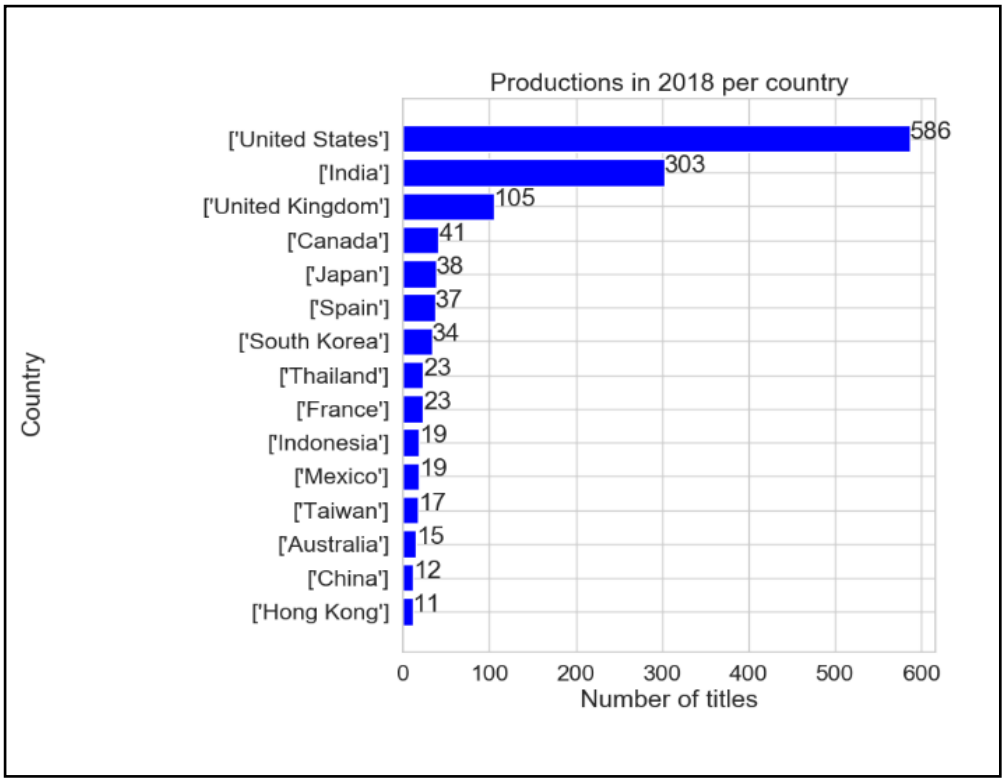
- *Appendix 9 - Titles added in 2016 per producer country*



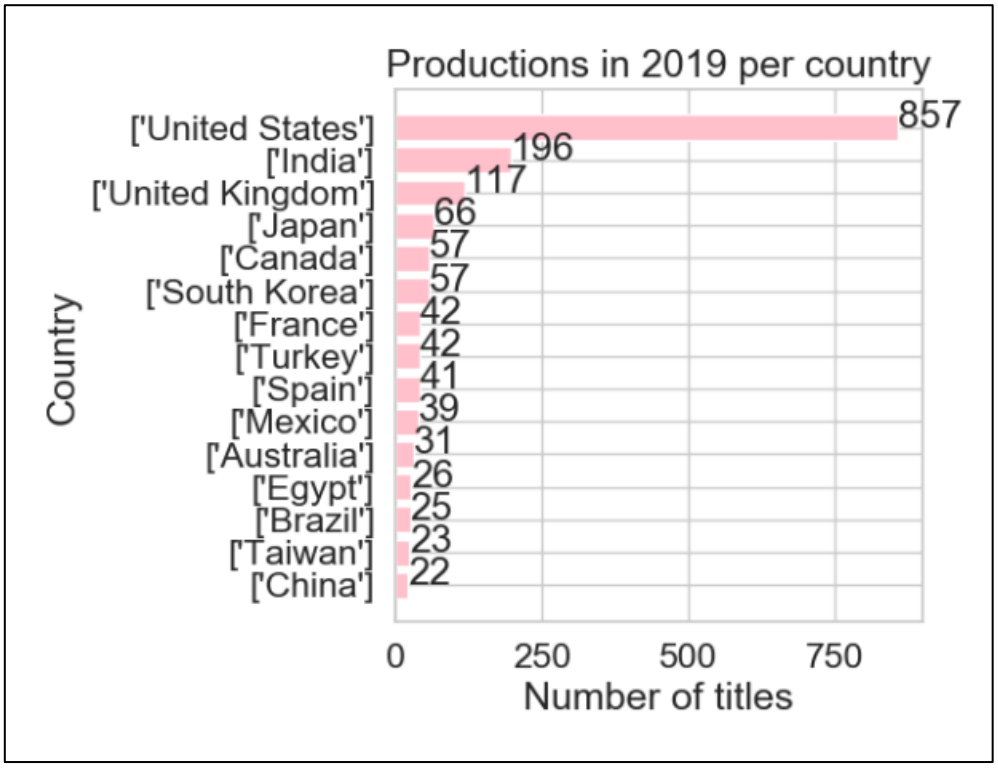
- *Appendix 10 - Titles added in 2017 per producer country*



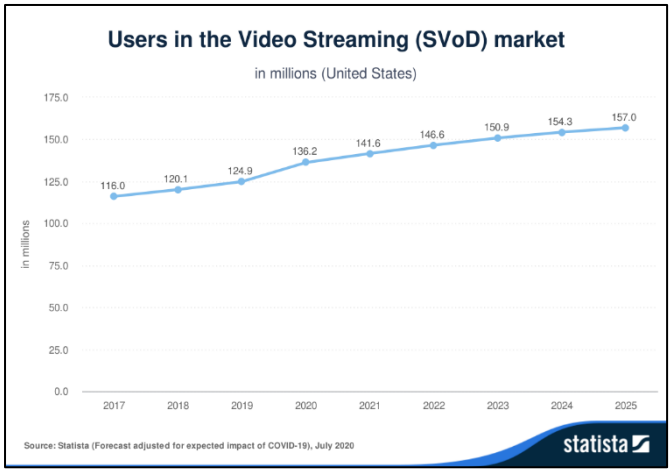
- *Appendix 11 - Titles added in 2018 per producer country*

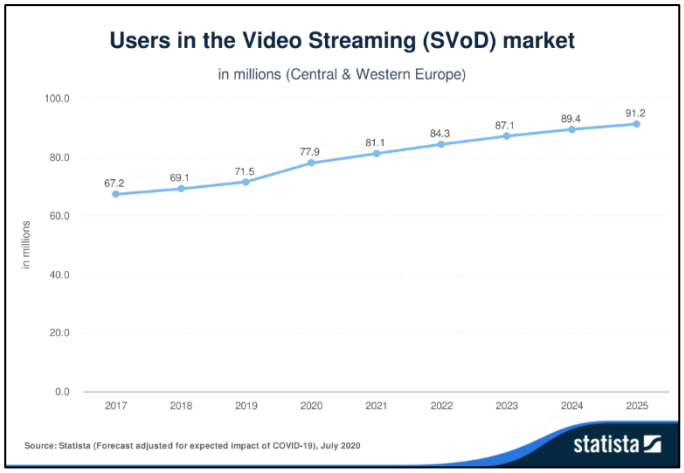


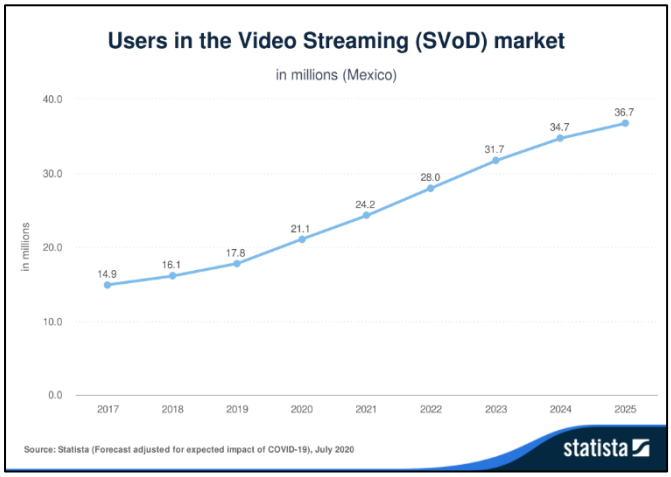
- *Appendix 12 - Titles added in 2019 per producer country*

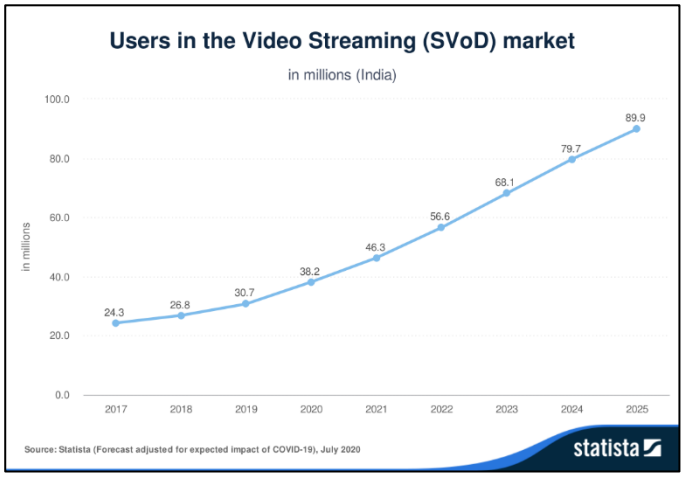


- *Appendix 13 - Growth in number of SVOD Users*



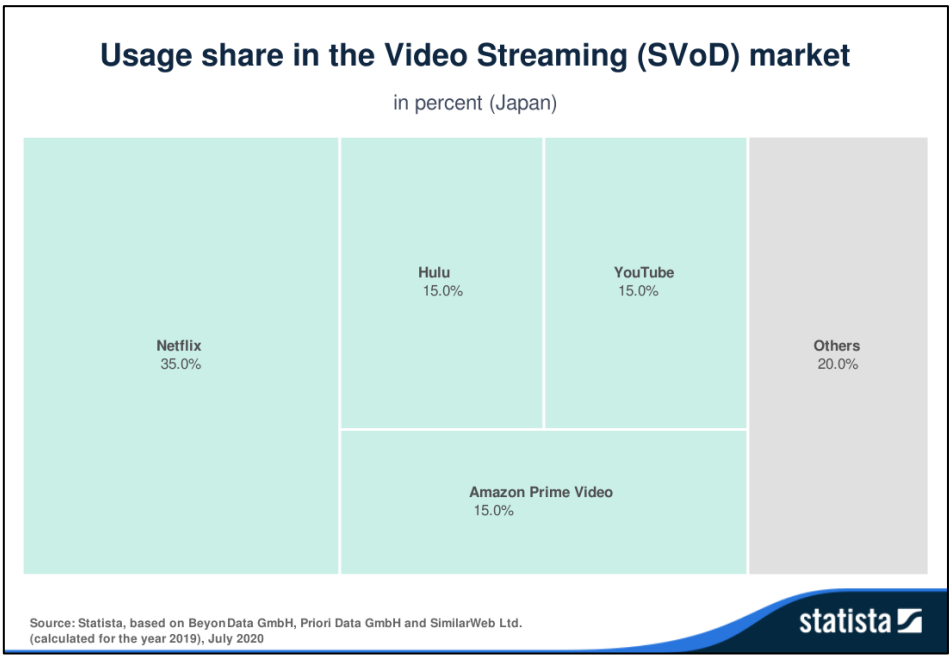






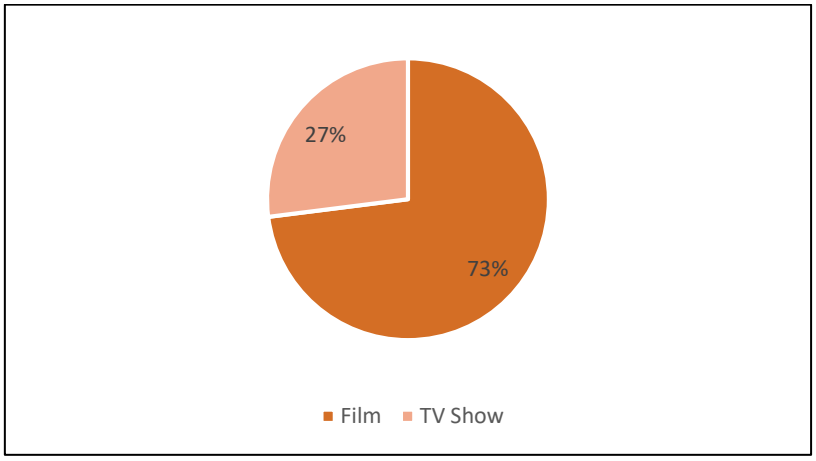
Source: Statista

- *Appendix 14 - Japan Market of SVOD*

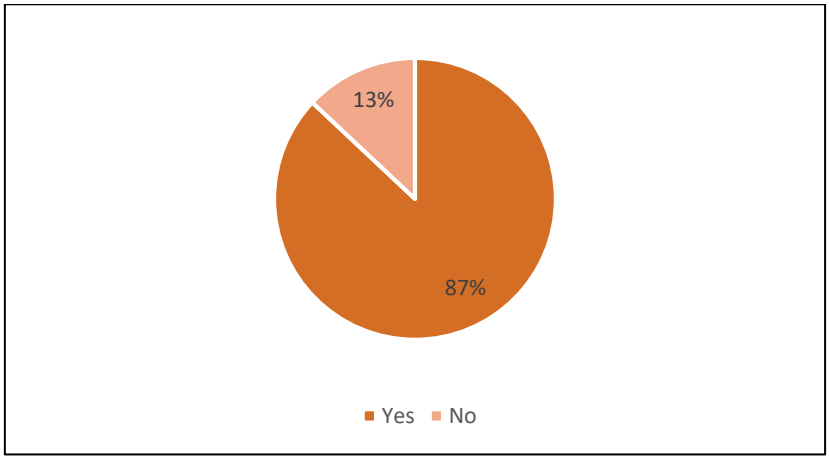


Source: Statista

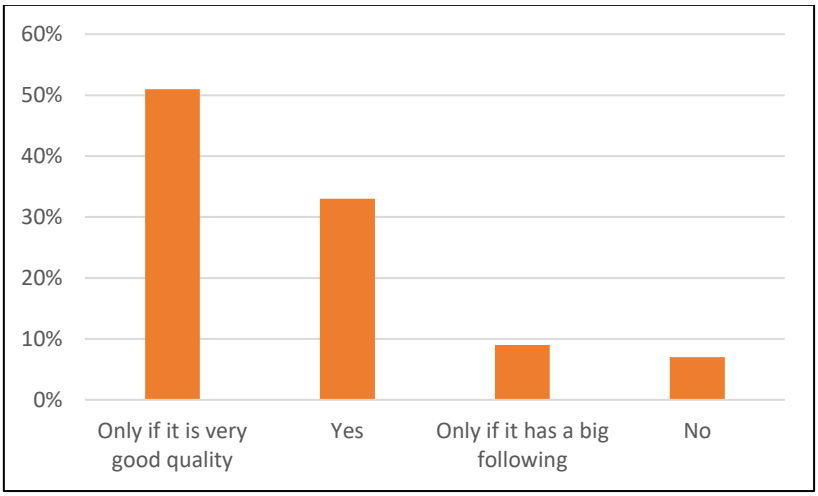
- *Appendix 15 - Survey: "What kind of content do you watch the most?"*



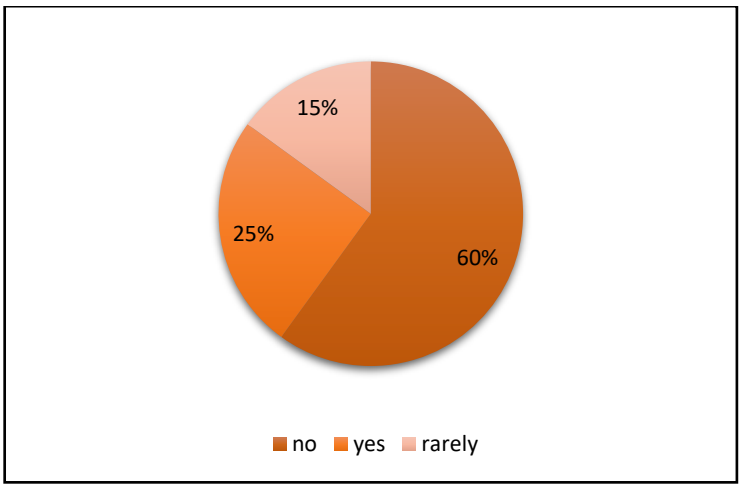
- *Appendix 16 - Survey: "Do you watch most content produced in English?"*



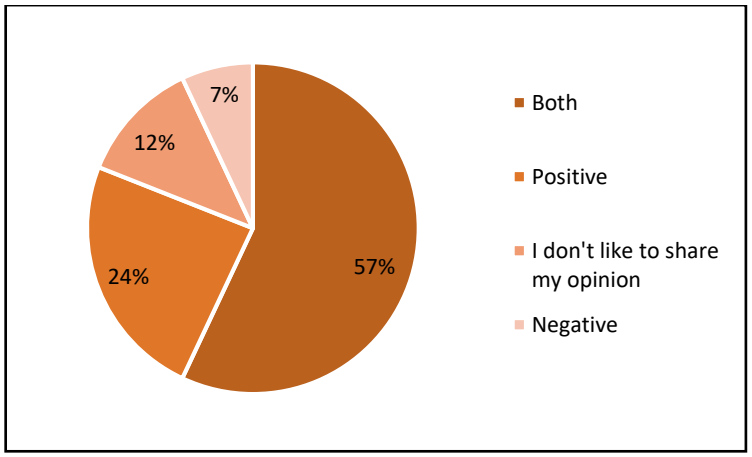
- *Appendix 17 - Survey: Would you watch more local content if available?"*



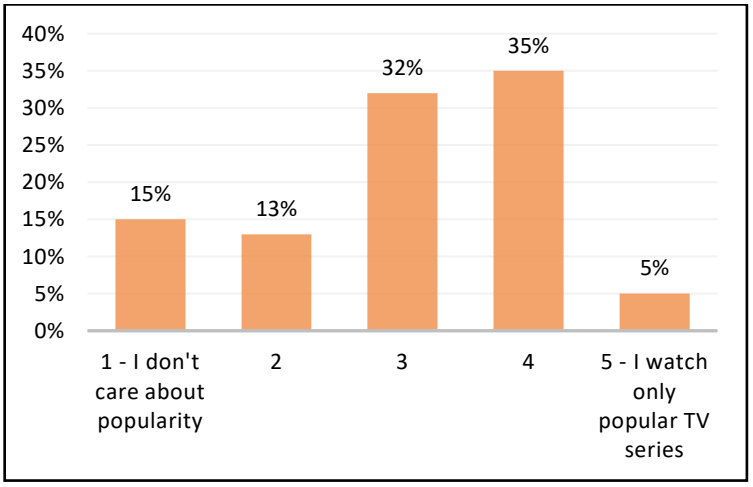
– *Appendix 18 - Survey: "Do you watch more than one series simultaneously?"*



– *Appendix 19 - Survey. "After consuming a product, are you more likely to share your experience, whether it was positive or negative?*



– *Appendix 20 - Survey. "How much does the popularity of a Tv series affect your decision-making process?"*



8. Bibliography

1. Camacho David, Panizo-LLedot Ángel, Bello-Orgaz Gema, Gonzalez-Pardo Antonio, Cambria Erik, “The four dimensions of social network analysis: An overview of research methods, applications, and software tools”, in *Information Fusion* 63, 88–120, 2020.
2. D.Chaffey, “*Global Social media research summary”*, 2019.
3. Guille Adrien, Hacid Hakim, Favre Cécile, Abdelkader Zighed Djamel. “Information Diffusion in Online Social Networks: A Survey”. SIGMOD record, ACM, 42 (2), pp.17-28, 2013.
4. Aggarwal C.C. “An Introduction to Social Network Data Analytics”. In: Aggarwal C. (eds) *Social Network Data Analytics*. Springer, Boston, MA, 2011.
5. Bernard J. Jansen and Mimi Zhang, Kate Sobel, Abdur Chowdury, “Twitter Power: Tweets as Electronic Word of Mouth”, in *Journal of the American Society for Information Science and Technology*, November 2009.
6. Coulter Keith S., Roggeveen Anne," Like it or not": Consumer responses to word-of-mouth communication in online social networks", in *Management Research Review*, Vol. 35 Iss: 9 pp. 878 – 899, 2012.
7. Kanoje Sumitkumar, Girase Sheetal, Mukhopadhyay Debajyoti, “User Profiling Trends, Techniques and Applications”, in *International Journal of Advance Foundation and Research in Computer (IJAFRC)* Volume 1, Issue 1, Jan 2014.
8. Allison Paul D., “*Multiple Regression: A Primer Research Methods and Statistics*”, 1st Edition, 1999
9. Nugus Sue, “Regression Analysis”, in “*Financial Planning Using Excel*”, Elsevier, chapter 5, 2009
10. Rennhoff Adam, Kenneth Wilbur, “The Effectiveness of Post-Release Movie Advertising*”* in *International Journal of Advertising*, 30 (2), 305–328, 2011
11. Buschow Christopher, Schneider Beate, Ueberheide Simon, “Tweeting television: Exploring communication activities on Twitter while watching TV”, in *Communications*, Volume 39: Issue 2, 2014
12. P. Hedlund, “*Sport Brand Community*”, p. 13, The Florida State University, 2011

13. Lee Kyung-Tag, Koo Dong-Mo, "*Effects of attribute and valence of e-WOM on message adoption: Moderating roles of subjective knowledge and regulatory focus*", School of Management, Kyungpook National University, 2012
14. Kolchyna Olga, Souza Thàrsis T. P., Treleaven Philip C. and Aste Tomaso, "*Methodology for Twitter Sentiment Analysis*", Department of Computer Science, UCL,2015
15. Doughty Mark, Rowland Duncan, Lawson Shaun, "*Who is on Your Sofa? TV Audience Communities and Second Screening Social Networks*", School of Computer Science, University of Lincoln, 2014
16. Cheung Christy M.K., Thadani Dimple R., "*The impact of electronic word-of-mouth communication: A literature analysis and integrative model*", Department of Finance and Decision Sciences, Hong Kong Baptist University, 2012
17. Hung K.H., Li SY, "The influence of eWOM on virtual consumer communities: social capital, consumer learning, and behavioral outcomes", in *Journal of Advertising Research* 47 (4) 485–495, 2007
18. Wood, W., Kallgren, C.A., Preisler, R.M., "*Access to attitude-relevant information in memory as a determinant of persuasion: the role of message attributes*", J. Exp. Soc. Psychol. 21 (1), 73–85 1985
19. Baber Alina, Thurasamy Ramayah, Malik Muhammad Imran, Sadiq Bushra, Islam Samina, Sajjad Muhammad, "*Online word-of-mouth antecedents, attitude and intention-topurchase electronic products in Pakistan*", Department of Management Sciences, COMSATS Institute of Information Technology, Attock, Pakistan School of Management, Universiti Sains Malaysia.
20. Brown Jo, Broderick Amanda J., and Lee Nick, "Word of Mouth communication Within online communities: conceptualizing the Online social network", in *Journal of Interactive Marketing* Volume 21 / Number 3 / Summer 2007
21. Rogers E.M. "Diffusion of Innovations". In *New York: Free Press*, 1983.
22. Lee J., Park D.H., Han I., "*The effect of negative online consumer reviews on product attitude: an information processing view, Electronic Commerce Research and Applications*" 341–352, 2008
23. Hennig-Thurau T., Gwinner K.P., Walsh G., Gremler DD, "Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet?", in *Journal of Interactive Marketing* 38–52, 2004.

24. Li Yung-Ming, Lin Chia-Hao, Lai Cheng-Yang, "*Identifying influential reviewers for word-of-mouth marketing*", Institute of Information Management, National Chiao Tung University, 2010
25. Ji Qihao, Zhao Danyang, "*Tweeting Live Shows: A Content Analysis of Live-Tweets from Three Entertainment Programs*", School of Communication Florida State University, 2015
26. Chen Yubo, Liu Yong, Zhang Jurui, "When Do Third-Party Product Reviews Affect Firm Value and What Can Firms Do? The Case of Media Critics and Professional Movie Reviews", in *Journal of Marketing* Vol. 75, 116 –134, 2011
27. Heinonen Kristina, "*Consumer activity in social media: Managerial approaches to consumers 'social media behavior*", Centre for Relationship Marketing and Service Management (CERS), Hanken School of Economics, Department of Marketing, Helsinki, 2011
28. Cheung Christy M.K., Thadani Dimple R., "*The impact of electronic word-of-mouth communication: A literature analysis and integrative model*", Hong Kong Baptist University, City University of Hong Kong, 2011
29. Shang R-A, Chen Y-C, Liao H-J., "*The value of participation in virtual consumer communities on brand loyalty. Internet Research*" 16(4): 398–418, 2006
30. Shao G., "Understanding the Appeal of User-Generated Media: A Uses and Gratification Perspective", in *Internet Research* 19(1): 7–25, 2009
31. Krishnamurthy S, Dou W., "Advertising with User-Generated Content: A Framework and Research Agenda", in *Journal Advertising* 8(2): 1–7, 2008
32. Brodie R., Ilic A., Juric B., Hollebeek L., "*Consumer engagement in a virtual brand community: an exploratory analysis*", J. Bus. Res. 66 (1) 105–114, 2013
33. Abdul-Ghani E., Hyde K., Marshall R., "*Emic and etic interpretations of engagement with a consumer-to-consumer online auction site*", J. Bus. Res. 64 (10) 1060–1066, 2011
34. Rui H., Liu Y., Whinston A., "Whose and what chatter matters? The effect of tweets on movie sales*,"in Decisive. Support Systems*, 55 (4) 863–870, 2013
35. Rishika R., Kumar A., Janakiraman R., Bezawada R., "The effect of consumers' social media participation on customer visit frequency and profitability: an empirical investigation", in *Inf. Syst. Res.* 24 (1) (2013) 108–127.

36. Wu J., Huang L., Zhao J., Hua Z., "The deeper, the better? Effect of online brand community activity on customer purchase frequency," in *Inf. Manage*. 52 (7) 813–823, 2015
37. Hoffman D.L., Fodor M., "*Can you measure the ROI of your social media marketing?*" MIT Sloan Manage. Rev. 52 (1) 41–49. 2010
38. Oha Chong, Roumanib Yaman, Nwankpac Joseph K., Hu Han-Fen, "Beyond likes and tweets: Consumer engagement behavior and movie box office in social media ", in *Information and management*, 25-37, 2017
39. Erickson Sarah E., Dal Cin Sonya, Byl Hannah, "An Experimental Examination of Binge-Watching and Narrative Engagement", in *www.mdpi.com/journal/socsci,* 2019
40. Yu Yinan, Ramaprasad Jui, "Engagement on Digital Platforms: A Theoretical Perspective", in *Fortieth International Conference on Information Systems*, Munich 2019
41. Jessica Braojos-Gomez, Jose Benitez-Amado, F. Javier Llorens-Montes, "Impact of IT Infrastructure on Customer Service Performance: The Role of Micro-IT Capabilities and Online Customer Engagement" in *Association for Information Systems AIS Electronic Library* (AISeL), 2015
42. Mareike Jenner, "Binge-watching: Video-on-demand, quality TV and mainstreaming fandom", Article in *International Journal of Cultural Studies* · September 2015
43. Vijay Viswanathan and Edward C. Malthouse, Ewa Maslowska, Steven Hoornaert and Dirk Van den Poel," Dynamics between social media engagement, firm-generated content, and live and time-shifted TV viewing", in *Journal of Service Management* 2018
44. Sidneyeve Matrix, "The Netflix Effect: Teens, Binge Watching, and On-Demand Digital Media Trends", in *Jeunesse: Young People, Texts, Cultures*, Volume 6, Issue 1, Summer 2014, pp. 119-138
45. Constantino Stavros, Matthew D. Meng, Kate Westberg, Francis Farrelly, "Understanding fan motivation for interacting on social media", in *Sport Management Review* 17 (2014) 455–469
46. Azza Abdel-Azim Mohamed Ahmed, "New era of TV-watching behavior: Binge-watching and its psychological effects", Article in *Media Watch*, January 2017
47. Jae-Hyeon Ahn, "*Previous Satisfaction and Positive Word-of-Mouth Communication as Antecedents to Purchase Intention in Transmedia Storytelling*", 2010 Korea Advanced Institute of Science and Technology

48. Hans Ouwersloot, Gaby Odekerken-Schröder, “Who’s who in brand communities – and why?” in *Eur. J. Marketing* 42, 2008, pp. 571–585
49. V. Madupu, D.O. Cooley,” Antecedents and consequences of online brand community participation: a conceptual framework”, in *J. Internet Comm*. 9, 2010, pp. 127–147
50. Ji Wu, Liqiang Huang, Jianliang Leon Zhao, Zhongsheng Hua, “*The deeper, the better? Effect of online brand community activity on customer purchase frequency*”, 2015
51. J.H. McAlexander, J.W. Schouten, H.F. Koenig, “Building brand community”, in *J. Marketing* 66, 2002, pp. 38–54
52. Sidneyeve Matrix 2014, “*The Netflix Effect: Teens, Binge Watching, and On-Demand Digital Media Trends*”, Queen's University
53. Arti J. Ugale, P. S. Mohod, "Business Intelligence Using Data Mining Techniques on Very Large Datasets", in *International Journal of Science and Research* (IJSR), Volume 4 Issue 6, June 2015, pp2932-2937
54. Oded Maimon, Lior Rokach “*Data Mining and Knowledge Discovery Handbook*”, 2nd ed 2010
55. Brojo Kishore Mishra, Deepannita Hazra, Kahkashan Tarannum and Manas Kumar, “Business Intelligence using Data Mining Techniques and Business Analytics”, in *5th International Conference on System Modeling & Advancement in Research Trends*, 2016
56. Ruxandra Petre,” Data Mining Solutions for the Business Environment “, in *Database Systems Journal* vol. IV, no. 4/2013
57. José Braga de Vasconcelos - Álvaro Rocha, “Business Analytics and Big Data”, in *International Journal of Information Management* 46 (2019) 250–251
58. Tim O'Reilly, “What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software”, 2005 in *https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html*
59. Hsinchun Chen, Roger H. L. Chiang and Veda C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact”, in *MIS Quarterly,* Vol. 36, No. 4 (December 2012), pp. 1165-1188 (24 pages)
60. Pang, B. and Lee, L., 2008. "Opinion Mining and Sentiment Analysis", in *Foundations and Trends in Information Retrieval* (2:1-2), 1-135
61. Tambe, P. (2014). “Big data investment, skills, and firm value,”, in *Management Science*, 60(6), 1452–1469.

62. Davenport, T., “Big data at work: Dispelling the myths, uncovering the opportunities”, in *Harvard Business Review Press,* 2014
63. Elisabetta Raguseo, “Big data technologies: An empirical investigation on their adoption, benefits and risks for companies”, in *International Journal of Information Management* 38 (2018) 187–195
64. Pak Irina, The Phoey Lee, “Value of Expressions behind the Letter Capitalization in Product Reviews”, in *Proceedings of the 7th International Conference on Software and Computer Applications*, Pages 147–152, 2018
65. Elragal Ahmed, Elgendy Nada, “Big Data Analytics: A Literature Review”, in *Paper Conference in Computer Science*, August 2014
66. Shangsong Liang, Xiangliang Zhang, Zhaochun Ren, Evangelos Kanoulas, “*Dynamic Embeddings for User Profiling in Twitter KDD*” August 19-23, 2018
67. G. Farnadi, J. Tang, M. De Cock, and M.-F. Moens, ‘‘User profiling through deep multimodal fusion,’’ in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, pp. 171–179, Feb. 2018.
68. Webb G. I., Pazzani M. J., and Billsus D., ‘‘*Machine learning for user modeling, User Model. User-Adapted Interact”*, vol. 11, nos. 1–2, pp. 19–29, Mar. 2001
69. Raghu T., Kannan P., Rao H. R., and Whinston A. B., ‘‘Dynamic profiling of consumers for customized offerings over the Internet: A model and analysis,’’ in *Decis. Support Syst.*, vol. 32, no. 2, pp. 117–134, Dec. 2001.
70. C. I. Eke, A. A. Norman, L. Shuib and H. F. Nweke, "*A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions*," in *IEEE Access*, vol. 7, pp. 144907-144924, 2019.
71. Sumitkumar Kanoje, Sheetal Girase, Debajyoti Mukhopadhyay, “User Profiling Trends, Techniques and Applications”, in *International Journal of Advance Foundation and Research in Computer (IJAFRC*) Volume 1, Issue 1, Jan 2014.
72. M. Gao, K. Liu, and Z. Wu, ‘‘Personalisation in Web computing and informatics: Theories, techniques, applications, and future research,’’ in *Inf. Syst. Frontiers*, vol. 12, no. 5, pp. 607–629, Nov. 2010.
73. Mostafa Mohamed M., “More than words: Social networks’ text mining for consumer brand sentiments”, in *Expert Systems with Applications*, Elsevier, 2013

74. Liu Yang, Huang Xiangji, An Aijun, Yu Xiaohui, "*ARSA: a sentiment-aware model for predicting sales performance using blogs*", 2007.
75. Bing Liu, "*Sentiment Analysis and Opinion Mining,*" Morgan & Claypool Publishers, May 2012.
76. Asur Sitaram, Huberman Bernardo, "*Predicting the Future with Social Media*", 2010
77. Yano, Tae et al. "*Predicting Response to Political Blog Posts with Topic Models*." *HLT-NAACL* (2009).
78. Mary McGlohon, Natalie S. Glance, Zach Reiter, "*Star Quality: Aggregating Reviews to Rank Products and Merchants*.",January 2010
79. Conference: "Proceedings of the Fourth International Conference on Weblogs and Social Media", in *ICWSM* 2010, Washington, DC, USA, May 23-26, 2010
80. Ye Sun, "How conversational ties are formed in an online community: a social network analysis of a tweet chat group", in *Information, Communication & Society*, 2019
81. Blei DM, Ng AY, MI Jordan, "*Latent Dirichlet allocation*", J Mach Learn Res 3:993 - 1022, 2003
82. Haddaway Neal R. "The Use of Web-scraping Software in Searching for Grey Literature", in *TGJ* Volume 11, Number 3 2015
83. Alsaeedi Abdullah, Khan Mohammad Zubair, "A Study on Sentiment Analysis Techniques of Twitter Data ", in *IJACSA, International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 2, 2019.
84. Cambria E., Poria S., Bisio F., Bajpai R., Chaturvedi I.," *The CLSA Model: A Novel Framework for Concept-Level Sentiment Analysis*", Gelbukh A. (eds) "Computational Linguistics and Intelligent Text Processing", in *CICLing, Lecture Notes in Computer Science*, vol 9042. Springer, Cham, 2015
85. Annett M., Kondrak G.," A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs", Bergler S. (eds) "Advances in Artificial Intelligence", Canadian AI. In *Lecture Notes in Computer Science*, vol 5032. Springer, Berlin, Heidelberg, 2008
86. Das, Cambria, Bandyopadhyay, Feraco, A "Practical Guide to Sentiment Analysis," Chapter 5, in *Socio-Affective Computing,* p.67.80, 2017).
87. Colbaugh Richard, Glass Kristin, "Emerging Topic Detection for Business Intelligence via Predictive Analysis of 'Meme' Dynamics Artificial Intelligence for Business Agility", Papers from the *AAAI Spring Symposium* (SS-11) 2011

88. Alnajran Noufa, Crockett Keeley, McLean David and Latham Annabel, “Cluster Analysis of Twitter Data: A Review of Algorithms”, in *ICAART - 9th International Conference on Agents and Artificial Intelligence* 2017
89. Steinskog, A., Therkelsen, J., Gamb¨ack, B., “Twitter topic modeling by tweet aggregation”, in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Association for Computational Linguistics. P. 77–86, 2017.
90. Wang C. and Zhang J. X., ‘‘Improved K-means algorithm based on latent Dirichlet allocation for text clustering,’’ *J. Comput. Appl*., vol. 34, no. 1, pp. 249–254, Jan. 2014
91. Xu Guixian, Meng Yueting, Chen Zhan, Qiu Xiaoyu, Wang Changzhi and Yao Haishen, “Research on Topic Detection and Tracking for Online News Texts”, in *Special Section on Artificial Intelligence and Cognitive Computing for Communication and Network*, 2019
92. Qiu L. and Yu J., ‘‘*CLDA: An effective topic model for mining user interest preference under big data background*,’’ Art. no. 2503816, vol. 2018
93. Liu J., Peng Y., Zhang L., Zhang Y., and Deng J., ‘‘*LDA-K-means algorithm of network food safety topic detection,*’’ Eng. J. Wuhan Univ., vol. 50, no. 2, pp. 307–310, Apr. 2017
94. Gropp Christopher, Herzog Alexander, Safro d Ilya, Wilson Paul W., Apon Amy W., “*Clustered Latent Dirichlet Allocation for Scientific Discovery*”, arXiv:1610.07703v3 [cs.IR] 4 Oct 2019.
95. Kaufman L., and Rousseeuw P., “*Finding Groups in Data: An Introduction to Cluster Analysis*”, New York: J. Wiley & Son, 1990.
96. Shreya Tripathi, Aditya Bhardwaj, Poovammal E, “Approaches to Clustering in Customer Segmentation”, in *International Journal of Engineering & Technology* 7(3.12):802, July 2018
97. Trupti M. Kodinariya, Dr. Prashant R. Makwana, “Review on determining number of Cluster in K-Means Clustering”, in Volume 1, Issue 6, *International Journal of Advance Research in Computer Science and Management Studies Research*, November 2013
98. Tonella, Tiella, Nguyen,” Interpolated *N-Grams for Model Based Testing*”, ICSE ’14, May 31 – June 7, 2014, Hyderabad, India
99. Kumari Singh Anita, Shashi Mogalla, “Vectorization of Text Documents for Identifying Unifiable News Articles”, in (IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 7, 2019

100. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya et al., “Preprocessing Techniques for Text Mining- An Overview”, in *International Journal of Computer Science & Communication Networks,* Vol 5(1),7-16, 2015

101. Curiskis Stephan A., Drake Barry, Thomas R., Osborn Paul J. Kennedy, “An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit”, in *Journal of Information Processing and Management* April 7, 2019

102. Xue Linyao, Jianguo,” Improved K-means Algorithm Based on optimizing Initial Cluster Centers and Its Application”, in *International Journal of Advanced Network Monitoring and Controls* Volume 02, No.2, 2017

103. https://www.forbes.com/sites/greatspeculations/2019/03/08/loss-of-licensed-content-is-an-underrated-crisis-for-netflix/#284126421117

104. https://www.dhs.gov/sites/default/files/publications/Lawful_Permanent_Residents_2017.pdf

105. https://www.nytimes.com/2018/12/04/business/media/netflix-friends.html

106. https://www.forbes.com/sites/greatspeculations/2019/03/08/loss-of-licensed-content-is-an-underrated-crisis-for-netflix/#3742f00f2111

This page intentionally left blank.