ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

JOANA PACHECO SCHERER

VISUAL ANALYSIS APPROACH FOR BRANDS PERCEPTION ON SOCIAL MEDIA

Porto Alegre
2021

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

VISUAL ANALYSIS APPROACH FOR BRANDS PERCEPTION ON SOCIAL MEDIA

JOANA PACHECO SCHERER

Master Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Isabel Harb Manssour

Porto Alegre
2021

Ficha Catalográfica

S326v Scherer, Joana Pacheco

Visual Analysis Approach for Brands Perception on Social Media / Joana Pacheco Scherer. – 2021.
88.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Isabel Harb Manssour.

1. Social Media. 2. Visual Analysis. 3. Interactive Visualizations. I. Manssour, Isabel Harb. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

JOANA PACHECO SCHERER

VISUAL ANALYSIS APPROACH FOR BRANDS PERCEPTION ON SOCIAL MEDIA

This Master Thesis has been submitted in partial fulfillment of the requirements for the degree of Master in Computer Science, of the Computer Science Graduate Program, School of Technology of the Pontifical Catholic University of Rio Grande do Sul

Sanctioned on Janeiro 18, 2021.

COMMITTEE MEMBERS:

Prof. Márcio Sarroglia Pinho (PPGCC/PUCRS)

Prof. Eduardo Campos Pellanda (PPGCOM/PUCRS)

Prof. Isabel Harb Manssour (PPGCC/PUCRS - Advisor)

For all those who in some way contribute to this journey.

“Science never solves a problem without creating ten more.”
(George Bernard Shaw)

ACKNOWLEDGMENTS

To my family for all their support.

To everyone in the DaVint Labs, for all the support, help, talks, and undocumented work arounds.

To everyone in the Praias research group, for the support.

To Caio, my first friend on this journey.

To the group "Friends that I got in PUCRS", for all the studies, support, and friendship.

To all PPGCC professors, for the lessons and support.

Lastly, but certainly not least, for all the support, help, and especially patience to my professor advisor Isabel Harb Manssour, that welcomed and guided me on this journey.

ABORDAGEM DE ANÁLISE VISUAL PARA A ANÁLISE DE MARCAS EM REDES SOCIAIS

RESUMO

Devido ao seu crescimento exponencial e sua rápida capacidade de prover feedback, as redes sociais tornaram-se importantes fontes de informação para diversas áreas. A grande quantidade de dados gerados diariamente fez das redes sociais fontes de dados confiáveis, rápidas e de baixo custo. Desta forma, as marcas perceberam que poderiam utilizá-las como ferramentas de marketing para obter um rápido retorno a respeito de seus produtos e serviços. Todavia, a análise uma marca através de suas redes sociais não é trivial e apresenta desafios tais como a coleta, análise, filtrage e organização dos dados. Para que a marca possa beneficiar-se dos dados obtidos através de de redes sociais, é necessários o desenvolvimento de ferramentas que auxiliem no seu entendimento. Essas ferramentas devem ser de fácil utilização pelos gestores das marcas, sem que seja necessário noções de programação. Neste contexto, o objetivo deste trabalho é prover uma abordagem de análise visual interativa, composta por várias técnicas de visualização, que auxilie a marca a obter vantagem dos dados provenientes de três redes sociais: Twitter, Instagram e YouTube. Nossa abordagem provê um pipeline que pode ser facilmente atualizado, sem a necessidade de programar. Além disso, são apresentados três estudos de caso que demonstram a possibilidade de obter várias informações a respeito dos dados coletados através do uso da nossa abordagem.

Palavras-Chave: Redes Sociais, Análise Visual, Visualizações interativas.

VISUAL ANALYSIS APPROACH FOR BRANDS PERCEPTION ON SOCIAL MEDIA

ABSTRACT

Due to the exponential growth and the quick feedback provided, Social Media has become an important information source for many areas. The thousands of data generated daily transformed Social Media into a reliable, fast, and relatively low-cost data source. So, brands note that they could use Social Media data as a marketing tool to obtain quick feedback about their products and services. However, analyzing a brand thru its Social Media is not a trivial task and raises challenges like data gathering and data analysis. To benefit from Social Media data, brands need tools that help them understand the vast amount of generated data. These tools need to be easy-to-use for brands managers that do not have programming knowledge. Thus, the objective of this work is to provide a visual analysis approach with several interactive visualization techniques to help brands obtain insights about the collected data from three social networks: Twitter, Instagram, and YouTube. Our approach provides a pipeline that can be easily extended and used without needing programming knowledge. Furthermore, three case studies are presented to demonstrate possible insights that can be identified using our approach.

Keywords: Social Media, Visual Analysis, Interactive Visualizations.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

API – Application Programming Interface

VADER – Valence Aware Dictionary for sEntiment Reasoning

CONTENTS

1. INTRODUCTION

Since its creation, social media has experienced exponential growth. Nowadays, social media fosters millions of users, is well integrated in our daily life, and became one important communication source. In the beginning, the focus of social media was to bring people together. But over time, it started to be used as a tool to express opinions about different subjects, like politics, products, brands, within others. This behavior ended up increasing the proximity among brands and users and also empowered users' opinions

This proximity is helpful for brands given the quick[28] and less costly feedback it provides when compared to traditional surveys. Especially since, according to Connor [44], "A standard telephone poll of one thousand respondents easily costs tens of thousands of dollars to run". The feedback obtained through social media is a rich source for marketing campaigns and strategies.

Likewise, social media also has empowered customers [10]. A recent example of social media power was the release of the Sonic movie trailer. The trailer[1] garnered so many negative comments regarding its main character appearance, that the studio decided to spend five million dollars on redesigning it. Redesigning the character was costly for the studio but probably avoided more losses by preventing a box office flop. This episode illustrates the importance of social media vigilance for brands and users' empowerment.

Moreover, the dynamic nature of social media can cause a range of problems for a brand, given its public perception can change very quickly. An inadequate product launch, a wrong choice of words in an announcement, or even an erroneous post, for example, can generate backlash very fast, causing damages and prejudice for the brand.

Given the present situation, we can imply the importance of monitoring social media for a brand. Observing its social media, brands can provide a proper response as soon as such events occur. It is worth mentioning that, for this work, we are using the term brand with a more broadly meaning. We consider that a person can also be referenced as a brand, considering that, many artists, celebrities, athletes, and YouTubers get monetized through their social media accounts. For instance, Neymar[2], Gisele Bündchen[3], and Anitta[4], who have 144, 16,5, and 50,4 million followers on Instagram, respectively, are some examples of celebrities that can be considered a brand.

However, brand analysis through its social media is not a trivial task and poses several challenges. At first, social media generates vast amounts of data that needs to be collected and properly stored to be later used. Gathering all this data can be very prob-

[1] https://www.cinemablend.com/news/2485139/sonic-the-hedgehogs-redesign-reportedly-cost-a-ton-of-money

[2] https://www.instagram.com/neymarjr/

[3] https://www.instagram.com/gisele/

[4] https://www.instagram.com/anitta/

lematic. Second, the collected data needs to be understandable for the brand, i.e., easy to analyze and interpret, or, otherwise, it might become useless. One useful solution for this challenge is to combine several visualization techniques and build visualization tools in which brands can obtain significant insights about the data.

Although some works similar to ours have already been developed [5, 33], they have been limited to Twitter and Glassdoor (for [33] only) analysis. Given this limitation, they do not provide means to compare the brand perception through different social media, which can be desirable for brands due to their necessity to reach a wider audience. Moreover, they do not provide analysis throughout time-series or comment chains, which can be useful features to assess the repercussion of controversial posts.

In this context, the main goal of this work is to provide a visual analysis approach with several interactive visualization techniques for helping brands obtain insights about the collected data of three social networks: Twitter, Instagram, and Youtube. Our approach allows the brand to inspect public perception, identifying which posts got more attention (good and bad), the most successful ones, and how the comments of a post behave, besides providing comparisons of how it performs in different social media. Thus, our approach focuses on the provided features and visualizations for brands perception analysis and management.

Our main contribution is the visual analysis approach to explore and compare three social media, which has a well-defined pipeline that can be easily extended and used by users without programming knowledge. Another contribution is the script for YouTube data collection. To the best of our best knowledge, this is the first work to analyze three social media together and provide comparisons between them.

The remainder of this work is organized as follows. In Chapter 2 we describe the related work for this research. Chapter 3 describes the developed approach including the research methodology and the research goals, the development environment, data collection and processing, along with our data sources, and a full description of our interactive visualizations. The case study used to validate the visual analysis approach is describes in Chapter 4. Chapter 5 presents our final remarks and future enhancements of this work.

2. RELATED WORK

In this chapter, we present the research methodology used and the related work selected for this study. Firstly, we introduce our research strategy along with the keywords and inclusion and exclusion criteria. Our strategy resulted in 23 papers that were entirely read. These papers were grouped and presented according to which social media was used as source data. Secondly, we analyzed some tools that presented functionalities related to our work. Finally, we provide a comparative table of the selected articles and discuss their advantages, disadvantages, and gaps.

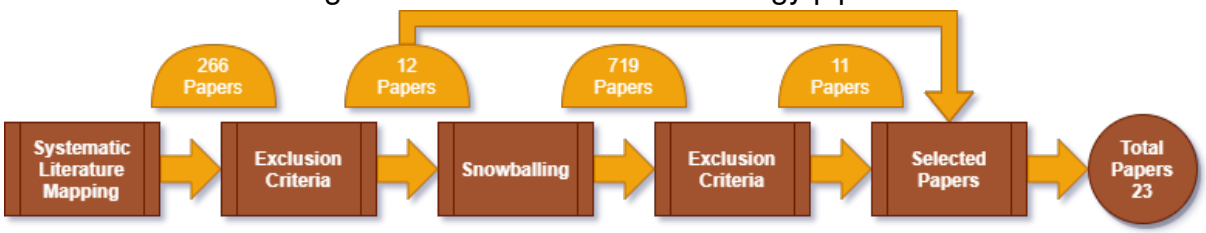
2.1 Research Methodology

One significant step of research work is to get an overview of the area that will be studied. This step is essential to identify which studies were already developed and find the main issues to be researched for the area's enrichment.

For this work, especially, this step is very relevant because several studies regarding social media analysis already exist. Thus, we needed to know those studies before we could propose a research topic to be developed.

We divided our research methodology into two steps. The first step was a Systematic Literature Mapping, where we used computer science libraries to select studies. The second step was a Snowballing process to complement the Systematic Literature Mapping. Our research division is presented in Figure 2.1.

Figure 2.1 – Research methodology pipeline.



A Systematic Literature Mapping is a research methodology used in research studies, which origin is medical researches [46]. The Systematic Literature Mapping aims to identify evidence available for a topic, provide an overview of an area, and identify the existing gaps in it. For this research, we opted to use the methodology proposed by Petersen et al [46]. This methodology was chosen because the authors present a literature mapping review methodology adapted to software engineering. Our completely Systematic Literature Mapping is presented in Subsection 2.1.1.

Snowballing is a qualitative research method used to identify papers that were not selected by other researches methods. This method consists of, starting from a group of initial studies, identify further relevant articles for the research subject. For this research, we used the methodology proposed by Wholin [62]. We decided to use this methodology because the author presents a Snowballing process adapted to software engineering. Our entire Snowballing description is presented in Subsection Subsection 2.1.2.

2.1.1 Systematic Literature Mapping

The Systematic Literature Mapping had the objective to identify which topics have already been researched regarding social media brands' perception. To achieve this goal, we also needed to understand which aspects of social media are analyzed to build the brand image. Furthermore, we would like to investigate which types of visualization are used by them and, in case they present sentiment analysis techniques[1], which are them. A[7]

To guide this study, we defined the main research question as: What has already been developed regarding brands' perception through social media? We also developed the following secondary research questions:

- Q1-Which aspects of social media should be considered to identify the brands' perception?
- Q2-Which data visualization techniques are being used to visualize the brand perception?
- Q3-Which sentiment analysis techniques are being used to analyze the brands' perception?
- Q4-Which studies compare brands' perception through different networks?
- Q5-Which studies analyze the brands' perception through time?
- Q6-Which studies analyze the brands' perception during real-time?

We chose to use ACM Digital Library, IEEE Explorer, and Science Direct libraries. We selected ACM Digital Library because its papers are all about Computer Science. IEEE Explorer was chosen because it indexes IEEE articles, and IEEE is one of the most used libraries for Computer Science, especially for publications in the data visualization area. Science Direct was selected because of the vast number of papers in Computer Science and Engineering, and because it indexes, also, other areas studies. Scopus was chosen because it indexes papers from several areas, not limited to computer science.

[1] According to Bing, sentiment analysis is "the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language".

Analyzing our objective and research questions, we defined the following keywords for our search: Brands, Branding, Visualization, Visual Analytics, Visual Analysis, Social Network, social media, Social Network Sites, Twitter, Facebook, and Instagram.

We present the search strings used for this work in Table 2.1. For IEEE Explorer and ACM Digital Library, we performed two searches, one including the visual word and one without it. The search containing the word "visual*" was not performed for the Science Direct library because wildcards are no longer accepted on their website. The searches were executed using the fields abstract, title, and keywords.

Table 2.1 – Search Strings

Data Source Search Strings	**Search Strings**
ACM Digital Library	(+visual* +brand* +("social media" "Twitter" "Social Network*" "Instagram" "facebook")) (+brand* +("social media" "Twitter" "Social Network*" "Instagram" "facebook"))
IEEE Explorer	(("Abstract":Visual* AND "Abstract":Brand* AND ("Abstract":"social media" OR "Abstract":"Twitter" OR "Abstract":. Social Network*. OR "Abstract":"Instagram" OR "Abstract":"facebook"))) (("Abstract":Brand* AND ("Abstract":"social media" OR "Abstract":"Twitter" OR "Abstract":. Social Network*. OR "Abstract":"Instagram" OR "Abstract":"facebook")))
Science Direct	Title-Abstr-Key(Visual* and Brand* and ("Social Network*" or "Twitter" or "social media" or "Facebook" or "Instagram"))

The search strings used in the three libraries resulted in 644 papers that were analyzed, and 12 of them were selected. The exclusion criteria were: Paper duplicity; Paper availability; Paper language must be in English or Portuguese, and Paper must be related to our subject. These same criteria were also used in the snowballing process. The complete list of selected papers in the Systematic Literature Review is presented in Table 2.2.

Table 2.2 – Systematic Literature Review

Dataset	**Title**	**Year**
Microblogs	Brand data gathering from live social media streams [18]	2014
Microblogs	Branty: A social media Ranking Tool for Brands [5]	2014
Microblogs	Filtering of Brand-Related Microblogs Using Social-Smooth Multiview Embedding [19]	2016
Microblogs	Insights from twitter analytics: Modeling social media personality dimensions and impact of breakthrough events [31]	2016
Microblogs, Glassdoor	SocialBrands: Visual analysis of public perceptions of brands on social media [33]	2016
Microblogs	Geo-localized public perception visualization using GLOPP for social media [2]	2017
Microblogs	The engagement strategy of Netflix Spain in twitter [16]	2018
Facebook	Attention Prediction on social media Brand Pages [32]	2011
Facebook	Digital traces for business intelligence: A case study of mobile telecoms service brands in Greece [38]	2014
Facebook	Estudo comparativo de mineração de opiniões em rede varejista [21]	2017
Instagram	Multimodal Popularity Prediction of Brand-related social media Posts [36]	2016
Internet Blogs	Brandmap: An information visualization platform for brand association in blogosphere [9]	2012

2.1.2 Snowballing

The objective of the Snowballing was to complement the results of Systematic Literature by adding more relevant papers. The Snowballing process includes two steps: Backwards and Forwards.

For the backward process, we verified the references of the previously selected papers presented in Table 2.2. In total, we analyzed 453 references, and we selected two more works related to our research subject.

For the forward process, we verified the papers that cited the previously selected works. We got a list of 266 papers and, after the exclusion criteria, we added another nine papers.

In total, for the Snowballing, we selected 11 papers. The complete list is provided in Table 2.3.

Table 2.3 – Snowballing Selected Papers

Dataset	Title	Year
Microblogs	The design of a live social observatory system [34]	2014
Microblogs	Live multimedia brand-related data identification in microblog [51]	2015
Microblogs	Applying brand equity theory to understand consumer opinion in social media [26]	2016
Microblogs	Multi-modal microblog classification via multi-task learning [65]	2016
Microblogs	Analyzing the startup ecosystem of India: a Twitter analytics perspective [52]	2019
Microblogs	The efficiency of social network services management in organizations. an in-depth analysis applying machine learning algorithms and multiple linear regressions [35]	2020
Various	StanceVis Prime: visual analysis of sentiment and stance in social media texts [29]	2020
Instagram	A Spatio-Temporal Category Representation for Brand Popularity Prediction [45]	2017
Facebook	Identification of the factors that affect the user reaction to posts on Facebook brand pages [25]	2018
Facebook	Toward maximizing the visibility of content in social media brand pages: a temporal analysis [30]	2018
Website	Predicting the Brand Popularity from the Brand Metadata [6]	2018

2.2 Microblogging Analysis

Microblogging platforms, such as Twitter and Sina Weibo, are commonly used social media for data analysis, and consequently, provide a very high number of articles. For example, Singh et al. [52] provide a methodology that can be used to extract relevant information from Twitter. Other works like [18, 34, 19, 51] focus on live streaming data gather and cleaning. The work presented by Luan et al. [34] also implemented event detection during live streaming. Two other papers provide data classification. One is Zhao et al. [65] which classify microblogs data according to the brand it belongs to, and Kalampokis et al. [26], which classify tweets data into relevant categories to enable automatic marketing metrics computation.

Regarding brand perception analysis, Aggarwal and Singh [2] evaluate the perception of a brand in different geographical regions, and other authors [31, 33] use microblogging data to assess brands according to brands' personality dimensions. Liu et al. [33] and Arvanitidis et al. [5][2] provide an analytical platform for monitoring brand social media data. In a different approach, Fernández-Gómez and Martín-Quevedo [16] attempt to discover the kind of posts that attract more attention for a brand. While another study, Matosas-López

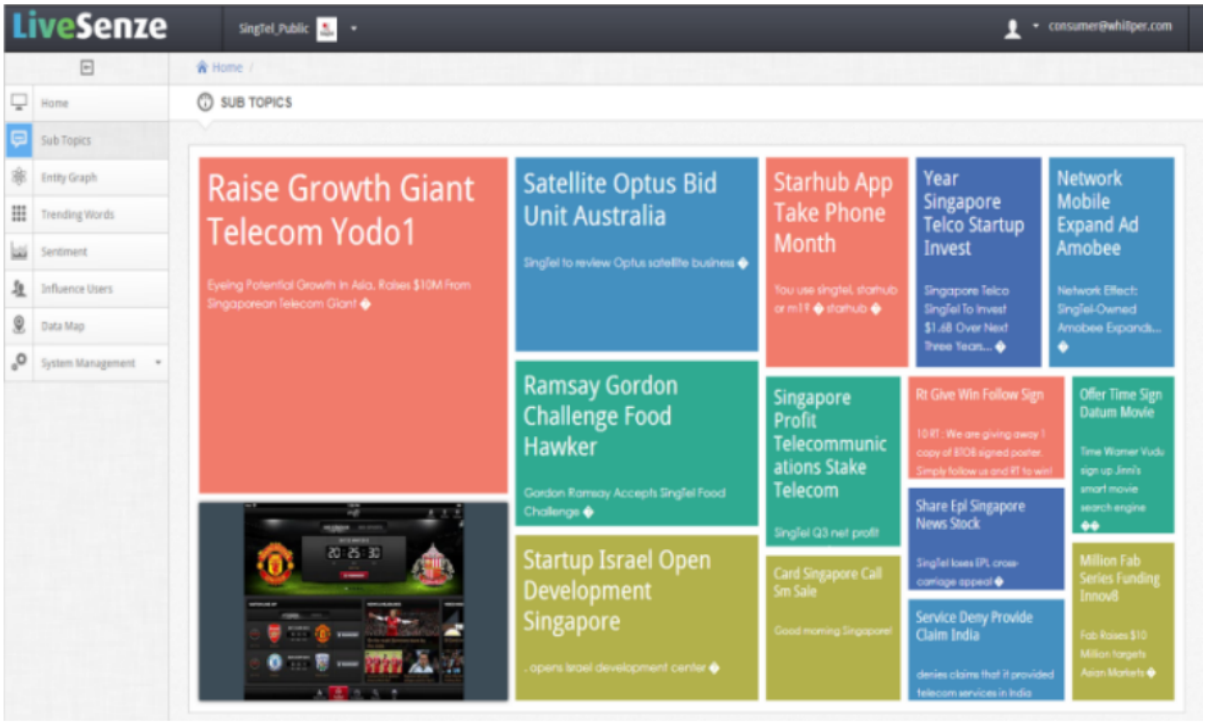
[2] Although the article mentions http://branty.org/ as Branty web application, this site was not available during the development of this work

and Romero-Ania [35], identifies which variables allow organizations to manage their social network services efficiently.

Of the selected papers for microblogging, five of them implemented visualization techniques. The works presented by Singh et al. [52] and by Arvanitidis et al. [5] implement simple visualization techniques. Singh et al. [52] implement bar and pie charts along with word clouds, and Arvanitidis et al. [5] provide just a horizontal bar chart graphic showing the rank of the most popular brands.

In another study, Luan et al. [34] present a Treemap for the events detected during the post-collection. Figure 2.2 presents one of the visualizations available for it. In this visualization, the box size indicates the event's importance based on the amount of data available.

Figure 2.2 – LiveSenze [34] - Events detection TreeMap



Geo-map visualization techniques are presented by Aggarwal and Singh et al. [2] and can be seen in Figure [2]. This figure presents a geo-map combined with a heatmap to show how the brand is perceived through an analyzed area. The green color is used where the brand is perceived positively, and the red color where the brand is perceived negatively.

Among the selected papers, the completest visualization technique is provided by Liu et al. [33]. This visualization presents the BrandWheels, a type of donut graph that illustrates how a brand is perceived among its personality traits. The authors also provide a visualization that summarizes the distribution of the brand over the personality traits. An example of the visualization provided is available in Figure 2.4.

Figure 2.3 – GloPP [2] - Brand Sentiment HeatMap

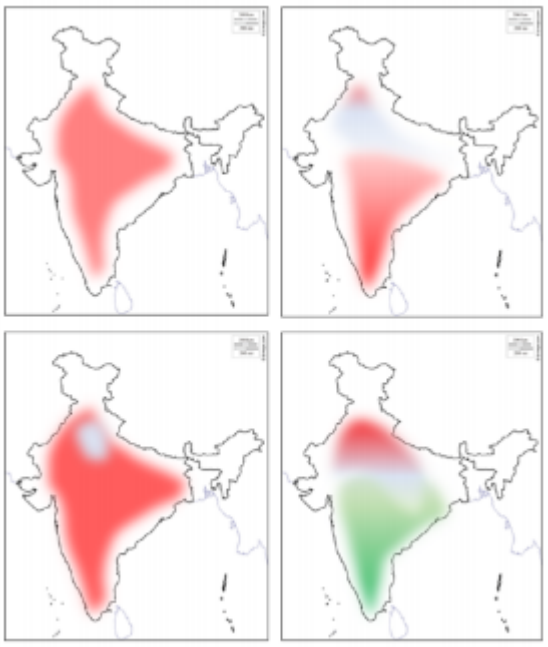
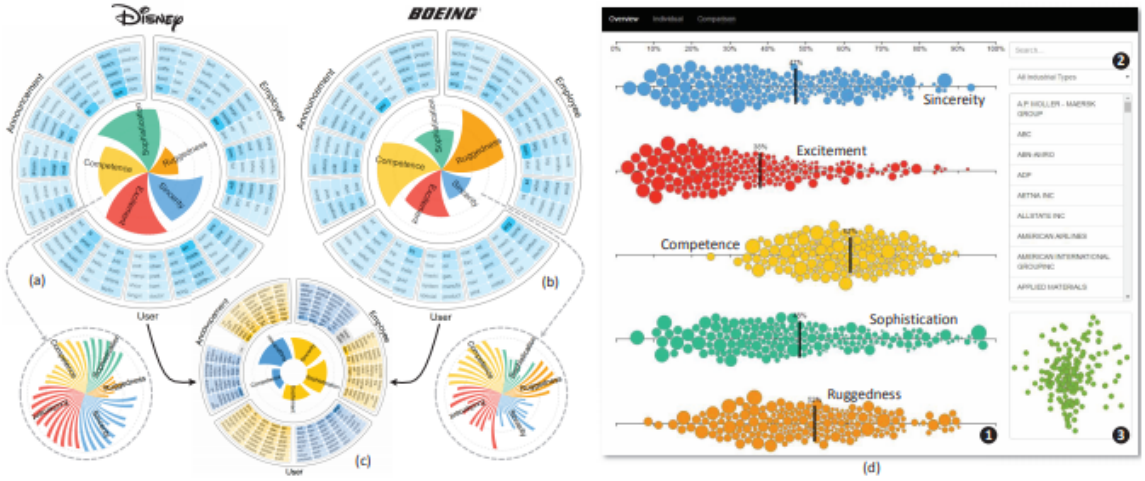


Figure 2.4 – Social Brands [33] - BrandWheel and Brand Perceptions



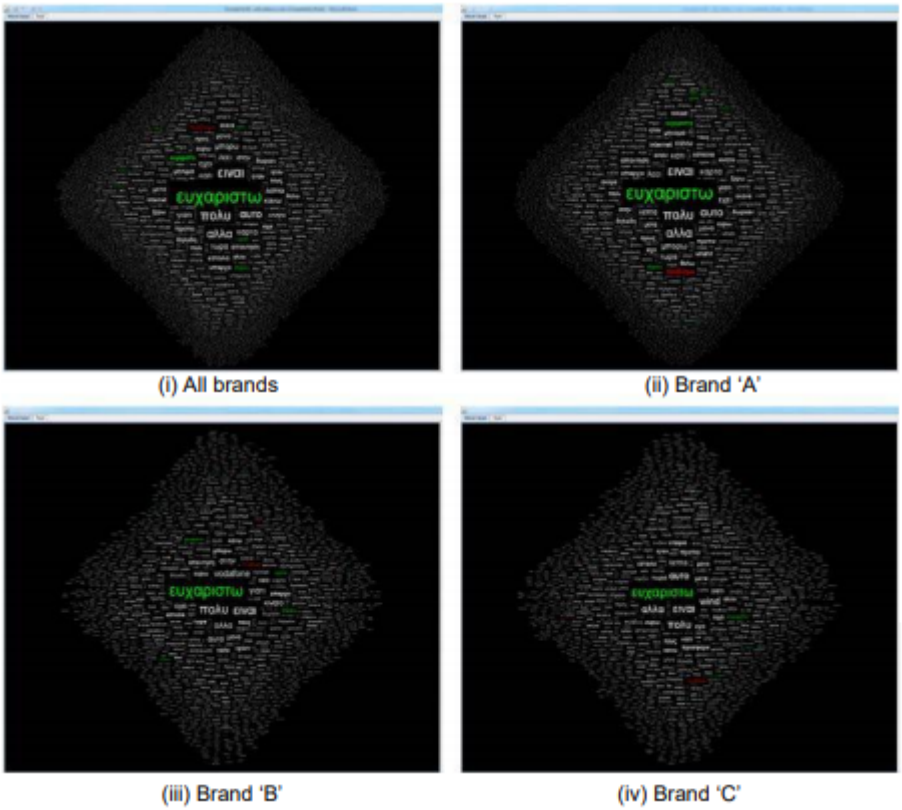
2.3 Facebook Analysis

Besides microblogging platforms, Facebook was another social network that appeared several times among our selected papers. Considering the papers selected for Facebook, the majority of them focus in attention prediction of a post. For example, Lakkaraju and Ajmera et al. [32] attempt to predict how much attention a new post will get. Authors Jeon and Ahn [25] attempt to discover the types of posts that attract more attention. Also, Kumar et al. [30] attempt to predict the best times of the day to maximize the attention and engagement a post will receive.

Of the two remaining articles, Hecksher and Ebecken [21] compare results between sentiment analysis and survey results, and Milolidakis et al. [38] describe a methodology to transform data gathered into valuable BI.

The work presented by Milolidakis et al. [38] is the only one that presents visualization techniques. One visualization developed by them is shown in Figure 2.5. This visualization shows a positive/negative word cloud for the most common words used in the brand's comments. Positive words are displayed in green, negative words are displayed in red, and neutral ones remain white.

Figure 2.5 – Digital traces Visualizations[38]



2.4 Instagram Analysis

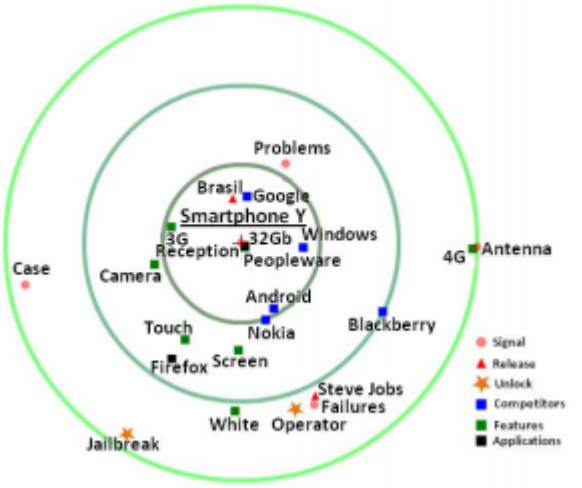
Different from microblogging and Facebook, Instagram only appeared in two of our selected papers. But, like Facebook, both articles focus on popularity prediction. Mazloom et al. [36] present the popularity prediction of a brand-related post using engagement parameters (sentiment, vividness, and entertainment) for better prediction. Meanwhile, Overgoor et al. [45] present the brand's overall popularity through incorporating spatio-temporal features. However, neither of them implemented Visualization techniques.

2.5 Other Social Networks Analysis

Besides microblogging, Facebook and Instagram, three papers rely on websites and internet blogs as source data. In their work, Bhargavi et al. [6] also attempt to predict a brand's popularity. They gather brands' comments from internet sites and classify them as favored or disfavored. Next, the comments are combined with the brand metadata to predict the brand's popularity. Meanwhile, Kucher et al. [29] provide a visual analytics solution that supports analysis for texts from multiple sources. They developed a visualization that shows the brand's posts in a timeline and provided a tool that compares the contents of different texts. The work presented by Campos Filho et al. [9] uses data collected from blogs to provide a Solar System graphic that shows the most relevant brands' topics among the collected data.

Figure 2.6 presents the visualization developed by Campos Filho et al. [9]. This visualization displays the brand as the "center" of the solar system and the terms used to describe the brand around the center like planets. The most commonly used terms are localized close to the center, and the fewer common ones are spread in solar system surroundings. The different physical characteristics (color, shape, and size) represent different brand dimensions like product attributes and related concepts.

Figure 2.6 – BrandMap Interface [9]



2.6 Tools

To enrich our research, we also felt the need to evaluate some existing social media analytics tools. Our goal in this step was to verify what kind of analysis these tools implemented and which visualizations they provided. We also focused on tools that went beyond

statistical analysis because while we recognize that statistical analysis is crucial, we wanted to see what was implemented besides it. For these reasons, we investigated the three tools presented next.

The first analyzed tool was Mention[3]. This tool presents several statistics visualizations for the brand. But besides the statistical data, it also implements other significant functionalities. For example, this tool provides a feature that shows if the brand's mentions on the internet are increasing or decreasing. Another valuable feature provided shows how many people are reached by a comment and where this comment originates. Currently, this tool analyzes data from Facebook, Instagram, Twitter, and Internet Blogs.

The second analyzed tool was Smain.io[4]. This project is no longer available, but it focused on building an API (Application Programming Interface) instead of a tool. Therefore, Smain.io was technically not a tool but an API used to develop other tools. The benefit of this approach was that the user could build his application according to his needs. This API provided some compelling functionalities like POP Score, POP Rank, Emotional Context, and others.

POP Score measures the popularity of profiles on social media. It follows the performance of brands and individuals on social media, assigning a score for them. They use different weights for each interaction type (a share has high weight than one like). POP Rank obtains all people that an individual or a brand has influenced, creating your direct influence net. It identifies who are the ones that most interact with you or your brand. The Emotional Context analysis provides the emotional context of people's comments. The comments analysis returns which sentiment and emotion are predominant on them.

Buzzmonitor was the third analyzed tool. This tool provides social media monitoring instruments, using user-selected terms and geo-localization. Another feature provided by this tool is the possibility to identify digital influences according to your objectives. For instance, if the brand launches a new marketing campaign targeting a specific public, it can find the best influencers for this project. This tool also monitors news websites and informs the brand of the latest news regarding them. This tool was used by Campos Filho et al. [9] to obtain the data needed for their project.

2.7 Discussion

In this section, we present a comparative analysis of the selected works. We also provide the answers to our research questions and the research topics to be explored in future developments.

[3]https://mention.com
[4]http://snam.io/

2.7.1 Works Comparison

We provide a comparative table showing the relevant features of the articles from different social media we have selected for this section. The full table is presented in Table 2.4 and displays the following features: title, dataset, publication year, objective, implemented sentiment analysis, and implemented visualizations. We also provide a general comment about its contents.

One aspect that we notice is that most papers analyze only one social media at a time. Only two of the selected articles [33, 29] gather data from more than one social media. These articles combine data of different social media to provide an overall brand's perception and do not analyze if the brand perception changes according to social media. The only exception is the study developed by Kucher et al. [29], but this tool only compares post texts.

We found only two papers that work with time series analysis. One of them is the work provided by Lakhiwal et al. [31], but this work is restricted to the occurrence of a major event like a strike. The other one is the study developed by Overgoor et al. [45]; that incorporates time series analysis into popularity prediction.

Another aspect we notice is that very few papers present visualization techniques. Among the ones that do present, the majority only provide simple visualizations. Only four papers implemented more elaborated visualizations.

2.7.2 Answers to research questions

Q1-Which aspects of social media should be considered to identify the brands' perception?

Among the selected paper, the most common aspect used to identify the brand's perception is the analysis of users' comments. The works developed by Milolidakis et al. [38] and by Campos Filho et al. [9] use the comments to count word frequency and display the most common ones in word clouds. Another study provided by Aggarwal and Aggarwal and Singh [2] analyzes users' comments as a whole, classifying them either as positive or negative. They use this classification to provide a colored map showing the locations where the brand is perceived positively and negatively.

Some authors like Kalampokis et al. [26] opted to develop their own classification method. Their method classified the comments in the following categories: Satisfaction, Image, Intention, and None, and they claim that this classification allows automatic marketing computations.

Table 2.4 – Articles Overview Table

Title	Dataset	Year	Objective	SentimentAnalysis	Visualization Techniques
Brand data gathering from live social media streams [18]	Microblogs	2014	Improve social media data gathering.	Mixed text analysis and image analysis	
Branty: A social media Ranking Tool for Brands [5]	Microblogs	2014	social media monitoring platform that analyzes, ranks and visualizes the social presence of brands on Twitter.	SentiWordNet	BarChart
The design of a live social observatory system [34]	Microblogs	2014	Continuous social media data gathering to detect events.		Time Line, Treemap
Filtering of brand-related microblogs using social-smooth multiview embedding [19]	Microblogs	2015	Improve social media data gathering by incorporating brand and social relations.	Mixed text analysis and image analysis	
Live multimedia brand-related data identification in microblog [51]	Microblogs	2015	Improve social media data gathering using pre-analyzed offline data.		
Applying brand equity theory to understand consumer opinion in social media [26]	Microblogs	2016	Classifies tweets into relevant categories (Satisfaction, Image, Intention, None) to enable automatic marketing metrics computation.	Category classification algorithm implemented by the authors.	
Insights from twitter analytics: Modeling social media personality dimensions and impact of breakthrough events [31]	Microblogs	2016	Evaluate brands according to brands personality dimensions.	Classify brands on brands personality dimensions	
Multi-modal microblog classification via multi-task learning [65]	Microblogs	2016	Classifies microblogs into different brands.		
SocialBrands: Visual analysis of public perceptions of brands on social media [33]	Microblogs, Glassdoor	2016	Present an analytic tool to assess and analyze public perceptions of brands on social media.	Latent Dirichlet Allocation, Classify brands on brands personality dimensions Sunburst Graph	Sunburst Graph
Geo-localized public perception visualization using GLOPP for social media [2]	Microblogs	2017	Evaluate brands by regions	Classify brands on brands personality dimensions	Geo Map, Bar chart,Wordcloud
The engagement strategy of Netflix Spain in twitter [16]	Microblogs	2018	Discover which kind of posts attract more attention	Visual-Sentiment, Textual-Sentiment	
Analyzing the startup ecosystem of India: a Twitter analytics perspective [52]	Microblogs	2019	Provide a methodology that can be used to extract relevant information from Twitter.	Latent Dirichlet Allocation	Bar chart, Pie Chart, Wordcloud
The efficiency of social network services management in organizations. an in-depth analysis applying machine learning algorithms and multiple linear regressions [35]	Microblogs	2020	Identify the variables (publication volumes, components, day of the week, time slot, topic, and recognition obtained by the publication) that allow organizations to manage their social network services efficiently.		
Attention Prediction on social media Brand Pages [32]	Facebook	2011	Predict the attention of a new post	SentiWordNet	
Digital traces for business intelligence: A case study of mobile telecoms service brands in Greece [38]	Facebook	2014	Describes a generic methodology for social media data gathering and transforming it into valuable BI.	Only word count was implemented	Colored Word Cloud,Table Lens,Interaction Graphic
Estudo comparativo de mineração de opiniões em rede varejista [63]	Facebook	2017	Compare sentiment analysis results with survey results	Apache, OpenNLP, Document Categorizer	
Identification of the factors that affect the user reaction to posts on Facebook brand pages [25]	Facebook	2018	Discover which kind of posts attract more attention		
Toward maximizing the visibility of content in social media brand pages: a temporal analysis [30]	Facebook	2018	Find the best posting time(s) to get high visibility.		
Multimodal Popularity Prediction of Brand-related social media Posts [65]	Instagram	2016	Predict brand-related post popularity	Visual-Sentiment, Textual-Sentimen	
A Spatio-Temporal Category Representation for Brand Popularity Prediction [45]	Instagram	2017	Predict brand-related popularity	Textual Sentiment Analysis, Keywords Extraction	
Brandmap: An information visualization platform for brand association in blogosphere [9]	Internet Blogs	2012	Provide a Solar System graphic showing the most relevant topics for a brand	Only count the most common words and classify them	Solar System
Predicting the Brand Popularity from the Brand Metadata [6]	Website	2018	Propose a novel framework to classify the comment's as favored or disfavored, and later combines them with the brand metadata to forecast the popularity of the brand.		
StanceVis Prime: visual analysis of sentiment and stance in social media texts [29]	Various	2020	Provide a visual analytics solution that supports the exploratory analysis of sentiment and stance classification results for temporal text data from multiple sources.	Vader	Time Series for text similarity.

Meanwhile, authors Liu et al. [33] and Lakhiwal et al. [31] presented the most complex sentiment analysis among the selected articles. Those two authors classify the com-

ments as one of the five brand personality traits areas (Sincerity, Excitement, Competence, Sophistication, and Ruggedness) which, are well known in the marketing area [1]. These traits are reliable and valid measurements for assessing brand personality.

Finally, articles by Fernández-Gómez and Martín-Quevedo [16]. and Jeon and Ahn [25] approached the brand perception problem differently using statistical data collected through social media to assess how the brand is perceived.

Q2-Which data visualization techniques are being used to visualize the brand perception?

Visualization techniques were developed by several papers evaluated in our research. Some studies took a simple approach and implemented classical visualization techniques like word clouds and bars charts. The word clouds visualization is present in papers by Milolidakis et al. [38], Aggarwal and Singh [2], and Singh et al. [52]. This visualization intends to provide a general idea of what is being said about a brand (topics, positive and negative words, among others).

The bar charts visualiztion technique were found in paper by [2, 5, 52]. This visualization technique intentions is to facilitate the comparison among data from different data sources.

Four authors presented more complex and more detailed visualization implementations. Authors Luan et al. [34] implement a tree view visualization technique. Different from Campos Filho et al. [9], which display their findings arranged in a solar system visualization, and from Liu et al. [33], which organize their data in a brand wheel. Moreover, author Kucher et al. [29] presents the texts gathered for a brand in a timeline.

Q3-Which sentiment analysis techniques are being used to analyze the brands' perception?

From the analyzed papers, six of them presented some type of Sentiment Analysis. From these, two papers [9, 38] did not fully implement Sentiment Analysis and were restricted to counting words. Papers by Mazloom et al. and by Overgoor et al. provided Sentiment Analysis evaluating combinations of text and image. Differently, papers by Liu et al. and by Lakhiwal et al. papers classified the brands according to the brand's personality dimensions: Sincerity, Excitement, Competence, Sophistication, and Ruggedness, instead of a simple positive and negative evaluation.

Q4-Which studies compare brands' perception through different networks?

Most of the papers we found only use one social media as a data source. However, three exceptions were presented. Campos Filho et al. [9] combined data from several internet blogs, and Liu et al. [33] combined data from microblogs and Glassdoor[5], to build general brand perception but do not compare the brand perception among them. Kucher et

[5]Glassdoor is a social media that provide insights about jobs and companies.

al. is the only paper that provides a comparison between different social media, but they only target microblogging texts and do not provide means to compare statistical data.

Q5-Which studies analyze the brands' perception through time?

Despite our efforts, we were not able to provide an answer to this question. Considering the selected papers, only two of them implemented temporal analysis. One of those papers is the study developed by Lakhiwal et al.[31]. But for this study, the analysis of the brands' perception thought time depends on the occurrence of a big event; thus, the continuous-time analysis does not occur.

The other paper that implements time analysis is the study presented by Lakkaraju and Ajmera [32] In this study, the authors presented the analysis of a brand's posts through time, using the posts data to predict the brand's popularity. Besides these two studies, the only other mention of real-time analysis occurs in the study by Liu et al. [33], but they mention it as a future enhancement of their work.

Q6-Which studies analyze the brands' perception during real-time?

Although several articles [18, 34, 19, 51] gather data in real-time, their only concern is to filter noisy data and detect events. They do not provide brand analysis in real-time. So, we consider that this question was not answered by any of the papers we found. In our understanding, this shows that an analysis of brands' perception through social media is a relatively new area and, due to this, still have gaps to be filled in.

2.7.3 Research Opportunities

Considering the selected paper, we could identify some research opportunities that future work can address. The first is concerning the comparison of brand perception on different social media. The majority of the papers we found only work with one social media at a time. Articles that use more than one social media often combine this data to provide an overall of the brand's perception. Only the paper by Kucher et al. [29] implements some comparison from different social media, but the authors only provide a comparison between posts' texts. The comparison of other features, like statistics ones, is not provided. Although they do not explicitly mention social media comparison, papers [2, 5, 33] cite expansions for other social media as future work.

The second one is about time series analysis. Considering the brand's perspective, the importance of monitoring their social media data for long periods is very high due to the valuable data they provide. For instance, brands can identify which products launched were successful and which ones failed, or which posts generated positive engagement, and which ones generated a negative one. Thus, time series analysis is an important feature to detect these types of events. Among the selected papers in our research, Lakhiwal et al. [31]

present time series analysis, and Liu et al. [33] describe time series analysis as a future work development.

The third one regards real-time analysis. Some of the works [18, 34, 19, 51] we analyzed implemented tools for real-time data gathering and event detection. But they do not provide brands' perception analysis in real-time. The paper developed by Liu et al. [33] mentions real-time analysis as future work, which would be a valuable feature for automatically detecting backlash events, so the brand can take some actions to minimize them.

Lastly, the fourth one is regarding predicting the popularity of a post. Several papers [32, 36, 45, 6] addressed this issue, but they only work with Instagram and Facebook social media. So, it might be convenient to have a similar prediction method for other social media, like Twitter. Also, it is possible to suggest improvements to the already developed methods.

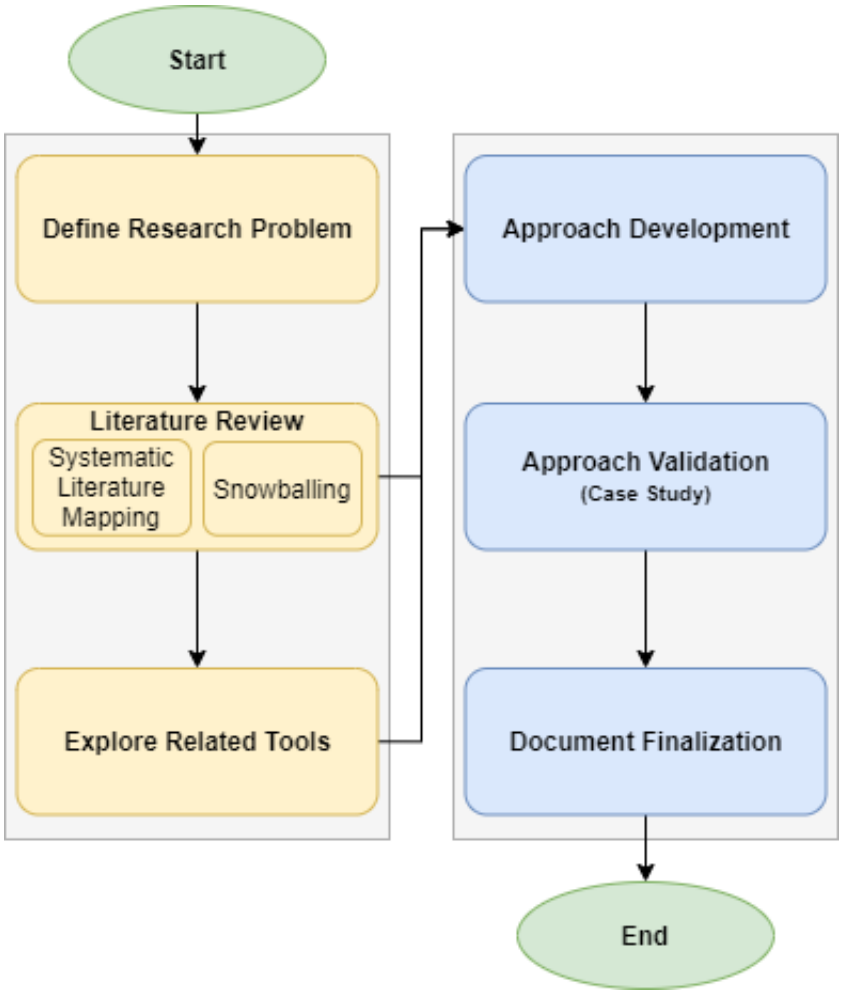
3. VISUAL ANALYSIS APPROACH DESCRIPTION

This section provides a complete description of how this work was developed, starting with the research methodology, research questions, and goals. After that, we present the development environment, describe the data input, storage, and processing, as well as the interactive visualization techniques developed.

3.1 Research Methodology

This study was divided into two phases, as shown in Figure 3.1. The first phase contains four tasks related to literature research. These tasks aim to provide the state of the art of this research area and identify possible contributions. The obtained results of this first phase were presented in Section 2 and served as guidelines for the definition of our research questions presented in Subsection 3.2.

Figure 3.1 – Research phases, along with tasks division.



The second phase consists of the development, validation, and writing tasks. The approach development task describes the design and the development of our approach, including used libraries, databases, and interactive visualization techniques. This task is described in detail in Subsection 3.5.

The validation task was included to identify the advantages and some limitations of our approach. According to Ward [59]: "...some visualization researchers attempt to validate the effectiveness of their techniques by showing real (or sometimes contrived) examples". Since the research subject is brand analysis through social media data, we chose to use a real data analysis approach. Once our focus was not system usability, instead of performing users test, we choose to analyze three social media profiles to provide the approach validation. Through these case studies, presented in Chapter 4, it was possible to obtain some interesting insights. Lastly, the final task was the document finalization, which consists of the writing of this volume.

3.2 Research Questions and Goals

Considering the literature review presented in Chapter 2, we could identify four main gaps regarding the use of visual analysis techniques for brand perception in social media: brands perception comparison among different social media; real-time analysis of brand perception; time series analysis for brands perception, and post's popularity prediction.

Thus, we decided to focus our efforts on two of these gaps: comparison between social media; and time series analysis for brand perception. We chose to cover these gaps because we believe that they are main issues in brand perception management. Also, time-series analysis is cited as future work in [33]. We will not provide real-time analysis because it involves other problems such as continuous data collection, which usually is a limitation of the social media APIs themselves [58]. Moreover, big data management and processing are outside the scope of this work.

In this context, we formulated the following research questions for this work:

- RQ1 - Which graphical representations facilitate the identification in which social networks a brand is more present?
- RQ2 - How can we identify which are the posting patterns that generate the highest number of interactions?
- RQ3 - How do the comments of a post behave over time?
- RQ4 - On which social network does the brand have a higher number of interactions?

Considering these research questions, the main goal of this work is to provide a visual analysis approach for data gathered from different social media. This approach aims to help brands to analyze their image across several social media. The study aims to provide statistical information and different visualization techniques to allow the user to obtain insights on how a brand can increase its social media presence. We hope that the proposed model not only facilitates social media data visualization but also helps to automatize the analysis process steps.

The visual analysis approach is the main contribution of this work, but we also provide the following contributions: a rank of the posts that attracted more interactions; social media statistical analysis including sentiment analysis; social media comparison; and behave of a post' comments over time.

3.3 Selected social media

We selected Twitter, Instagram, and YouTube, to be accepted in our approach because they have millions of users and have some similar features like the number of likes, shares, and comments. Due to these similar characteristics of those social media, it is easy to make comparisons among them. Moreover, Twitter and YouTube provide APIs for data collection, and Instagram allows data screening.

Twitter is the social network that people use to comment about many things, from daily events to politics. Regarding brands, Twitter is one of the first places used for compliments and complaints. Twitter also provides a feeling of proximity from the users with the brands. Twitter's collected data was split into two categories: profile and posts. Unfortunately, it was not possible to collect replies to tweets due to API restrictions.

Since Instagram posts contain, in its majority, visual communication, several brands use Instagram to show and advertise their products. People usually like pictures and include comments, generating visibility for the brand. Also, several smaller businesses use it as their main social media to show their products. For Instagram, the data we collected was divided into three categories: profile and posts, and comments.

Finally, YouTube, the last social media selected, is used by several brands for advertising products and launching commercials. These videos generate lots of views, comments, and shares, providing almost instant feedback for the brand. YouTube data collected was divided into three categories: channel profile, video, and video comments. The details of the collected data for the three social media are provided in Figure 3.4.

3.4 Development Environment

For this study we chose to use Python as our development language. According to Python's website [50]: "Python is a programming language that lets you work quickly and integrate systems more effectively.". Python is relatively easy to learn and use, is portable, and provides several libraries for data analysis and data visualization that are very helpful for this work [50].

We decided to use Jupyter Notebook [48] as the development environment. It is defined as "an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text". We chose to use it mainly because it provides a quick and simple environment for verification and tests since there is no need to compile or run our code through the prompt.

For data storage, we are using MongoDB [40], a NoSQL database that is "simple for developers to learn and use, while still providing all the capabilities needed to meet the most complex requirements at any scale". The main reason to select it as our database was the extensive use of emojis on social media. Since emojis are a significant resource for sentiment analysis, we could not remove them from the dataset. By the time we were developing this feature, even though a relational approach could have been more suitable, MongoDB was the only one (among Postgress, MySQL, and SQLDeveloper) with a driver prepared to handle UTF-32 in Python (used by Twitter).

Since Pandas is a library simple to use for data analysis and manipulation in Python [55], we used it as our data analysis library. Pandas [37] is widely used, documented, and provides the main statistics analysis functionalities we need like average, standard deviation, median, mode, among others.

For the implemented visualizations, we chose to use the Seaborn Library, "a Python data visualization library based on facts, which provides a high-level interface for drawing attractive and informative statistical graphics" [60]. Seaborn [61] was selected because it presents more refined charts, is easy to use, and provides a high degree of flexibility and interactivity. We have considered other visualization libraries, some of them supply prettier charts than Seaborn (like Plotly [47]), but those libraries do not provide the same degree of flexibilization and interactivity.

To overcome some limitations that Seaborn, unfortunately, has (as the lack of Pie chart support), we used Matplotlib, "a comprehensive library for creating static, animated, and interactive visualizations in Python" [54]. Matplotlib [22] is one of the most commonly used visualization libraries in Python. However, although it is flexible, it requires more development work, especially if we considered more elaborated charts.

Since we want to provide sentiment analysis for comments, we chose to use Vader (Valence Aware Dictionary for Sentiment Reasoning) [23], a library that has sentiment anal-

ysis algorithms based on lexicons that detects polarity. Vader divides the text into lexicons and, for each one of these lexicons, a positive or negative score is assigned. The algorithm also considers the lexicon intensity classifying it in either highly positive or highly negative. The algorithm then presents the score for the positive, negative, and neutral lexicons founded in the text along with an overall compound score.

Vader is used for sentiment analysis in social media because it considers several social media aspects that are usually ignored by other methods. For example, it considers emoticons, capital letters, and multiple exclamation points to classify the text. According to Hutto and Gilbert [23], Vader has an accuracy of 96% for tweet analysis, 61% for movie reviews analysis, and 63% for amazon products reviews, which are higher rates obtained so far for this kind of analysis.

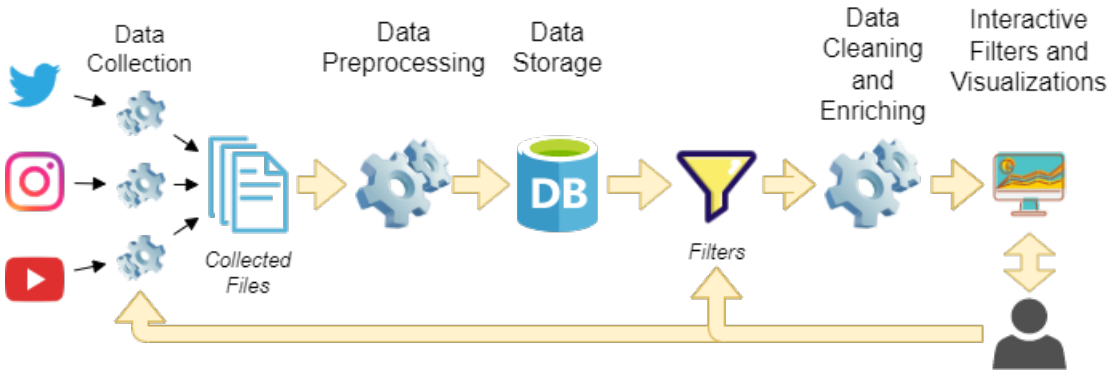
Since Vader is only currently available in English, and we are working just with English tweets and comments, we also needed to use libraries for language detection. For this task, we selected the libraries Langdetect and Langid. According to the official Python Library website [49], Langdetect "is a direct port of Google's language-detection library from Java to Python" and Langid "is a standalone Language Identification (LangID) tool". Both of them have benefits and limitations, so we decided to use a combination of them to verify the comments languages.

Finally, for word clouds generation we used WordCloud Python's library, which provides features as selecting the most relevant terms from a text and stopwords removal. It is easy to use, requires very little code, and has detailed documentation [42, 41].

3.5 Proposed Pipeline

The developed pipeline for our approach was divided into five main steps that correspond to the data flow, as illustrate Figure 3.2: (1) Data Collection; (2) Data Preprocessing; (3) Data Storage; (4) Data Cleaning and Enriching; (5) Interactive Filters and Visualizations. These steps correspond to the data flow in the proposed approach. The first step encompasses the collection of social media data. Next, the collected data is pre-processed through the execution of the sentiment analysis algorithms. We chose to run the sentiment analysis at this point to minimize performance issues during the interactive visualizations. After that, the data is stored in our database. The next step provides data cleaning for word cloud generation and processes statistical information. Finally, the development of the interactive filters and visualization techniques occurs in the last step. Sections 3.6, 3.7, 3.8, 3.9 and 3.10 present, respectively, a detailed description of these steps. The developed scripts for data collection are available at [13, 12, 14], and the script for the pipeline's remaining steps is available at [11].

Figure 3.2 – Details of the approach pipeline.



Through the proposed pipeline, the user can update the data without the need to make new implementations. The user only needs to provide new collected data files and execute the pipeline again to filter, visualize, and interact with the latest data.

3.6 Data Collection

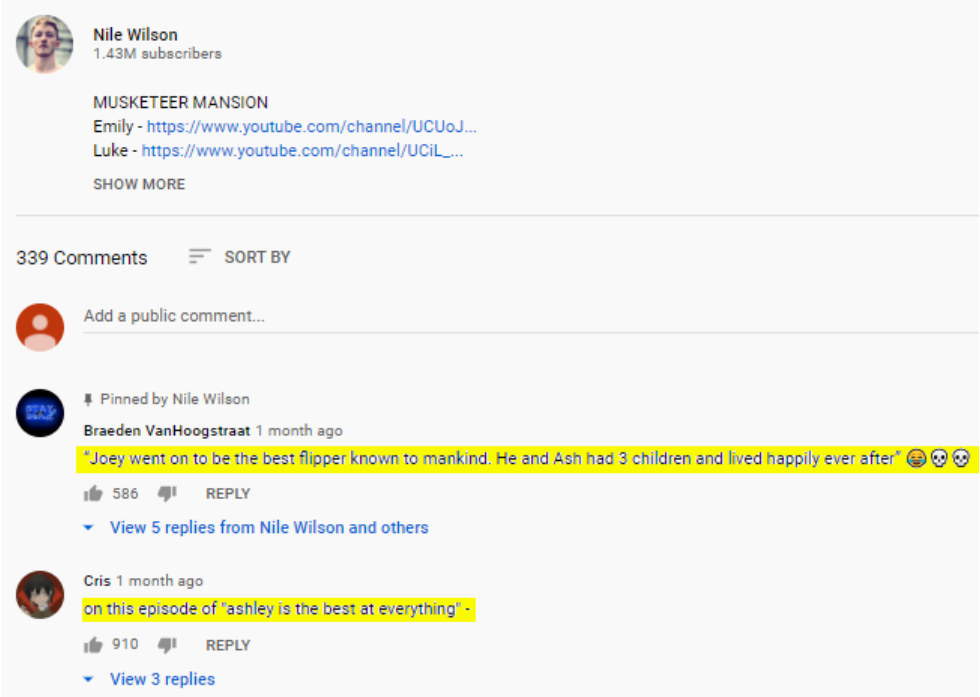
For the data collection, we developed three python scripts, one for each social media. This separation was necessary because social media and its APIs have different particularities that prevent data from being gathered at once.

For Twitter, we used the script previously developed at the DaVInt Lab at PUCRS. This script utilizes Twitter's official API and its complete documentation is available through its paper [53] and also at GitHub [13]. A detailed list of all the collected fields from Twitter is available in Figure 3.4, along with our database tables.

The Instagram script was developed with the support of the research group from DaVInt Lab. This script was split into two parts. The first one collects the profile post link for all the posts belonging to that profile. The second one gathers information regarding the link (number of likes and comments, comment text, among others). The scripts are available at GitHub [12], along with its documentation. Like the Twitter script, all the collected fields from Instagram are also available in Figure 3.4.

Lastly, the YouTube script was developed by using Google's official YouTube API. The YouTube script collects data from profiles and videos and also the video's comments. It is necessary to notice that we chose to get only the first level of the video's comments, i.e., we do not collect comments' answers. We chose this based on the researcher's impression that, in its majority, only the first level provides commentaries about the video. The other levels usually provide support or discussions regarding the first comment, and so they were not considered useful for this study. Figure 3.3 shows an example of the collected comments. We provided full documentation on how to use this script, along with its source code at GitHub [14]. Once again, the detailed data collected is available in Figure 3.4.

Figure 3.3 – YouTube collected comments example.



3.7 Data Preprocessing

After data collection, this study required some data processing for language detection and sentiment analysis before storing the data in the database. In this section, we provide a detailed description of this data processing.

For language detection, since both (Langid and Langdetect) libraries presented issues, there was the need to combine two libraries to obtain better results. The first step is to analyze the text using both of them. If both of them return the same language, then this language is assumed as the text language. If the results differ, then the reliability score of Langid is checked. If Langid score is higher than 80%, the Langid result is assumed to be the text language. Other than that, the text is disregarded. Even though Langid also has limitations, Langid presented more reliable results during the pre-implementation tests, so it was selected to be used in cases that the libraries return different results.

It is important to note that only English comments are stored in our database. Since it is hard to provide sentiment analysis for non-English texts, it was not necessary to store them in our database. This decision does not influence any other charts or statistics because the number of comments is gathered from the post itself.

As already stated in Section 3.4, we used Vader for sentiment analysis of posts comments. Vader provides the compound score that is a standardized score calculated through the lexicon scores. The compound scores vary between -1 and 1. For this work, we considered compound scores greater or equal to -0.5 as a negative text and compound

scores greater or equal to 0.5 as positive. Any value between -0.5 and 0.5 is considered neutral. This analysis happens during the database data registration to improve the chart's performance later on.

3.8 Database

We stored the data collected in a MongoDB database. The database was modeled using eight collections (the "equivalent" of tables in MongoDB), one for each category presented in Section 3.3. For Twitter data, two collections were created: profile and posts. For Instagram and YouTube, three collections were created: one for the profile; one for posts; and the third one for comments. The collections for profile information is unique and does not have a relationship with others. The posts and comments collections are related by unique ids. The detailed database model is available in Figure 3.4.

3.9 Data Cleaning and Enriching

This work provides several statistics data, such as the number of posts, the totals, and an average number of likes and replies between others. These statistics are dynamically calculated according to the user-selected filters. The complete description of all statistics provided by our approach is available in Section 3.10.

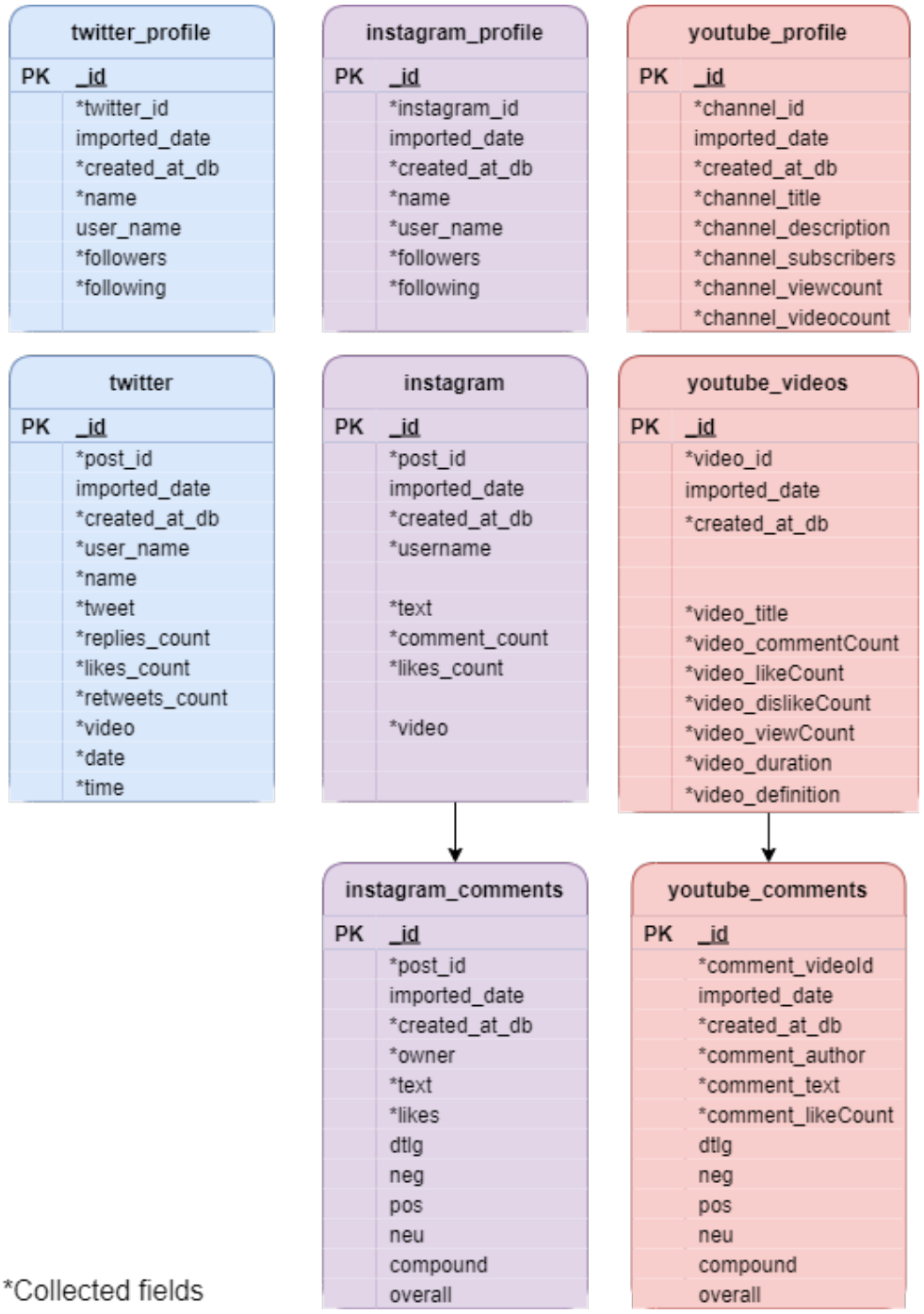
For word cloud generation, first, it is necessary to remove the text stopwords[1]. For this, we use the WordCloud library's stopword removal functionality (see Section 3.4 for description). Since our work only stores English text and the WordCloud library's stopword removal worked very well in our pre-implementation tests, it was unnecessary to implement our technique. This clean-up process happens dynamically each time a post is selected in the Sentiment Analysis Screen (see Section 3.10), and it only happens for Instagram and YouTube comments.

3.10 Interactive Filters and Visualizations

Since our target audience is the general public that might not have much computer knowledge, we chose to develop simple visualization and interface. The visualizations provided are based on classical charts that most people have some familiarity with them. We

[1]The term stopwords refer to words like articles, prepositions, pronouns, conjunctions, and others that do not add much information to the text. In English, words like "the", "a", "an", "so", and "what" are some examples of stopwords.

Figure 3.4 – Dababase model along with collected fields. Respected through social media are shown horizontally aligned.



also used the social media prominent color in our visualizations. This choice facilitates the user to identify to which social media the visualization belongs [20, 17].

It is important to note that, for this study, we define interaction as any reaction possible for a particular social media, such as likes, replies, and comments. Therefore, according to our definition: For Twitter, the interactions considered are likes, retweets, and replies. For Instagram, the interactions are likes and comments. Lastly, for YouTube, the interactions considered are likes, dislikes, views, and comments. In this section, we use interactions for

referring to social media interactions and approach interaction for the interactions provided by our approach.

Some of these interactions can be related to each other across the studied social media. For instance, the "like" interaction is present among the three social media. Meanwhile, the interaction "comment" for Instagram and YouTube is the same as a reply for Twitter. The remaining interactions are unique from their social media and cannot be related to the others.

For this study, we developed visualization based on pie charts, vertical and horizontal bar charts, line charts, heatmap charts, and word cloud charts. Pie charts are known to be used to show proportion among totals [27]. Therefore we used them for comparisons like the number of posts by social media and the total number of positive, negative, and neutral comments from a social media post.

According to Khan and Khan [27]: "Bar chart is use to represent a single data series...". They represent data on the horizontal axis and values on the vertical [27]. In our approach, we used them for comparing interaction values between the same social media, for daily comments classification, and for comparing comments with the highest likes number.

Also, according to Khan and Khan [27]: "The line chart is often used to visualize a trend in data over time interval...". In our approach, line charts were chosen as a form of comparing different time-series over time. We used them to show the values of the number of likes, comments, replies, views, dislikes in the same chart by social media (as long as the social media support it).

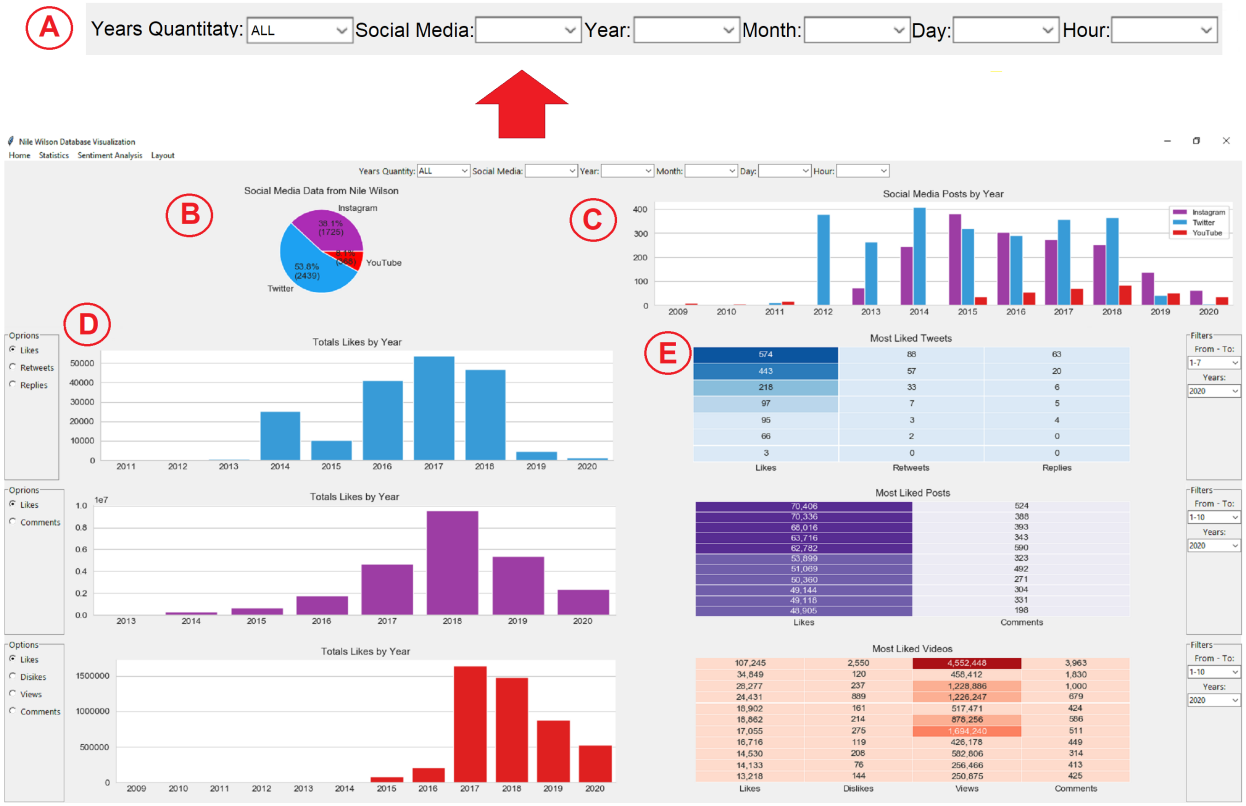
Bojko [8] defines heatmaps as: "two-dimensional graphical representations of data where the values of a variable are shown as colors.". He also states that: "Heatmaps help us quickly see “the big picture” including any patterns or trends that may exist in the data.". In our approach, heatmaps were selected for detailed comparisons between values, showing the number of social media Interaction by post, and allowing to rank the posts according to their number of interactions.

According to Xu et al. [64]: "Word clouds have been widely used to present the contents and themes in the text for summary and visualization". For this approach, they were used to show an overview of the most used words in the comments.

The developed charts were complemented by interactions and filters. Also, we provided a statistics board showing the main statistics of the collected data. The initial screen of our approach is available in Figure 3.5 for the Nile Wilson database. The Nile Wilson database will be used for the remainder of this section, and it is fully described in Section 4.4.

In Figure 3.5A, we provided the general filters available in our approach. Figure 3.5B presents a pie chart that shows the amount and percentage of posts made on

Figure 3.5 – Visual analysis approach's initial screen display: A shows the interactive filters; B shows the global pie chart for post number; C presents the details of B. D presents the interactions bar chart, and D presents the interaction heatmaps.



each social media. The bar chart on the right shows the pie chart data in detail by presenting the number of posts on the three social media during the specified period of time through the filters on 3.5A. It is important to note that the period reflects the selected filters. Both of these charts provide global data information and are a good start point for investigation. The pie chart allows comparing the number of posts on each social media, while the bar chart allows comparing posts number in a specif period.

We also want to mention that the chart's colors chosen in Figure 3.5 and on the statistics screen (Figure 3.10) reflect the social media's most prominent color. Thus the blue color was selected to represent Twitter, purple for Instagram, and red for YouTube.

After the general charts (Figure 3.5B and 3.5C), the screen is divided into three horizontal sections, one for each social media analyzed by our approach: Twitter, Instagram, and YouTube, from top to bottom. The bar charts in Figure 3.5D present the number of interactions in each social media: for Twitter, likes, retweets, and replies; for Instagram, likes and comments; for YouTube likes, dislikes, views, and comments.

The bar charts presented in Figure 3.5D provide the sum of each different interaction during a period of time. These charts are important to show the number of interactions increasing or decreasing, allowing future actions by the brand. For those charts, the default

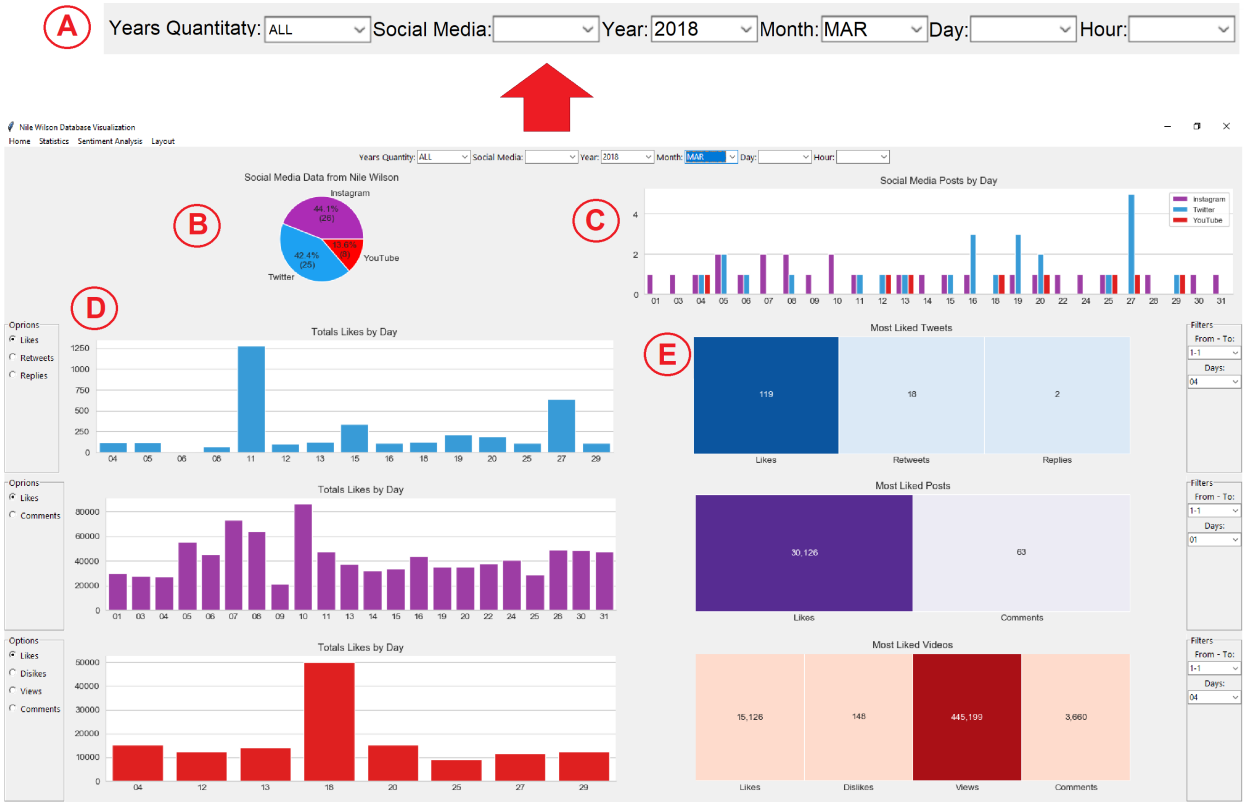
visualization is the number of likes, but it can be changed, e.g., to the number of comments in the radio button on the left of the screen.

The heatmap charts present posts with the highest or lowest number of interactions for each social media. The radio button controls which interaction will be used to sort the chart, e.g., likes or comments. Each chart shows 10 posts at the time except when the post number is lower than 10. This chart shows detailed information about the selected filter. For instance, if a year is selected on the filter, the heatmap chart will display the top posts according to the chosen month on its combo box.

The dashboard of our approach provides several filters (Figure 3.5A) that can be applied: Amount, social media, Year, Month, Day, and Hour. The Amount filter chooses how many years will be shown on the charts. For instance, the user might select that only the last three years are displayed, so the approach will only display data for 2018, 2019, 2020. The social media filter selects one social media to be shown. The other filters are for date periods. So it is possible to filter the graphics by year, month, day, and hour.

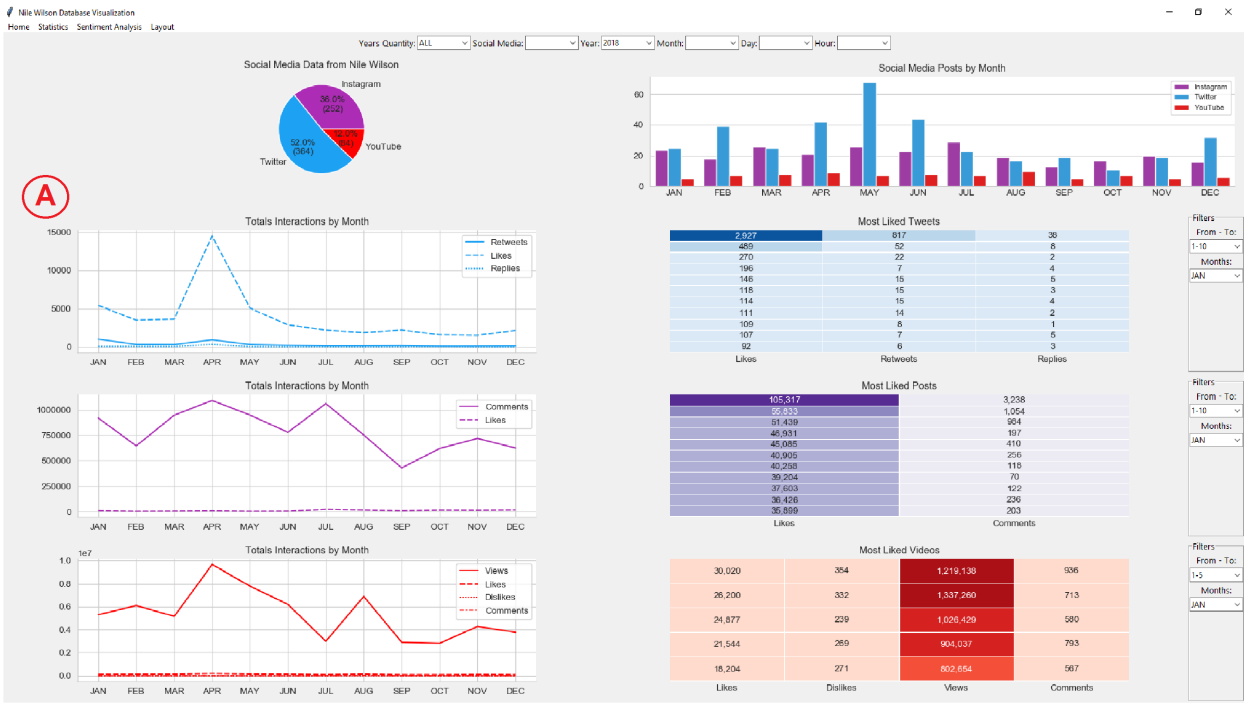
These filters can be applied by the drop-down fields on the top of the screen or by clicking on the graphics bars/pie slice. By clicking on the Heatmap chart is also possible to see the actual post on its social media. An example of the use of the filters is available in Figure 3.6. In Figure 3.6, filters 2018 for the year and march for the month were selected, and the charts reflect it according.

Figure 3.6 – Visual analysis approach's with filters example: A shows the selected filters, 2018 for year and March for month; B, C, D, and E show their respective charts for March 2018.



The bar charts presented in Figures 3.5 and 3.6 can be changed to line charts using the "Lines" option on the menu (Figure 3.8C). Different from the bar chart, the line charts will show a comparison between the totals of interactions. An example of the dashboard with the line charts is provided in Figure 3.7. In the screen provided in Figure 3.7, the radio buttons for changing the interaction type are not available.

Figure 3.7 – Visual analysis approach's screen display with lines charts: A shows line charts instead of bar charts.



The dashboard also provides two more interactions which are: the layout orientation (Figure 3.8D) and the selection of which social media to be shown (Figure 3.8E). The first one allows the user to change the dashboard orientation from horizontal to vertical. The second one allows the user to show or hide charts regarding specific social media data. For instance, the user may hide the charts that belong to Twitter. These interactions are available through menu items which are presented in Figure 3.8 and their interaction results are presented in Figure 3.9.

Besides the dashboard, our work also provides two other screens: statistics and sentiment analysis. Both of these screens support multiple instances and so can be used to compare data and both of them are available through menu options (Figure 3.8A and Figure 3.8B). The statistics screen is provided in Figure 3.10, and the sentiment analysis in Figure 3.11.

The statistics board provides statistical data regarding the collected data. The board is split into two segments: the upper part (Figure 3.10A) shows data about the database as a whole, and it does not change; the bottom part (Figure 3.10B) shows the statistics according to the selected filter on the dashboard. For example, if the year 2018

Figure 3.8 – Visual analysis approach's layout menu. A opens the statistic screen; B opens the sentiment analysis screen; C changes the layout from bar charts to line charts; D chooses the horizontal or vertical screen layout, and E selects which social media will appear

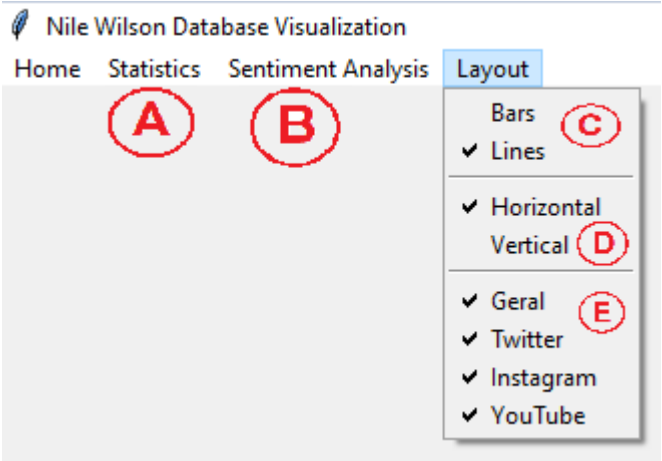
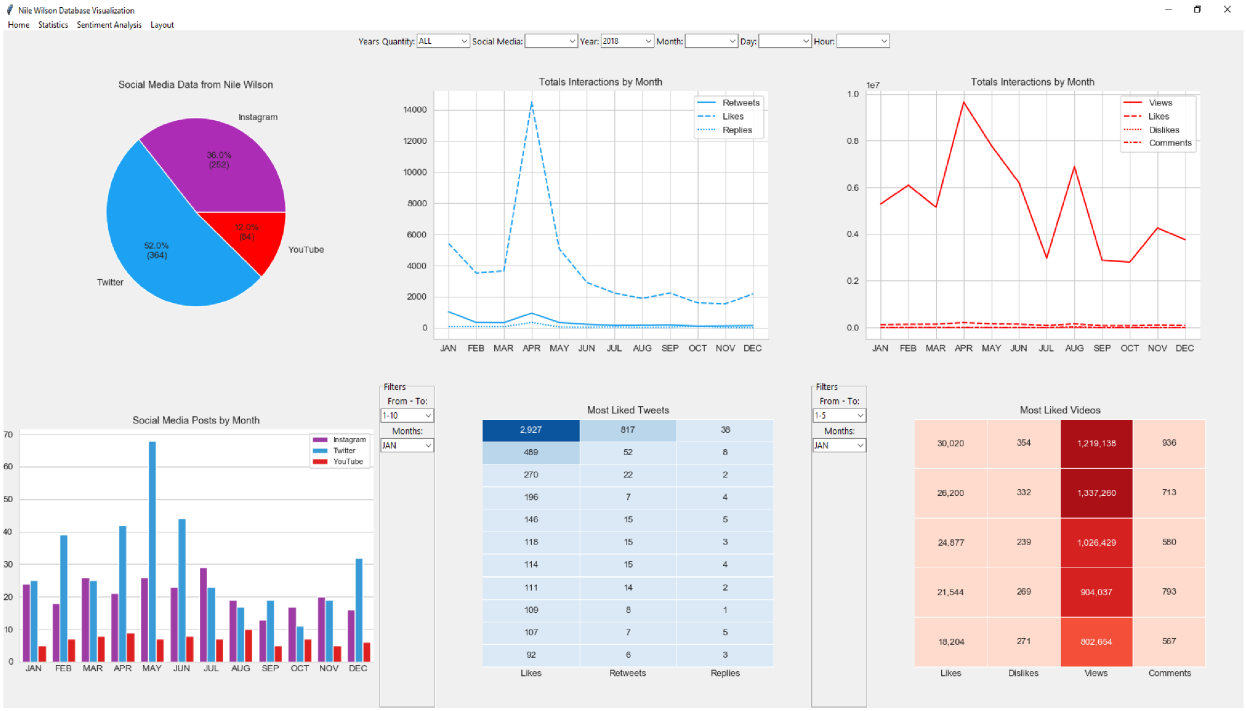
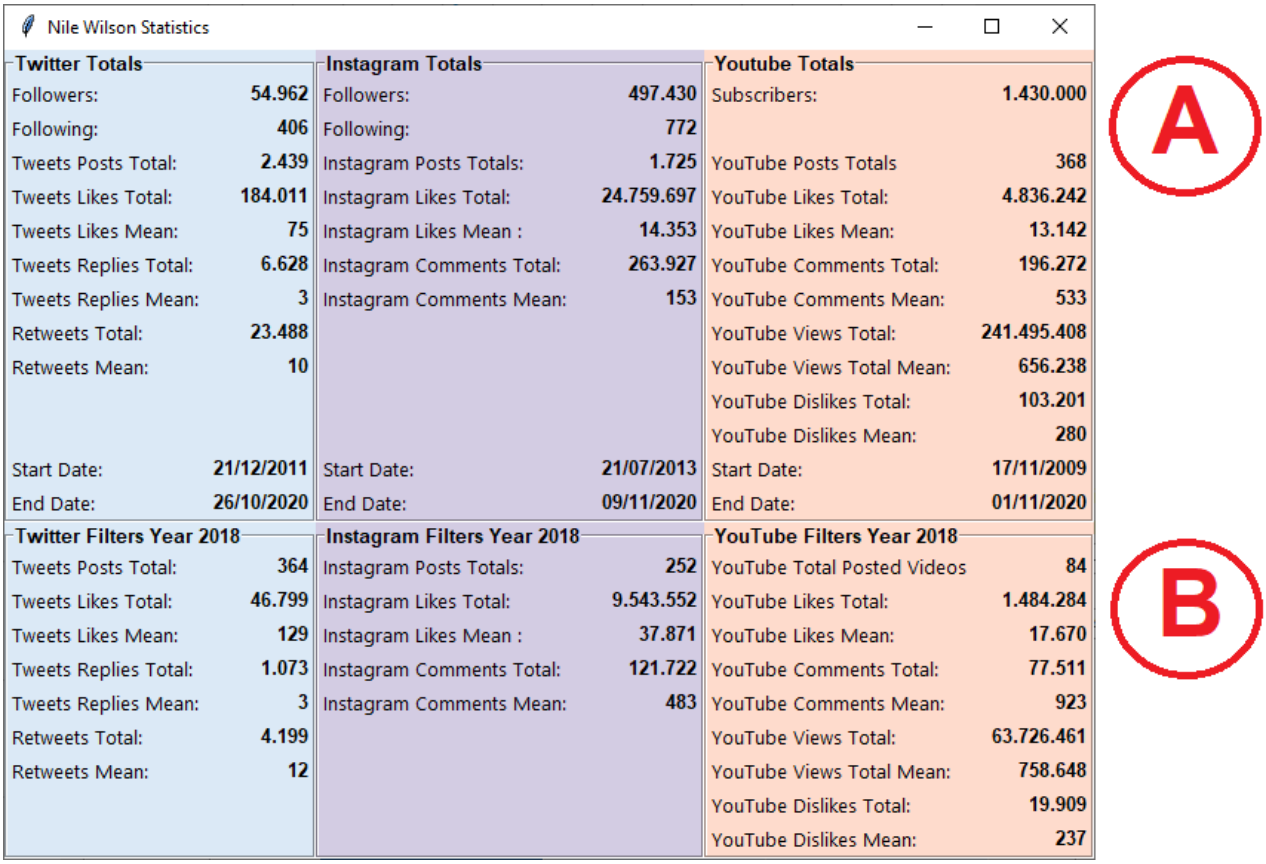


Figure 3.9 – Visual analysis approach with the vertical layout and with Instagram charts hidden.



is defined in the main panel, the bottom part will show statistics for 2018. The statistics we provide for Twitter are the number of followers, number of following, number of posts, retweets total and mean, likes total number and mean, and replies total number and mean. Instagram presents the following statistics number of followers, number of following, number of posts, comments total number and mean, and likes total number and mean. YouTube presents the following statistics number of subscribers, the number of posts, comments total number and mean, likes total number and mean, view total number and mean, and dislikes

Figure 3.10 – Visual analysis approach's statistics example: A presents the statistics for the database as a whole, and B presents the statistics according to the selected filter. In this case, the filter selected is the year 2018.



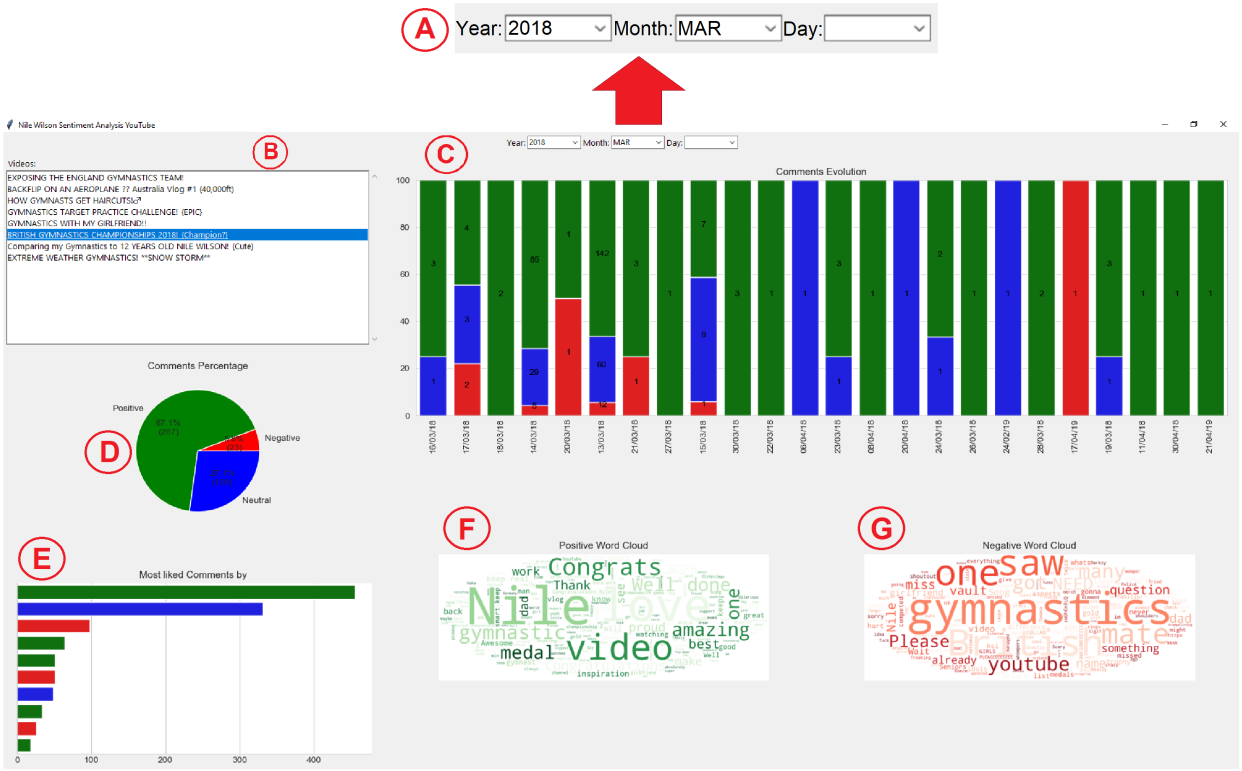
total number and mean. The statistics displayed on the same row are the ones that can be related through the selected social media as already presented.

Figure 3.11 presents our Sentiment Analysis screen. This screen shows in Figure 3.11B the list of posts for the selected social media on the menu. This list is built according to the dashboard filter or the filters (Figure 3.11A) provided by itself. It is important to note that, for this screen, the only filters available are Year, Mount, and Day. The main goal of the filters provided on this screen is to reduce and so clarify the list of posts.

In Figure 3.11D, a Pie Chart shows the percentage of positive, negative and neutral comments. Figure 3.11E presents the 10 comments with the highest number of likes with the comment classification represented by the bar color.

The Vertical bar chart, in Figure 3.11C, shows the comments "evolution", i.e., each bar contains the total number of positive, negative, and neutral comments received each day. In the case of videos with a high number of comments (for clarification purposes), the comments are then grouped by weeks or even two weeks, according to necessity. Finally, two word clouds are presented: one generated from positive (Figure 3.11F) comments; and the other from negative (Figure 3.11G) ones.

Figure 3.11 – Visual analysis approach's sentiment analysis: A presents the filters used to filter the list of posts in B; B displays the post list; C shows the number of positive, neutral, and negatives by day; D provides the overall of positive, neutral, and negatives for this post; E provides the rank of the most liked comments; F provides the positive comments word cloud, and G provides the negative comments word cloud.



4. CASE STUDY

This chapter presents the three case studies used to validate our approach. Two of them are about the major streaming platforms, Netflix and Amazon Prime Video, which are very active on social media. Moreover, since they also belong to the same segment, it is possible to make comparisons between them.

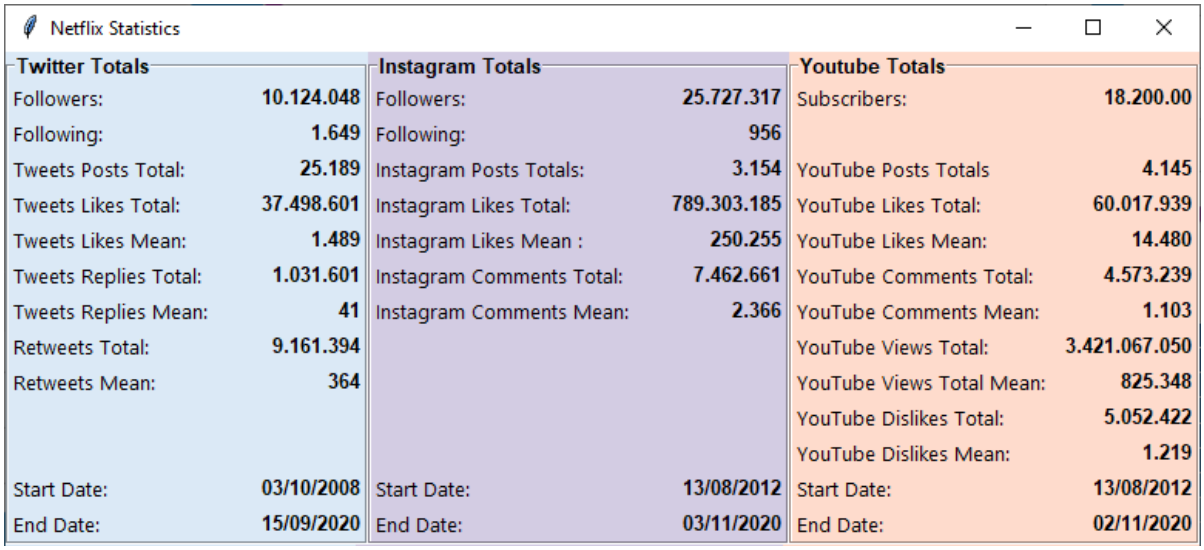
Since we consider that a person can also be treated as a brand, our third case study regards olympian athlete Nile Wilson. Nile Wilson is a British artistic gymnast that won a bronze medal at the Rio de Janeiro summer Olympic games. Combining the three analyzed social media, he has more than 1.000.000 followers and also has a massive social media presence. Section 4.1 presents the details about data collections, and the next sections detail our findings of each case study.

4.1 Data Collection

One important observation regarding the data collected is that it was gathered from different periods. For the start date, the reason is that we collected the entire profile data, and each profile was created on a different date. For the end date, since we use three separate collection scripts that worked at different speeds, it became hard to synchronize it.

The Netflix data were collected as follows: from 03/10/2008 to 15/09/2020 for Twitter; from 13/08/2012 to 03/11/2020 for Instagram; and from 17/06/2012 to 03/11/2020 for YouTube. The detailed information on the collected data can be seen in Figure 4.1. From this Figure, it is already possible to note that Twitter has the highest number of posts, but the highest number of interactions happens on Instagram.

Figure 4.1 – Netflix global statistics



Amazon Prime Video data were collected as follows: from 07/11/2020 to 29/08/2020 for Twitter; from 16/11/2016 to 06/09/2020 for Instagram; and from 06/02/2014 to 08/09/2020 for YouTube. The completed information for the data collected is displayed in Figure 4.2. For Amazon Prime Video, it is also possible to note that, like Netflix, the highest number of posts occur on Twitter, but Instagram has the highest interaction number.

Figure 4.2 – Amazon Prime Video global statistics

Amazon Prime Video Statistics

Twitter Totals		Instagram Totals		Youtube Totals	
Followers:	1.773.927	Followers:	1.421.339	Subscribers:	829.000
Following:	715	Following:	514		
Tweets Posts Total:	53.515	Instagram Posts Totals:	1.207	YouTube Posts Totals	620
Tweets Likes Total:	3.143.811	Instagram Likes Total:	27.760.122	YouTube Likes Total:	1.947.315
Tweets Likes Mean:	59	Instagram Likes Mean :	22.999	YouTube Likes Mean:	3.141
Tweets Replies Total:	148.037	Instagram Comments Total:	241.480	YouTube Comments Total:	152.095
Tweets Replies Mean:	3	Instagram Comments Mean:	200	YouTube Comments Mean:	245
Retweets Total:	564.876			YouTube Views Total:	466.737.168
Retweets Mean:	11			YouTube Views Total Mean:	752.802
				YouTube Dislikes Total:	103.852
				YouTube Dislikes Mean:	168
Start Date:	07/11/2008	Start Date:	16/11/2016	Start Date:	06/02/2014
End Date:	29/08/2020	End Date:	06/09/2020	End Date:	08/09/2020

The Nile Wilson data were collected as follows: from 21/12/2011 to 26/10/2020 for Twitter; from 21/07/2013 to 09/11/2020 for Instagram; and from 17/11/2009 to 01/11/2020 for YouTube. Details regarding the collected data can be seen in Figure 4.3. For Nile Wilson, the same phenomenon observed for Netflix and Amazon Prime Video also happens. The highest post number occurred on Twitter, but the highest interactions happened on Instagram.

Figure 4.3 – Nile Wilson global statistics

Nile Wilson Statistics

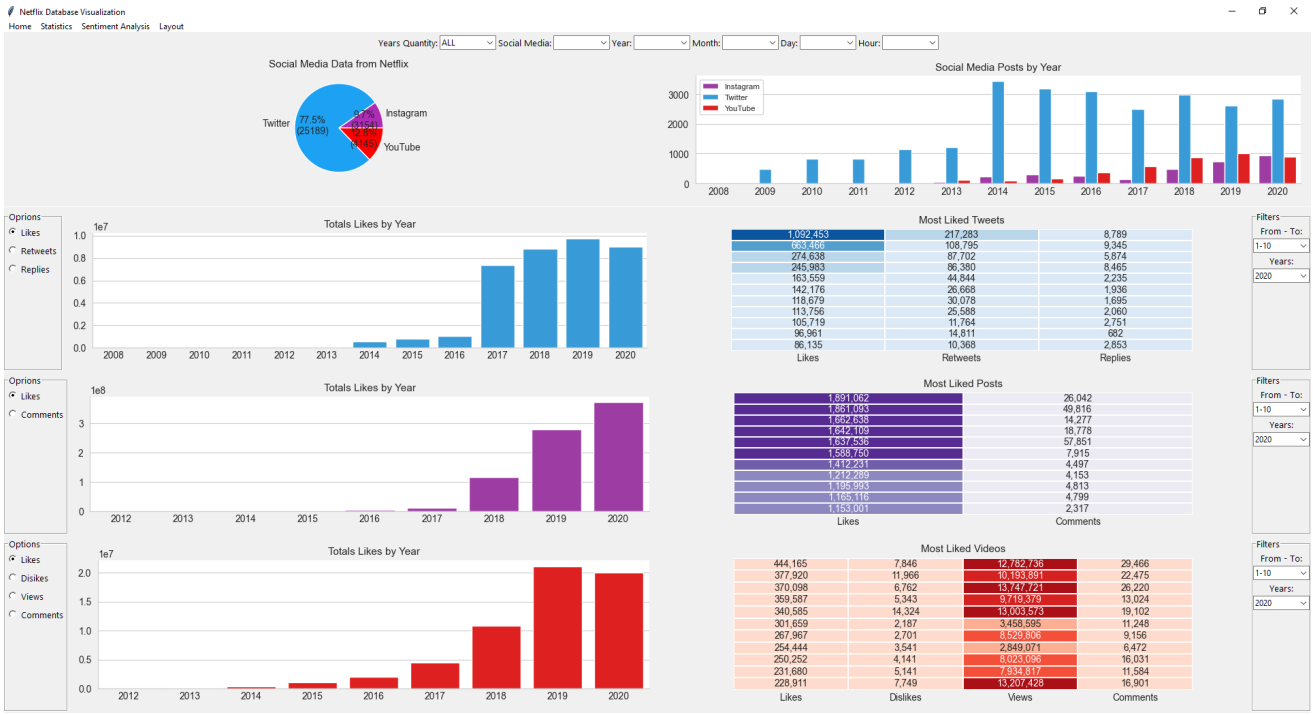
Twitter Totals		Instagram Totals		Youtube Totals	
Followers:	54.962	Followers:	497.430	Subscribers:	1.430.000
Following:	406	Following:	772		
Tweets Posts Total:	2.439	Instagram Posts Totals:	1.725	YouTube Posts Totals	368
Tweets Likes Total:	184.011	Instagram Likes Total:	24.759.697	YouTube Likes Total:	4.836.242
Tweets Likes Mean:	75	Instagram Likes Mean :	14.353	YouTube Likes Mean:	13.142
Tweets Replies Total:	6.628	Instagram Comments Total:	263.927	YouTube Comments Total:	196.272
Tweets Replies Mean:	3	Instagram Comments Mean:	153	YouTube Comments Mean:	533
Retweets Total:	23.488			YouTube Views Total:	241.495.408
Retweets Mean:	10			YouTube Views Total Mean:	656.238
				YouTube Dislikes Total:	103.201
				YouTube Dislikes Mean:	280
Start Date:	21/12/2011	Start Date:	21/07/2013	Start Date:	17/11/2009
End Date:	26/10/2020	End Date:	09/11/2020	End Date:	01/11/2020

4.2 Netflix

Netflix was founded in 1997 as a DVD rental service. But, in 2007, their business model was changed to media streaming [43]. Netflix is known to have close contact with its users through social media, confirmed by the high number of user interactions. It also has more than 50 million subscribers[1]. Netflix has joined Twitter in October 2009, Instagram in August 2012, and YouTube in July 2012.

Presented in Figure 4.4, the main dashboard from our approach allows the identification of points of interest. For instance, with the collected Netflix data, comparing the bar charts for the number of tweets with the one for the number of likes is possible to notice that the number of likes for tweets got a huge increase from 2016 to 2017, but the number of tweets decreased in the same period. Relating this information with the statistics data is possible to notice that in 2016 Netflix tweeted 3097 times but only got an average of 344 likes, 10 replies, and 105 retweets. In contrast, in 2017, the company tweeted 2,051 and received an average of 2,940 likes, 65 replies, and 958 retweets. Verifying the statistics for the following years, we can observe the same phenomenon for 2018, 2019, and 2020. From these years, 2018 had the higher number of tweets but the lowest number of interactions. Considering this information, we can conclude that a high number of posts does not necessarily translate into a high number of interactions for Twitter.

Figure 4.4 – Netflix's main dashboard.



[1] Data collected in December 2020 considering all of our social media combined.

Although this is true for Twitter, the dashboard did not indicate that this aspect would be replicated for other social media. Verifying the statistical data for Instagram and YouTube is possible to observe that, generally, the higher number of posts, the higher the number of interactions. There was only one exception between 2016 and 2017 for Instagram when the number of posts decreased (from 257 to 138), but the interaction number average increased (from 22,898 to 86,337 likes and from 658 to 2,225 comments).

Considering the number of likes for Instagram is possible to notice that the number of likes suffered a significant growth between the years 2017 to 2018 and once again from 2018 to 2019. Meanwhile, the number of comments only presented a significant growth between 2017 and 2018. From 2018 to 2019, the number of comments increased, but this rise was not that significant.

Like Instagram, YouTube also experienced significant growth in the likes number from 2017 to 2018 and from 2018 to 2019 when the number of likes was twice as high compared to previous years. But, in contrast, the number of views experienced a linear increase, and thus no significant rise could be observed. Regarding the number of comments, a substantial increase is noticeable between the years 2018 and 2019. Meanwhile, the number of dislikes behaved differently. At first, it presented its first significant growth from 2016 to 2017, like the others. But after this growth, the numbers remained stable until 2020, when the number of dislikes skyrocketed.This information is shown in Figures 4.5 and 4.6

We speculate that those growths may have occurred due to the increase in Netflix's popularity and, consequently, the increase of subscribers/followers. Unfortunately, we cannot support this theory due to a lack of subscribers' data from previous years.

Since our dashboard allows comparisons among social media is possible to notice that the largest number of posts occurs on Twitter, but Instagram and YouTube generate more interactions. Even if we compare the values by their average, Instagram and YouTube have much higher values for likes and comments than Twitter. For instance, Twitter has 2.833 posts in 2020, but the average of likes is 3.194 per post. In contrast, Instagram and YouTube have 950 and 907 posts, respectively, but their average number of likes is 391.347 for Instagram and 22.112 for YouTube. This information is shown in Figure 4.7.

We theorize that this difference occurs due to the different dynamics between those social media. Twitter is the most volatile social media among those three. Thus, the tweets only remain on the user's timeline for a short period. Instagram also has this dynamic, but posts update are not as fast as Twitter. On the other hand, YouTube's updates depend on new videos releases, and the feed is not as important as it is for the other two. In YouTube, the user generally searches for what he/she wants to see.

Since already stated, the Netflix dislike number in 2020 is much higher than the previous years. This outlier number of dislikes deserved to be examined in detail. Examining the chart closely, we could find that the major number of dislikes were concentrated in the "Mignonnes/Cuties" video. This video is the only one in our hole Netflix database with more

Figure 4.5 – Netflix statistics major raises.

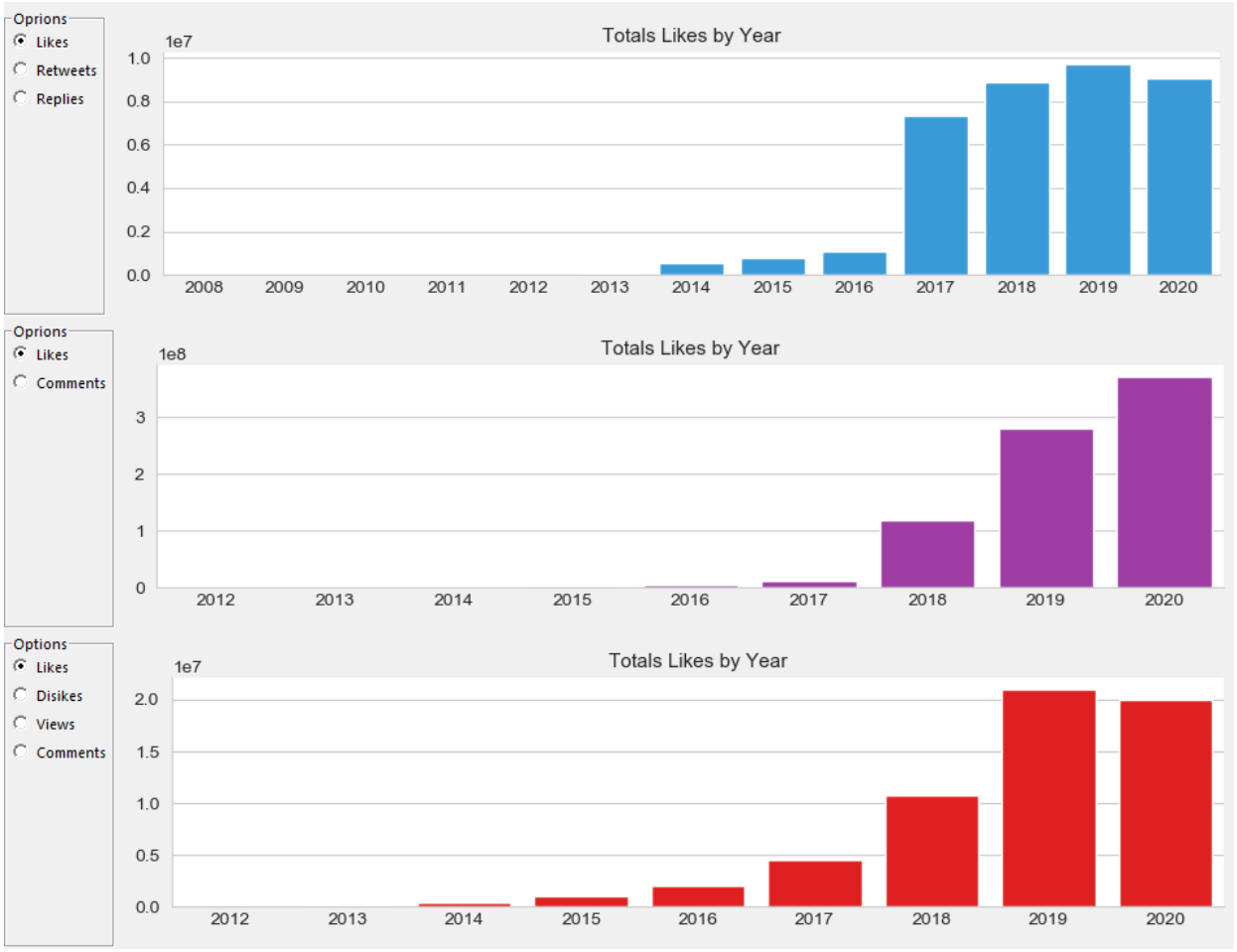


Figure 4.6 – Netflix statistics dislikes raises.

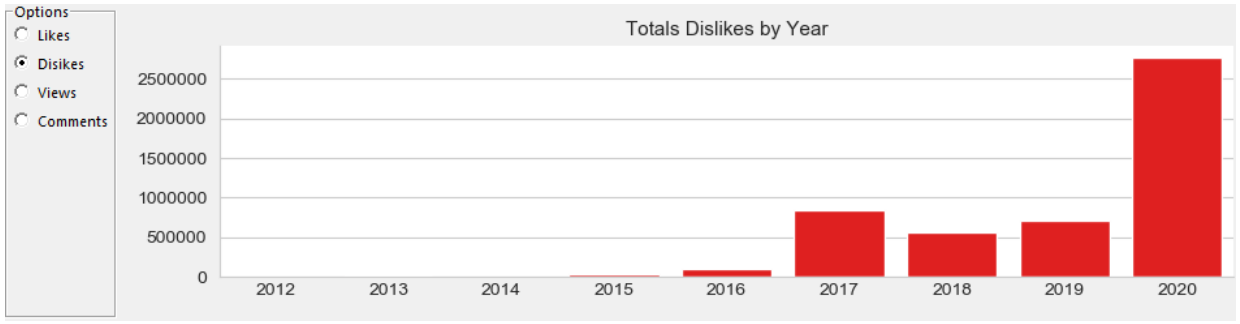
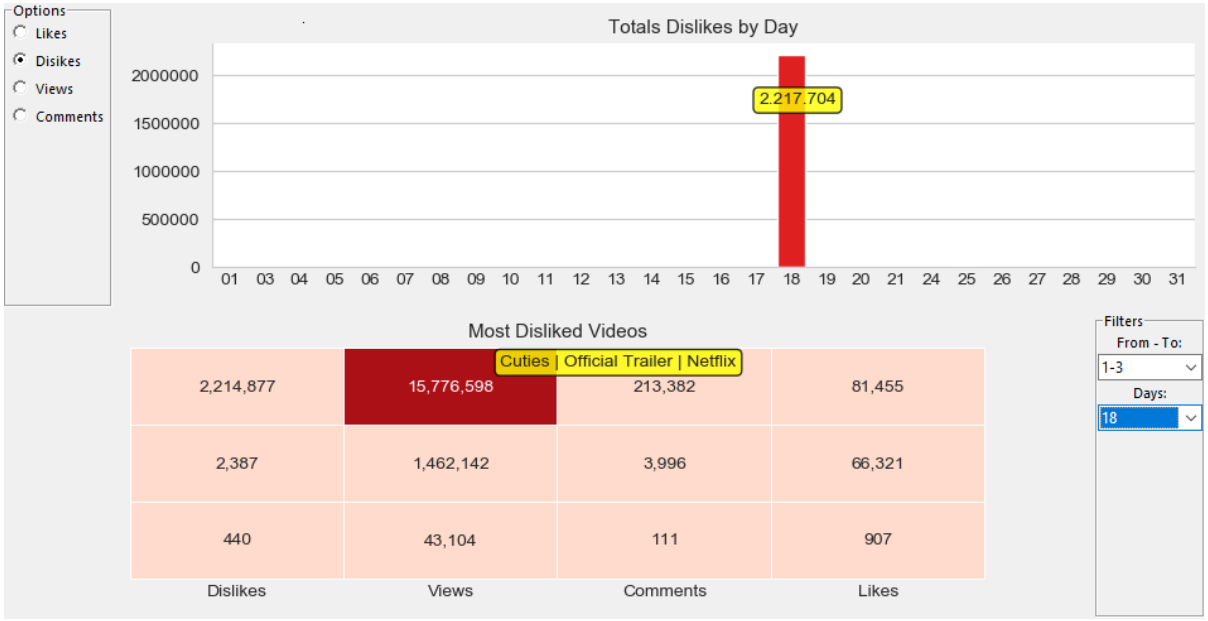


Figure 4.7 – Netflix statistics for 2020.

Twitter Filters Year 2020		Instagram Filters Year 2020		YouTube Filters Year 2020	
Tweets Posts Total:	2.833	Instagram Posts Totals:	950	YouTube Total Posted Videos	907
Retweets Total:	1.679.112	Instagram Comments Total:	2.527.621	YouTube Views Total:	831.663.985
Retweets Mean:	593	Instagram Comments Mean:	2.661	YouTube Views Total Mean:	916.939
Tweets Likes Total:	9.049.035	Instagram Likes Total:	371.780.403	YouTube Likes Total:	20.055.984
Tweets Likes Mean:	3.194	Instagram Likes Mean :	391.348	YouTube Likes Mean:	22.112
Tweets Replies Total:	265.056			YouTube Dislikes Total:	2.772.950
Tweets Replies Mean:	94			YouTube Dislikes Mean:	3.057
				YouTube Comments Total:	1.532.612
				YouTube Comments Mean:	1.690

than two million dislikes. All the other Netflix YouTube videos have less than 100 thousand dislikes except for the one “Dear white people” show. This video also generated controversy due to the diversity agenda. The charts shown in Figure 4.8 present the Mingones’ poor repercussion.

Figure 4.8 – Netflix dislike information.



Analyzing the comments for the “Mignonnes/Cuties” post, we can observe that most of them were negative. In fact, 47,7% of the overall comments were negative, 23% were neutral, and only 29,3% were positive. Considering the comments timeline is possible to notice that, only on the first day, the positive comments were higher than the negatives. For all the other days, the negative comments were higher. Also, its word cloud highlights very negative words like “bad”, “disgusting”, and “shit”. Comments analysis are available in Figure 4.9.

“Mignonnes/Cuties” is a french movie that generated lots of controversies due to Netflix marketing. Netflix has changed the original movie poster and description for another one more provocative. This new poster was accused of over sexualizing young girls and promoting pedophile. Netflix apologized for this change two days later on Twitter. But the apology did not stop the negative comments for the movie, as seen in Figure 4.9. As shown in Figure 4.10, the apology tweet was the most replied one from 2020.

We can also notice that a positive diversity agenda has a major impact on the number of likes. The two most liked and retweeted posts of 2020 are about the “Black lives matter” movement and support for the black community. Each of these posts had twice the number of likes than the one in third place. Even if we compare with the other years, they

Figure 4.9 – Netflix comments evolution for “Mignonnes/Cuties”.

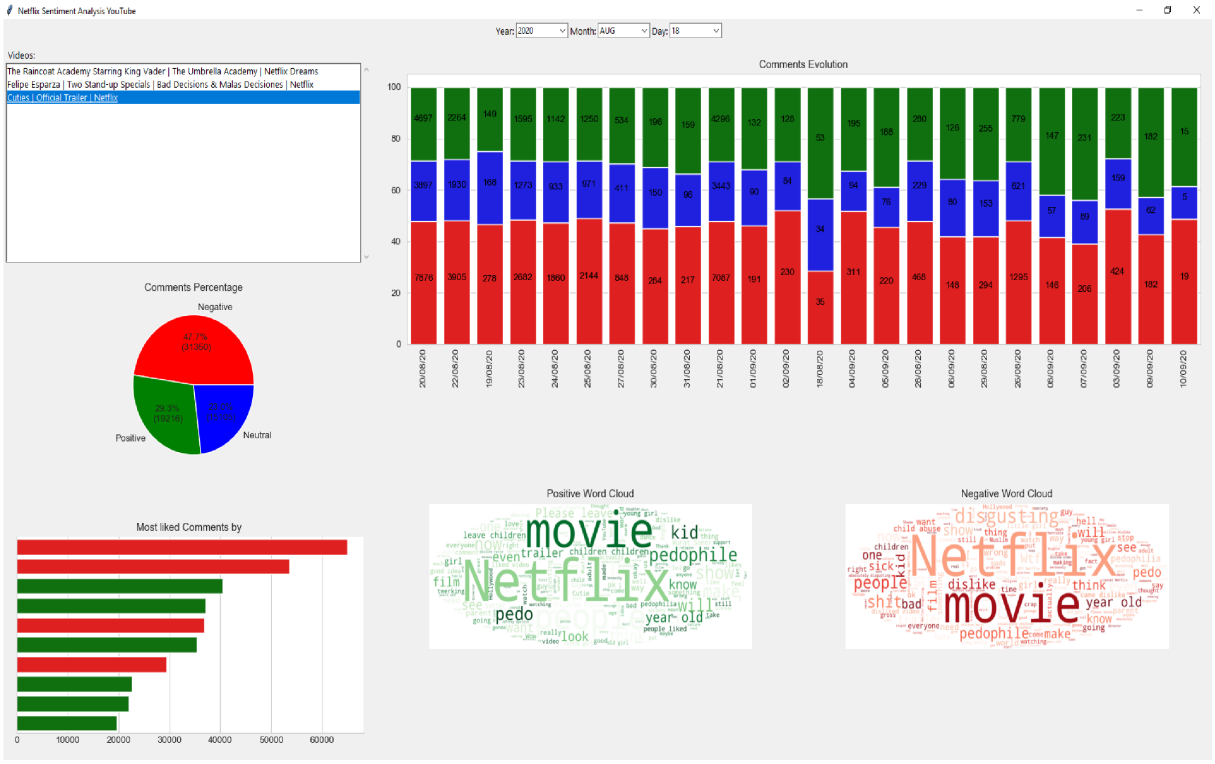
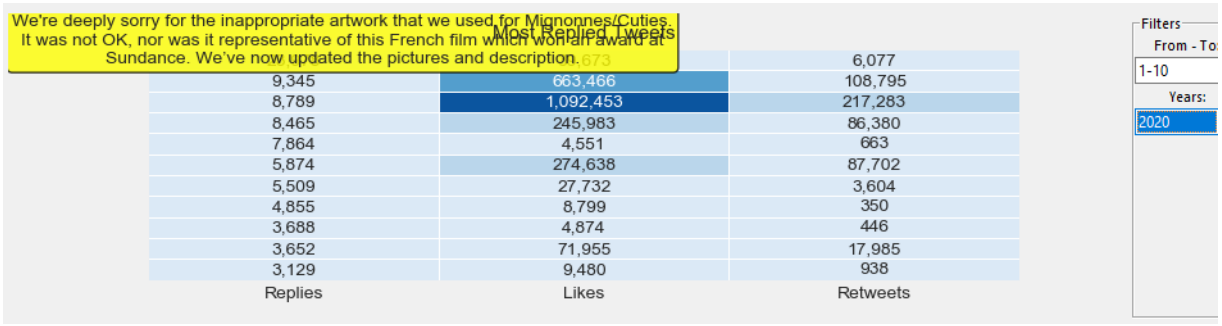


Figure 4.10 – Netflix comments evolution for “Mignonnes/Cuties”.

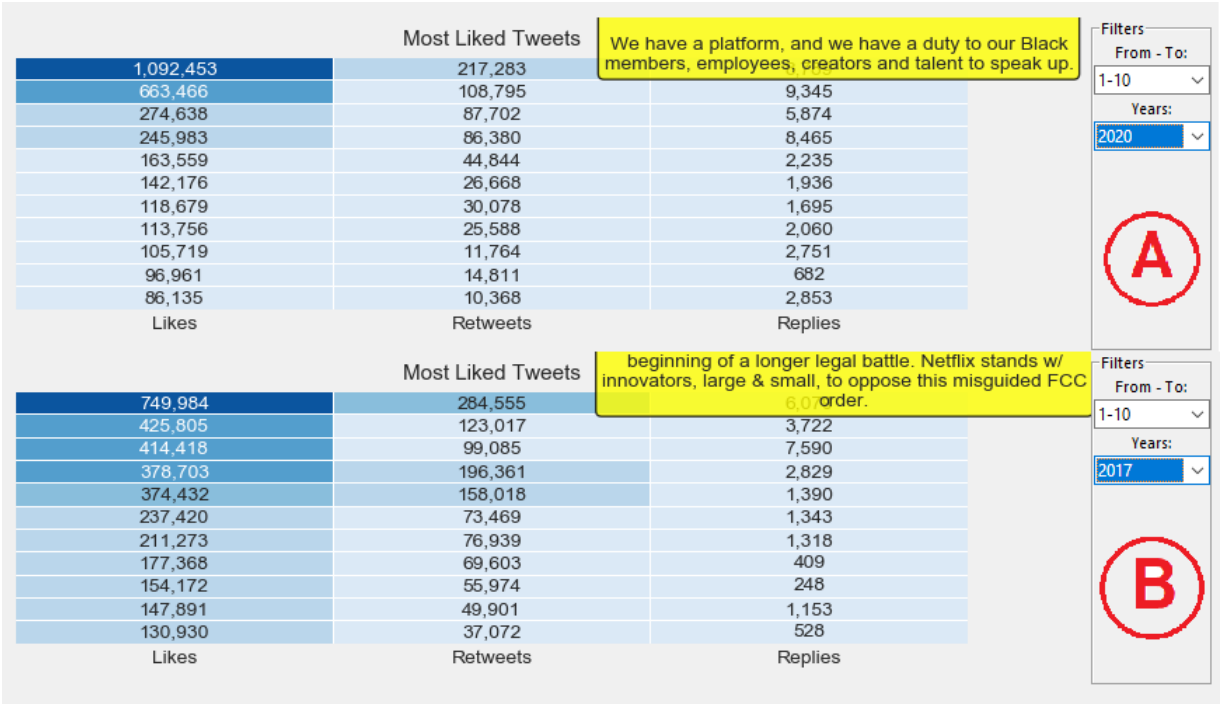


remain in the top positions. The only exception is a post from 2017 about “Net neutrality”[2] that has slightly more likes than the one about black community support. Even if we compare with the other years, they remain in the top positions 4.11.

On the other hand, the high likes and retweets number did not reflect a huge number of replies. Although these two posts occupy the second and third positions at the top replied posts, the post with the highest number of replies was regarding the movie “Mignonnes/Cuties”. This post had three times the number of replies than the second one. Thus we can infer that people use likes for support, but they use replies for controversy in the case of Netflix.

[2]Net neutrality is a movement that believes that "owners of the networks that compose and provide access to the internet should not control how consumers lawfully use that network, and they should not be able to discriminate against content provider access to that network" according to Gilroy [24]

Figure 4.11 – Netflix most liked tweets from 2020 and 2017. A represents 2020 and B represents 2017.



4.3 Amazon Prime Video

Amazon Prime Video [4] debut in September 2006, as Amazon Unbox. In September 2008, the service was renamed to Amazon Video on Demand, and in February 2011 to Amazon Instant Video. In this period, 5000 movies and TV shows were added to its catalog. In September 2015, the service assumed its definitive name Amazon Prime Video.

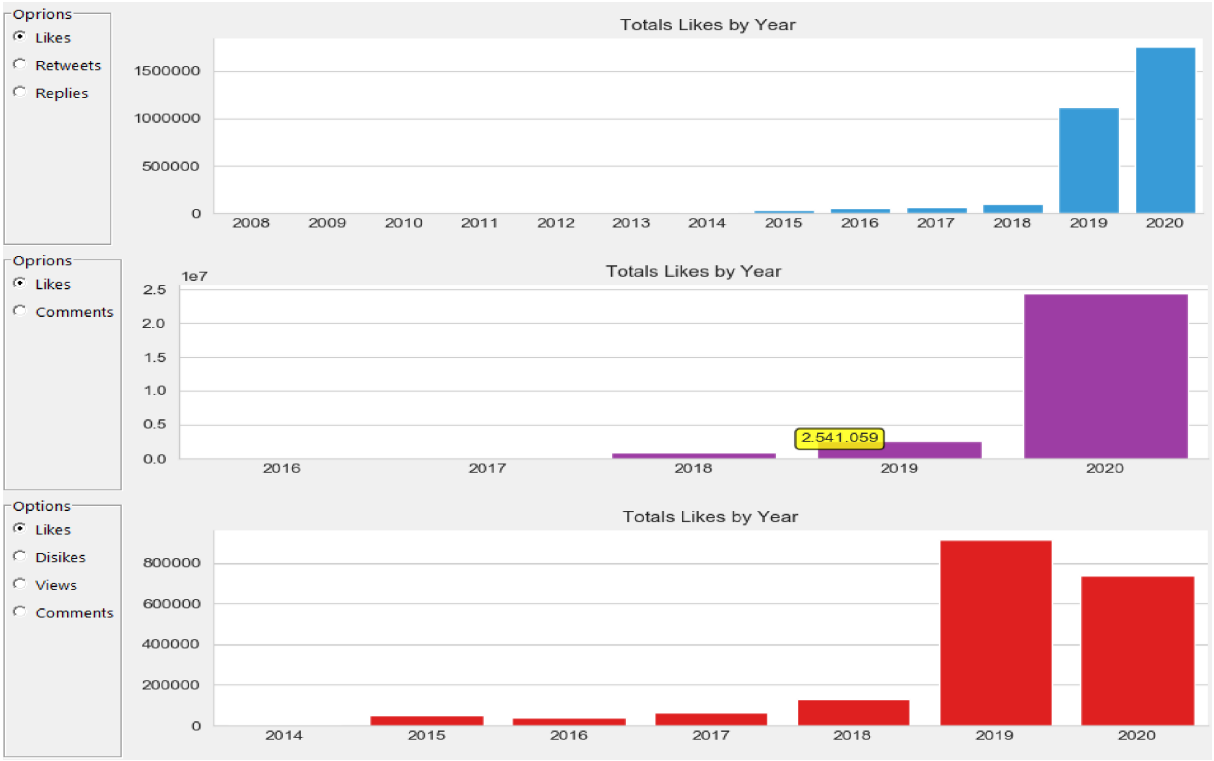
Amazon Prime data presents some of the same aspects as Netflix. For instance, like Netflix, Amazon Prime Video also experienced exponential growths, but they happened differently from Netflix. While Netflix experienced these growths spread in two years, Amazon Prime Video growth was condensed in only one year. For Twitter interactions, the growth happened from 2018 to 2019. However, for the number of replies, it is possible to notice a significant decrease from 2017 to 2018.

For Instagram, both the number of likes and comments skyrocketed from 2019 to 2020. For YouTube, it occurred from 2018 to 2019 for likes, views, and comments numbers. The number of dislikes had a linear growth and did not have a significant increase in any year. The only exception was 2016, where the number had slightly decreased.

Another aspect that presented similarities between Netflix and Amazon is that a high number of posts did not translate into a high number of likes. For example, the year 2017 had 38.829 tweets, but the number of likes, replies, and retweets was much lower

than the years 2019 and 2020, which had 3.846 and 2.603 tweets' respectively. Figure 4.12 shows Amazon Prime Video major growths.

Figure 4.12 – Amazon Prime Video statistics major raises.



For Amazon Prime Video, the highest number of posts occur on Twitter, but the ones that cause the highest number of interactions are the Instagram posts. This is also shown in Figure 4.12.

Similar to Netflix, the diversity agenda also generated a significant number of likes for Amazon Prime Video. The most liked tweet for 2020 (and second-most in general) and the most liked Instagram post regards the diversity agenda. The tweet had 152.746,00 likes against 51.776,00 of the second most liked. For Instagram, the most-liked post had 858.006,00 likes against 664.303,00 of the second one. This is shown in 4.13

Analyzing the comments evolution for the Instagram diversity post, it is possible to notice that most comments were considered positive or neutral. Also, the positive word cloud contains words like "black woman" and "love", suggesting support for diversity. See figure 4.14 for details.

But even though a lot of comments were made, it is possible to notice the same phenomenon observed for Netflix: Although the diversity agenda generates lots of likes, the number of replies remains average and does not stand out from the others. This reinforces our finding that people use likes to support (Figure 4.13).

In 2017 Amazon Prime Video made a promotion offering a free pizza for whoever made a tweet using two selected hashtags. At first glance, it seems that they had lots of replies on that subject. But, if we check the replies, it is possible to notice that several of

Figure 4.13 – Amazon Prime Video highest liked tweet and Instagram post.

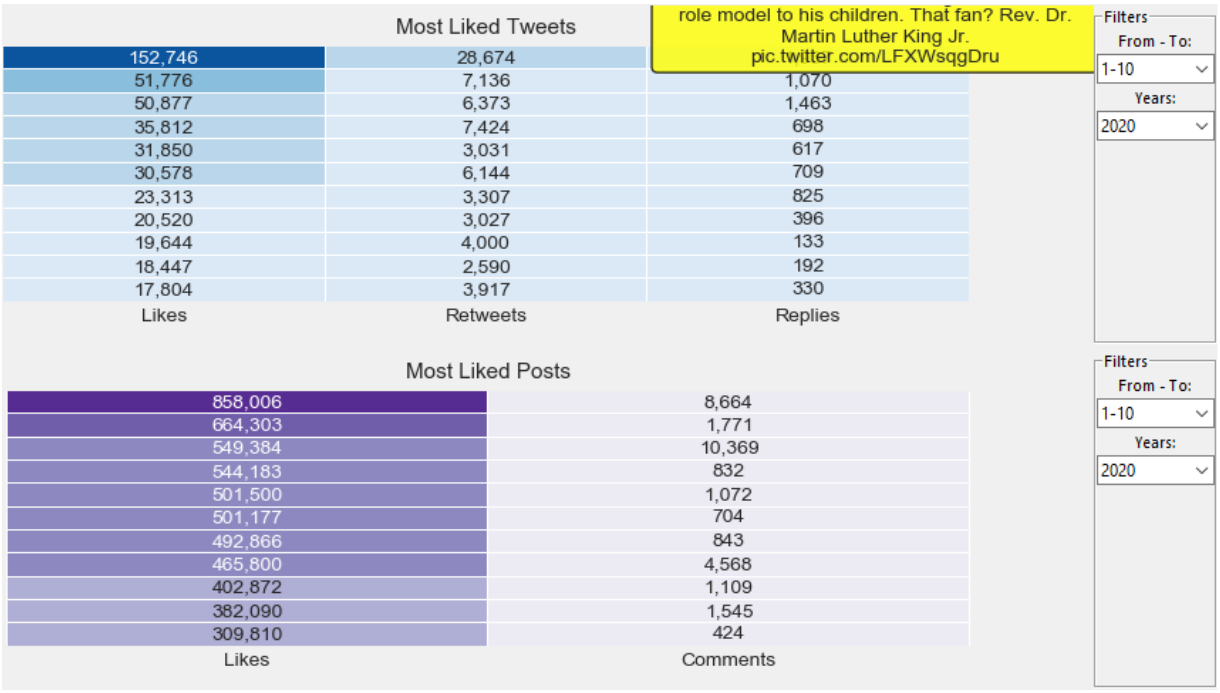
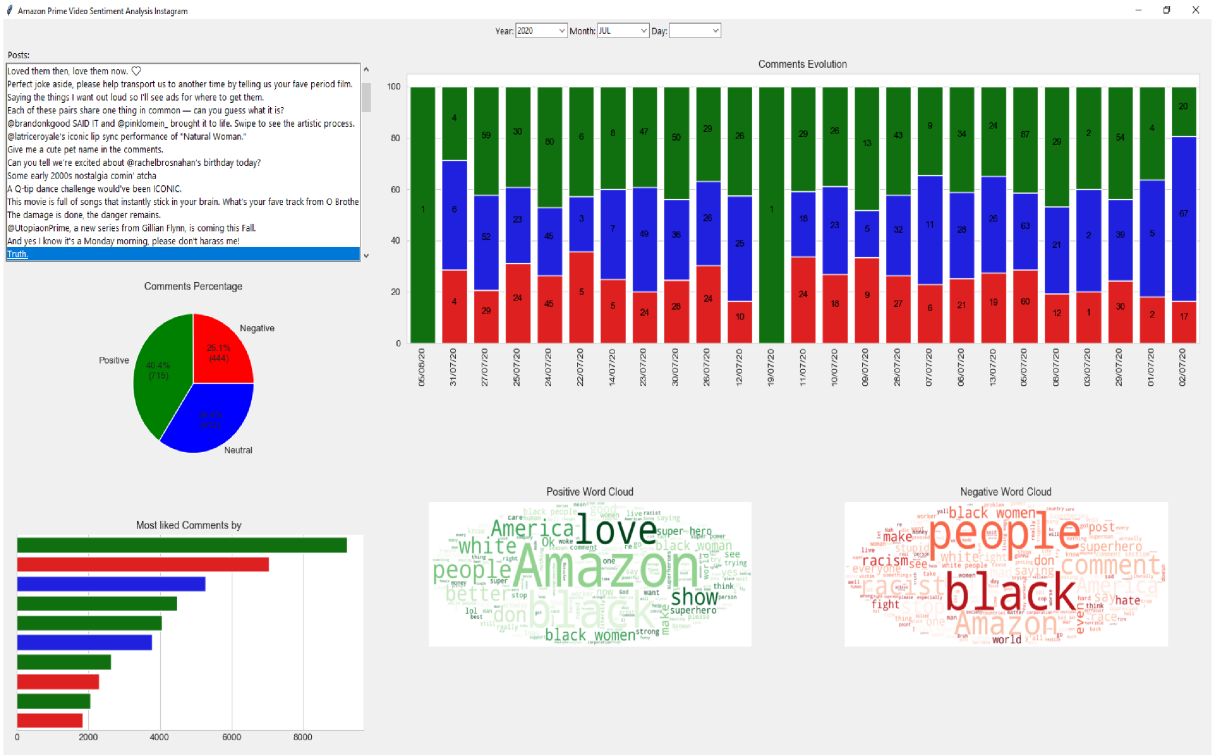


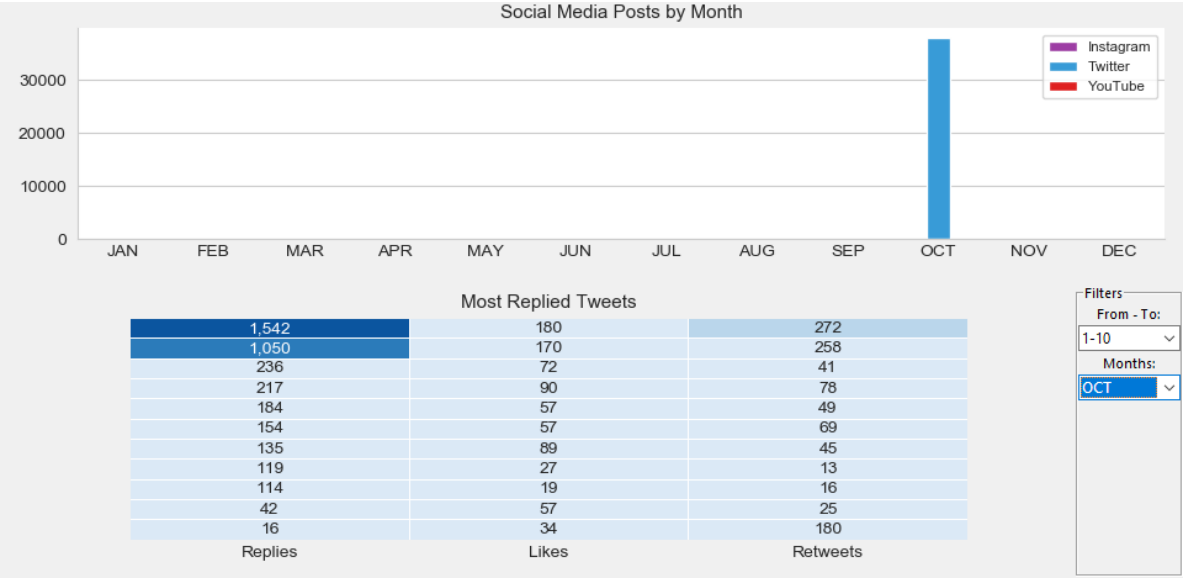
Figure 4.14 – Amazon Prime Video Instagram diversity's comment evolution.



them are from Amazon Prime itself, congratulating the winner customers or consoling losers. So this does not indicate that the promotion was either a success or a failure. This can be seen in Figure 4.15.

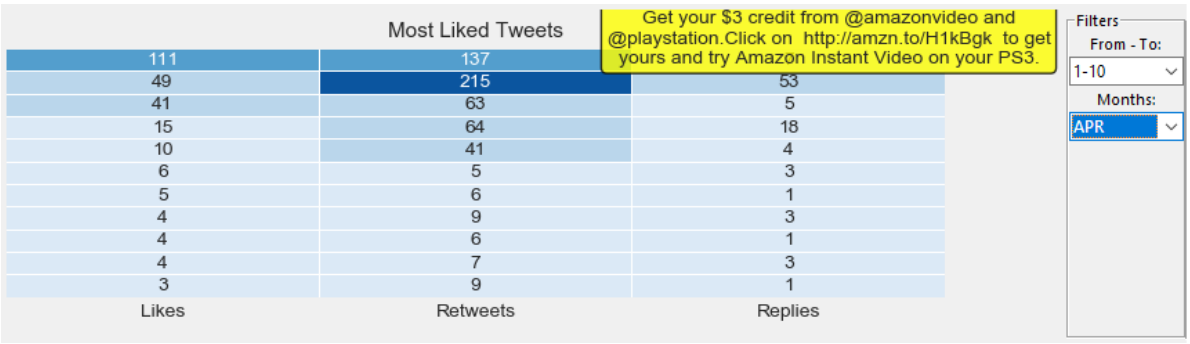
Another promotion regarding credit for using Amazon Prime Video on PlayStation was made in April 2012. However, the number of likes, replies, and retweets was low and

Figure 4.15 – Amazon Prime Video posts number and replies in October 2017 during the pizza promotion.



cannot be used to indicate if the promotion was either a success or a failure. So, regarding these promotions, we can conclude that they were not sufficient to generate a high number of interactions. See Figure 4.16 for details.

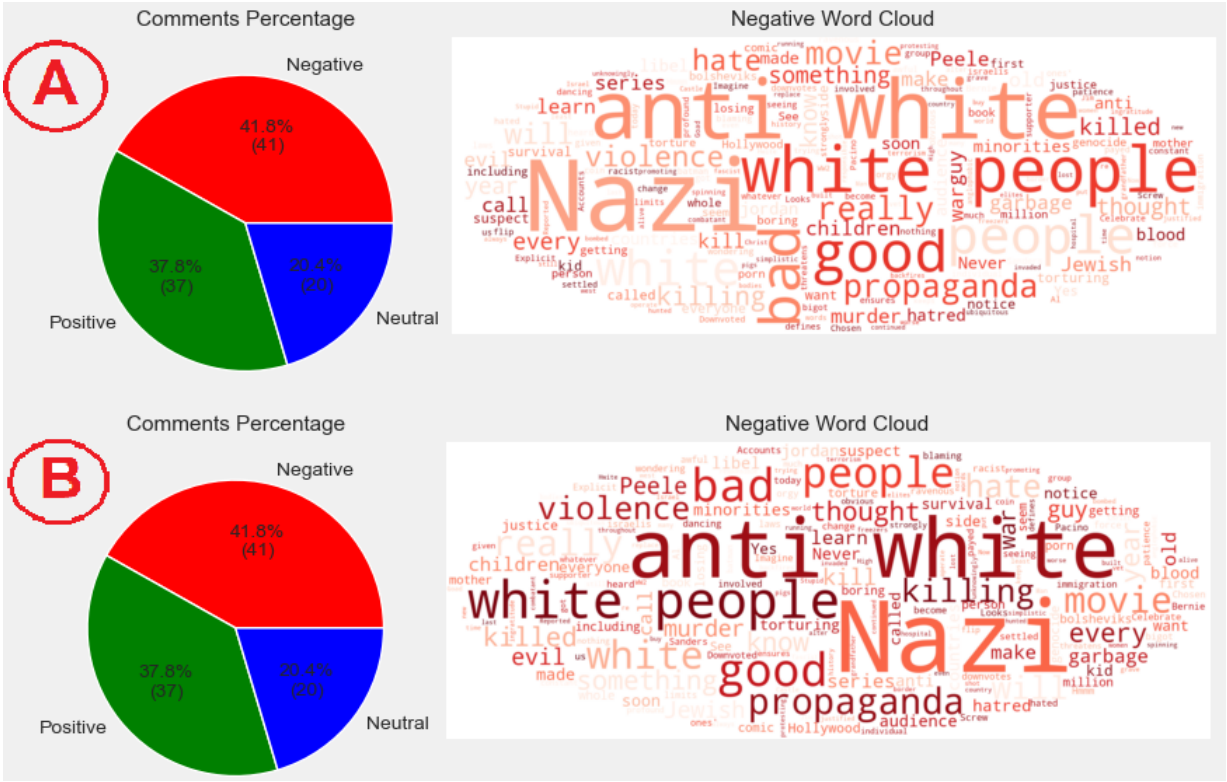
Figure 4.16 – Amazon Prime Video posts replies in April 2012 during the PS3 promotion.



Unlike Netflix, Amazon Prime did not present a clear outlier for the number of dislikes in a video. For YouTube, the two videos with more dislikes from 2020 were about the Hunters series. Hunter is a series that deals with the Nazism subject and which might be considered controversial. Analyzing the comments evolution is possible to notice that, for the two videos, the most comments were negatives. For the video released in January, 38,2% of its comments were negatives, and for the one released in February, 41,8%. On the word clouds, we can notice the use of harsh words like “hate”, “nazi”, “propaganda”, and “anti-white”. Details regarding comments’ behavior can be seen in Figure 4.17.

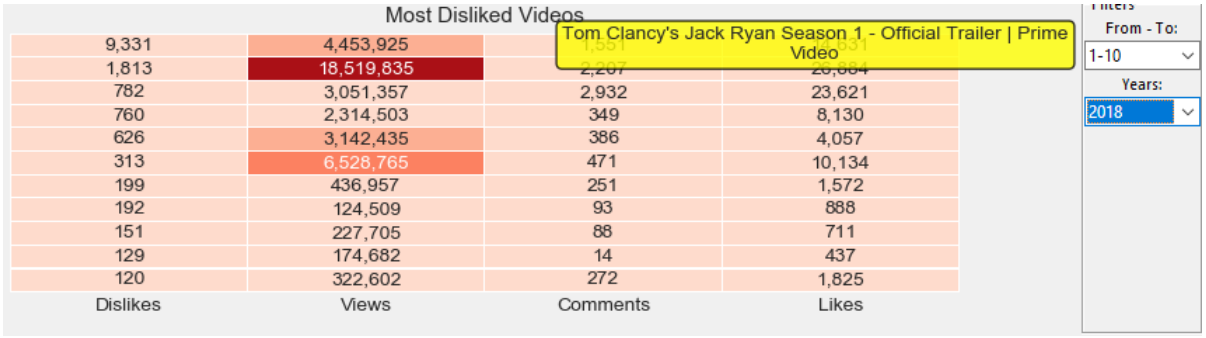
The other videos with a high dislike number are the first video released about Tom Clancy’s Jack Ryan from 2018 and the one about the NFL from 2017. The first Tom Clancy’s Jack Ryan video was removed from YouTube, and the other two regarding the series did not

Figure 4.17 – Amazon Prime Video Hunters' comments overall and negative word clouds. A represents the video from January and B represents the video from February.



generate a high number of dislikes. It suggests there was an issue with the first video, but this cannot be confirmed due to missing data. See Figure 4.18 for details.

Figure 4.18 – Amazon Prime Video Tom Clancy's Jack Ryan dislikes.



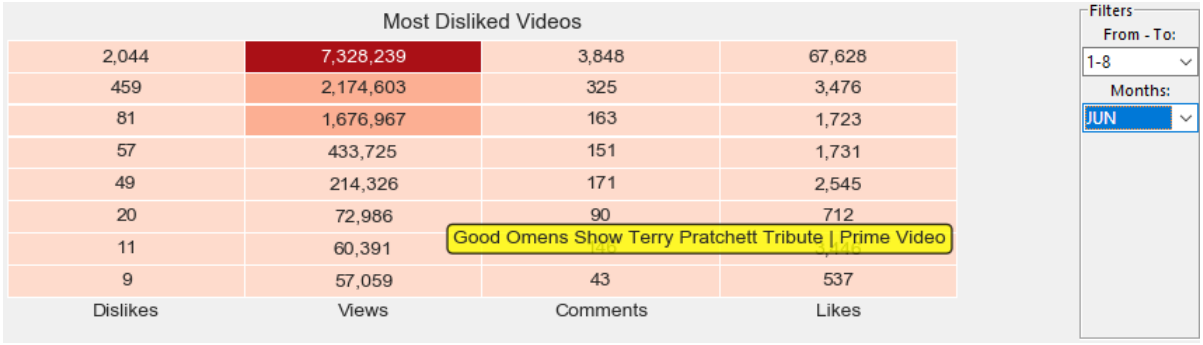
The NFL video presented a high number of dislikes, but the comments classification did not reflect that. The number of positive and negative comments were almost the same (57 for the positive and 52 for the negative). But if we check the most liked comments, the top five ones were negatives. Details can be seen in Figure 4.19.

In 2009, Amazon Prime Video released the "Good Omens" series. This series generated vast controversy in catolic groups that request to boycott the series. But these groups target the boycott to Netflix and not to Amazon Prime Video [57]. So, we decided to verify how the "Good Omens" videos were received, expecting to see lots of dislikes. But to our surprise, the videos had very little dislikes, as seen in Figure4.20. It might have

Figure 4.19 – Amazon Prime Video NFL's video comments evolution.

happened because the rage was direct to Netflix. Since Netflix did not post about the series on its social media, we could not confirm this.

Figure 4.20 – Amazon Prime Video Good Omens dislikes.



4.4 Nile Wilson

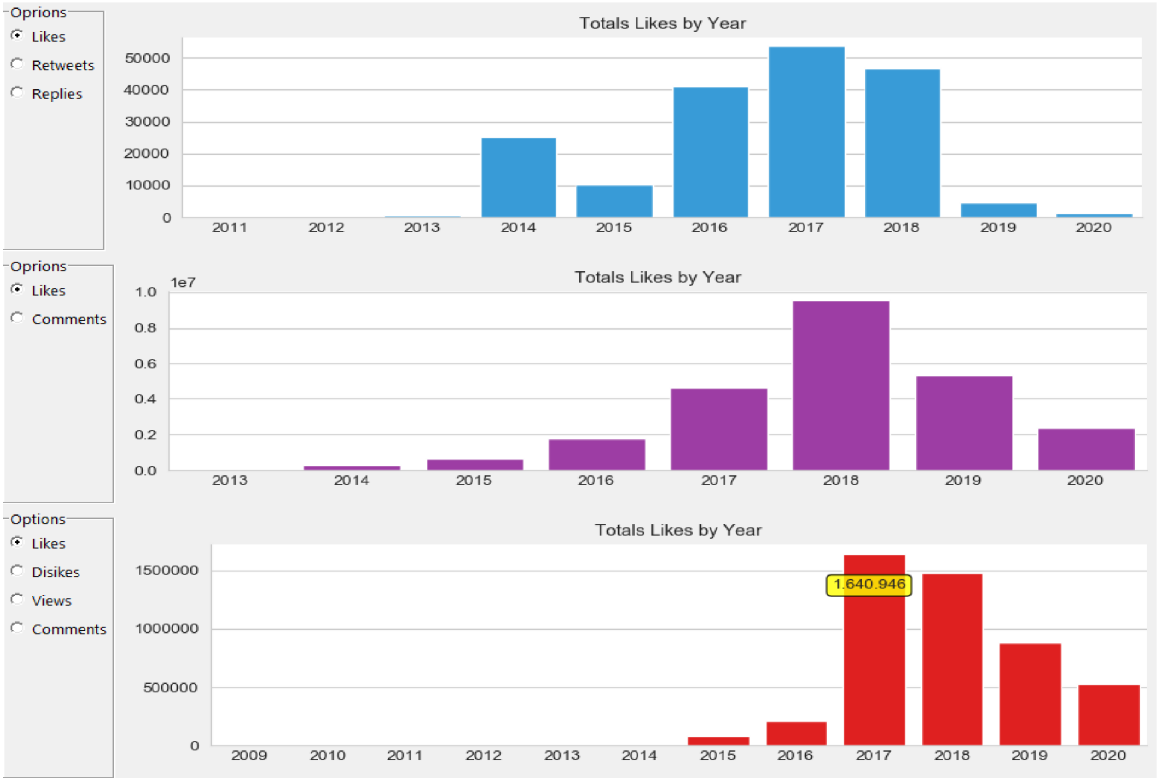
The third case we analyzed is about the profile of Nile Wilson on social networks. Nile Wilson [56] is a British artistic gymnast with a very consistent international career. As a junior competitor, he won the European and Youth Olympics. As a senior, he won medals at the 2014 Commonwealth Games, the 2015 World Championships, and the 2016 European Championships. In 2016, at the Olympic Games, he won a bronze medal in the individual

high-bar event. In 2017, after the Olympics, Nile suffered from injuries and decided to document his recovery on social media. That was, according to himself [15], when he gained social media Fame.

Like Amazon Prime Video and Netflix, Nile Wilson also experienced some growth spreads. For Twitter, significant growth in likes, retweets, and replies numbers happened from 2015 to 2016. Instagram behavior was different. For likes, the growth spread happened along two years, from 2016 to 2018. But, for comments, the growth was only observed from 2017 to 2018. It is interesting to notice that: the number of comments increased during the period he was recovering from his injury. This fact suggests public support in difficult times for the athlete.

For YouTube, the growth occurred from 2016 to 2017, when his number of views, likes, and comments increased more than 10 times. YouTube had the highest increase in interactions among the three social media. So we can agree with his statement that he gained social media fame in 2017. Information about the growths can be seen in Figure-4.21.

Figure 4.21 – Nile Wilson statistics major raises.



Unlikely Netflix and Amazon Prime Video, Nile Wilson experienced some significant declines in his interaction numbers totals. From 2018 to 2019, he has significantly decreased his numbers of posts on all his social medias. Consequently, the interaction totals also decreased.

We could observe a reduction of 88% for Twitter posts with a reduction of 90% in the likes numbers. For Instagram, the reduction was 48% for posts number and 42% for likes. And for YouTube, both the post number and the likes number were reduced by 40%.

Although the interaction totals have decreased, the interaction averages did not. They remained stable or even increased (Figure 4.22). This fact led us to believe that the drop in the totals happened due to a lower number of posts and not by the drop in his popularity.

Figure 4.22 – Nile Wilson statistics comparison between 2018 to 2019. A presents 2018 statistics and B presents 2019 statistics

Twitter Filters Year 2018	
Tweets Posts Total:	364
Tweets Likes Total:	46.799
Tweets Likes Mean:	129
Tweets Replies Total:	1.073
Tweets Replies Mean:	3
Retweets Total:	4.199
Retweets Mean:	12

A

Instagram Filters Year 2018	
Instagram Posts Totals:	252
Instagram Likes Total:	9.543.552
Instagram Likes Mean :	37.871
Instagram Comments Total:	121.722
Instagram Comments Mean:	483

YouTube Filters Year 2018	
YouTube Total Posted Videos	84
YouTube Likes Total:	1.484.284
YouTube Likes Mean:	17.670
YouTube Comments Total:	77.511
YouTube Comments Mean:	923
YouTube Views Total:	63.726.461
YouTube Views Total Mean:	758.648
YouTube Dislikes Total:	19.909
YouTube Dislikes Mean:	237

Twitter Filters Year 2019	
Tweets Posts Total:	41
Tweets Likes Total:	4.822
Tweets Likes Mean:	118
Tweets Replies Total:	153
Tweets Replies Mean:	4
Retweets Total:	448
Retweets Mean:	11

B

Instagram Filters Year 2019	
Instagram Posts Totals:	137
Instagram Likes Total:	5.376.622
Instagram Likes Mean :	39.245
Instagram Comments Total:	79.465
Instagram Comments Mean:	580

YouTube Filters Year 2019	
YouTube Total Posted Videos	51
YouTube Likes Total:	884.056
YouTube Likes Mean:	17.334
YouTube Comments Total:	28.366
YouTube Comments Mean:	556
YouTube Views Total:	37.515.662
YouTube Views Total Mean:	735.601
YouTube Dislikes Total:	28.539
YouTube Dislikes Mean:	560

From 2019 to 2020, the numbers of posts also decrease, and now, the average of interactions for 2020 also declined. It suggests that there is only a certain period that a person can remain out of social media before he/she starts to lose popularity. This information is provided in Figure 4.23

The year with the most disliked videos is 2017 when Nile Wilson released a series of videos called the ULTIMATE GYMNASTICS CHALLENGE. These videos were the most disliked videos of 2017, but they also got a high views number.

These videos regard a series of gymnastics challenges where gymnasts compete against each other or with athletes from other modalities. This tone of competition might be one of the reasons for a high dislike number. The other reason might be that, since more people watched the videos, the number of dislikes increased proportionally. If we check the proportion of views against the dislike number, we will note that it did not increase much. Figure 4.24 presents more information about dislikes.

The videos with high dislike numbers did not generate a high number of negative comments. It reinforces our finding that the dislikes number increased proportionally to the

Figure 4.23 – Nile Wilson statistics comparison between 2019 to 2020. A presents 2019 statistics and B presents 2020 statistics

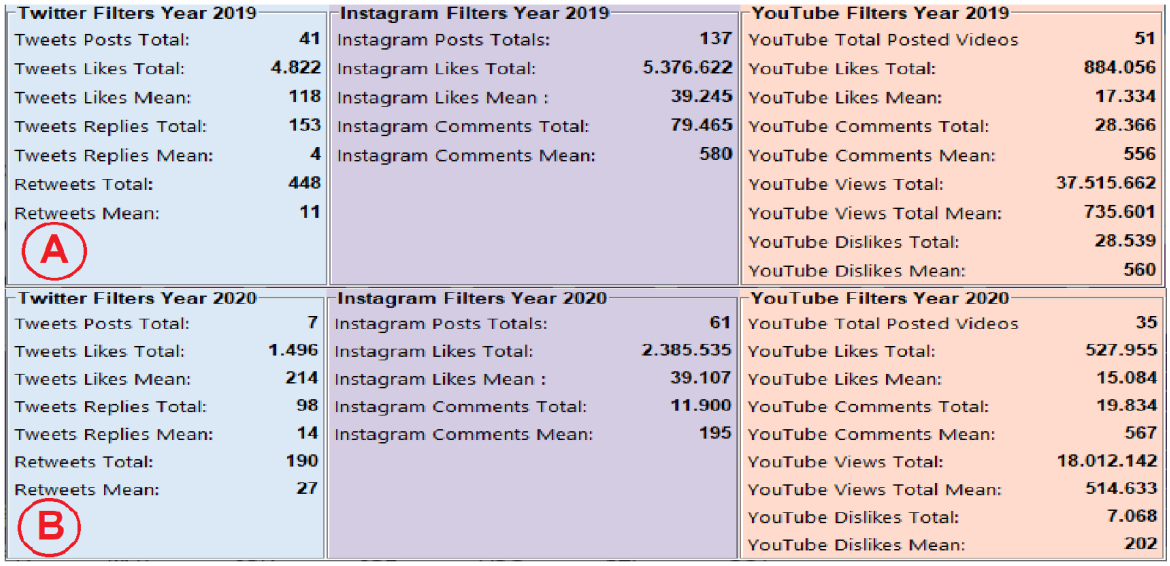
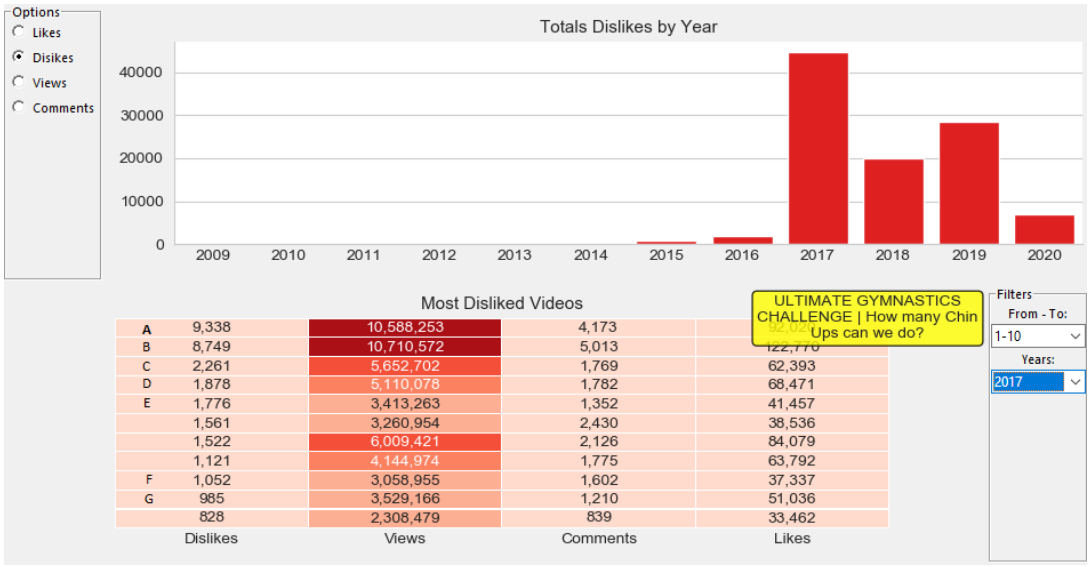


Figure 4.24 – Nile Wilson dislikes barchart and heatmap for 2017. Letters A-G indicate the posts related to the ULTIMATE GYMNASTICS CHALLENGE series.



views number. Figure 4.25 provides the overall comments charts for the two videos with more dislikes.

The most disliked video for Nile Wilson is the video titled “SETTING MY ’COACH’ A GYMNASTICS CHALLENGE for £10,000” from 2019. For this specific video, he asked for people to dislike it at the end of the video. We do not know his reasons for requesting it, but his viewers attended by making this the most disliked video. The commentary evolution chart supports (Figure 4.26) the fact that this video was only has a high dislike number due to his request.

Figure 4.25 – Nile Wilson dislikes overall for the most disliked videos of 2017. A indicates the January video, and B indicates the March.

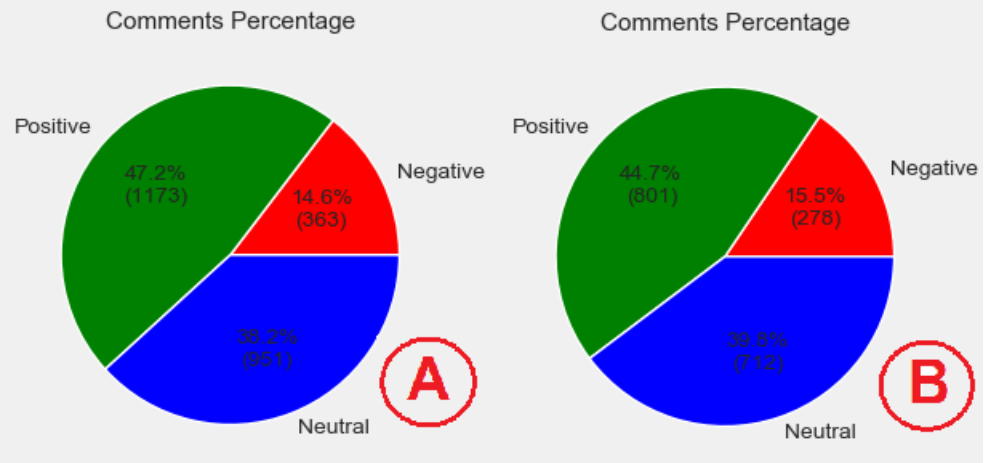
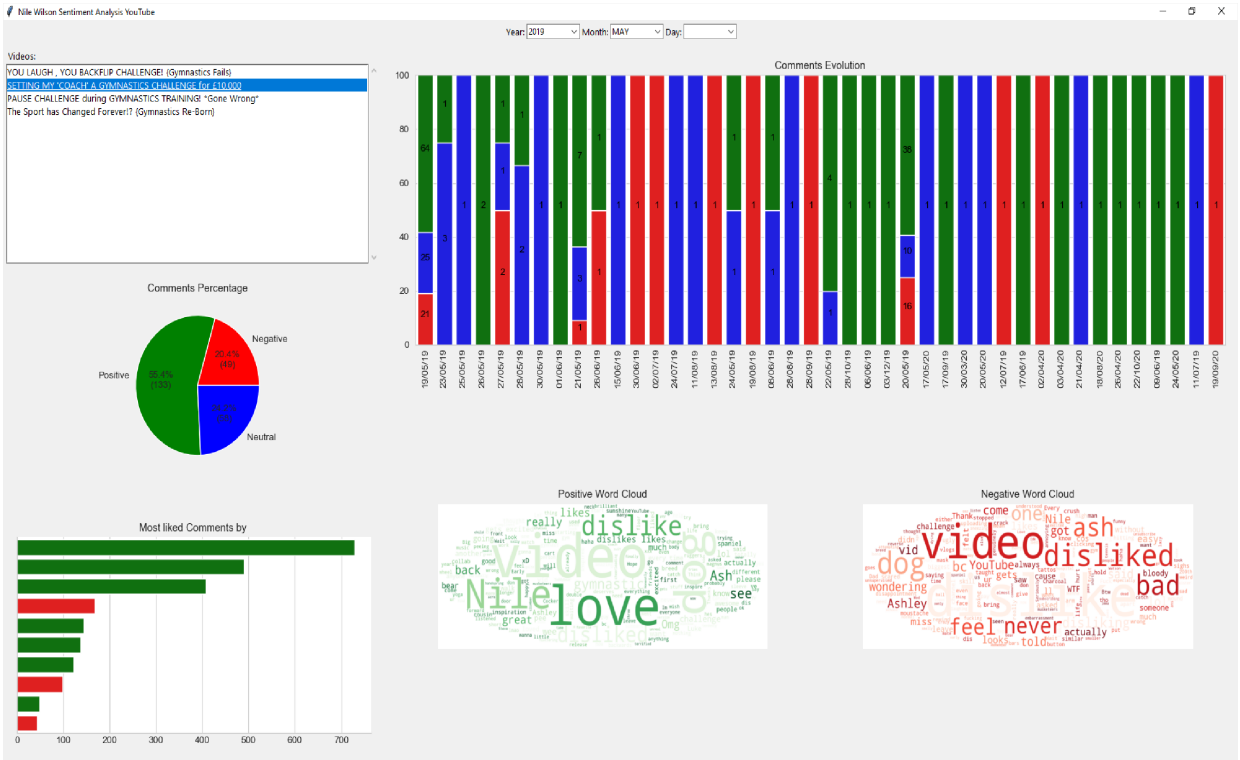
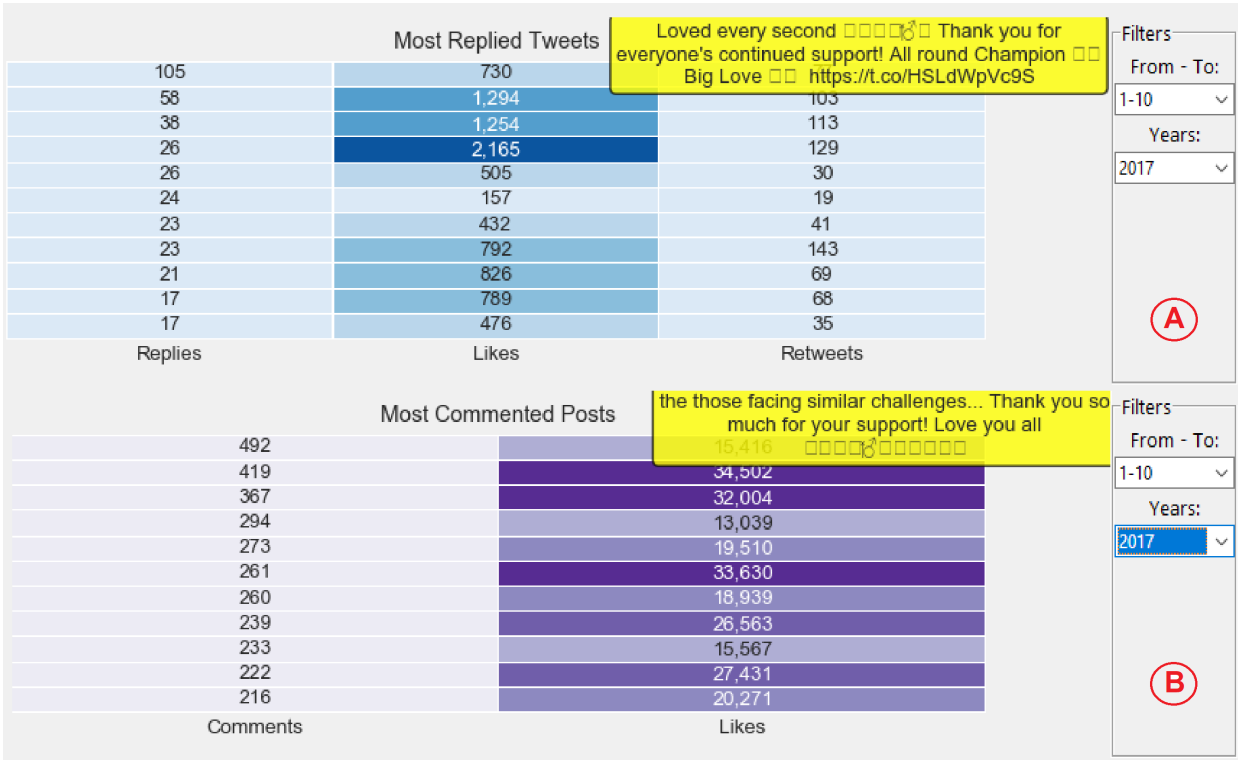


Figure 4.26 – Nile Wilson comments evolution for video “SETTING MY ’COACH’ A GYMNASTICS CHALLENGE for £10,000”.



Like Netflix and Amazon Prime Video, the Nile Wilson post with the highest number of likes and retweets is not the one with the highest number of replies. But, for him, we can observe a different phenomenon. For the other two brands, people use like for support and comments for controversies. For him, replies are used for supporting and the likes for posts where he performs funny things like opening a jar or jumping daily objects (cups, toilet papers, etc.) while performing gymnastic stunts. This phenomenon can be observed both for Twitter and Instagram, as illustrated in Figure 4.27.

Figure 4.27 – Nile Wilson most replied/commented posts. A presents the most replied posts for Twitter. B presents the most commented posts for Instagram.



4.5 Discussion

By analyzing the case studies, we can spot several similarities between Netflix and Amazon Prime data. The first is that the interactions are highly concentrated among individual high-performing posts rather than spread evenly across the data. This can be inferred from the scenarios in which a few select posts garner many likes and comments beyond the average number of interactions for each post. This phenomenon is replicated between both brands across different social platforms. Although this phenomenon is also observed for Nile Wilson, it occurs with less intensity. It suggests that social interactions behave differently when the posts are either made by a brand or by a person.

For future work, we foresee the possibility of extending the analysis of posts to discover common factors between them and predict which ones will maximize the attention and engagement of those posts. Also, we wanted to analyze other "persons' brands" to identify if the social interaction differs for brands in general.

Among all the analyzed social media, Instagram has consistently had the highest number of interactions. By observing Instagram data, we can easily spot a much higher interaction growth when compared with other social media. One possible explanation is the different dynamics in the way these platforms operate. As opposed to Twitter's real-time updated stream of tweets, posts on Instagram remain on users' feeds for a much longer

time, possibly enabling further interactions in scenarios where tweets would likely be missed by users. Moreover, when compared to YouTube, Instagram is still a more fast-paced environment. While on YouTube, each post is necessarily a video, which requires a higher time commitment for the user, given a video may take several minutes to be seen, on Instagram, each post comprises one or more pictures, which are visualized and read instantly by its users. We speculate that these differences have come to benefit Instagram regarding interaction numbers, but a more detailed analysis on exactly which factors enable this phenomenon is warranted for future works.

Lastly, another valid observation is that posts with high likes did not always translate into many comments and replies. This fact can also be observed in posts with a high number of replies but a relatively low number of likes. Our analysis verified that posts with a high number of likes generally meant public support regarding the posts and their contents. On the other hand, controversial topics generated many responses but a much lower number of likes. Therefore, we can infer that likes on social media are generally used to express support, but a higher number of replies can indicate controversy. In fact, the communities lingering in online spaces have noticed the dynamics associated with the balance of likes and replies, giving birth to the term “ratioed”. Minot et al. [39] describe the ratio value in Twitter as the balance of replies to likes and retweets, meaning posts mustering a high number of replies, yet low counts of likes will generally work as an indicator of controversy. Hence, the “ratio”. While the controversial tweets in our data verify the poor balance between replies and likes, we can also note that the ratio is not a phenomenon exclusive to Twitter, given the contentious YouTube videos in our data also sported an exceptionally high number of comments.

As the proposed visual analysis approach focuses on engagement, we argue that each of the evaluated metrics is able to show insights and pinpoint specific situations about how customers engage with and perceive brands in these online environments. For example, likes are strongly tied to notions of positive perception, such as brand loyalty. Algharabat [3] argues that the act of liking brands in social media is an indicator of brand love, in which by engaging with the brand, consumers then associate the brand as part of their online self-expression. On the other hand, the dynamics expressed by the fine balance between likes, shares, and replies can be an indicator of controversy and backlash.

Our case studies allowed us to infer other perceptions such as the content of a post is more important than the quantity of the posts; Instagram posts generate more interactions; it is important to avoid publishing controversial topics.

5. FINAL REMARKS

Due to the exponential growth and the quick feedback provided, social media has become an important information source for many areas. The thousands of data generated daily transformed social media into a reliable, fast, and relatively low-cost data source.

One area that can benefit from the use of social media data is brands. In the past, brands needed to rely on marketing surveys to obtain feedback for a product or service, but these surveys were costly and demanded time to be implemented. Nowadays, the world is very dynamic, and perceptions can change in a matter of seconds, so with the advancement of social media data, brands can get feedback quickly.

However, analyzing a brand through its social media raises challenges, especially regarding data gathering and data analysis. Social media data might become useless without the appropriate tools to exploit the data. It also needs to be easily updated by the user. To deal with these challenges, we need approaches that encompass the challenges as a whole.

Therefore, we developed an interactive visual analysis approach that provides a well-defined pipeline and simple but powerful visualization techniques that enable a direct analysis by any user without programming knowledge. The value of our approach is exemplified by three case studies, which provided valuable insights from the analyzed brands. To the best of our knowledge, this is the first work available at GitHub to analyze three social media together and provide comparisons between them. We also contribute with an easy-to-use script for YouTube data collection.

The interactive visualization techniques implemented by our approach might be simple. But they are powerful in providing a direct analysis by any users. It is exemplified in our case studies, which provided lots of insight into the three analyzed brands.

To improve our approach, we aim to develop a classification method that shows the mutual characteristics of the posts with higher interaction numbers and, therefore, predicts the ones that generate more engagement. Also, we aim to provide a real-time system analysis that can be configured, by the user, to issue alerts in specific scenarios (e.g., a high number of negative comments). Finally, we intend to interview brand managers to get feedback about other features we should consider for extending our approach.

REFERENCES

[1] Aaker, J. “Dimensions of brand personality”, *Journal of Marketing Research*, vol. 34–3, Aug 1997, pp. 347–356.

[2] Aggarwal, A.; Singh, A. “Geo-localized public perception visualization using glopp for social media”. In: Proceedings of the IEEE Annual Information Technology, Electronics and Mobile Communication Conference, 2017, pp. 439–445.

[3] Algharabat, R. “Linking social media marketing activities with brand love”, *Kybernetes*, vol. 46–10, Nov 2017, pp. 1801–1819.

[4] Amazon Prime Video. “Amazon”. Source: https://www.primevideo.com/, December 2020.

[5] Arvanitidis, A.; Serafi, A.; Vakali, A.; Tsoumakas, G. “Branty: A social media ranking tool for brands”. In: Proceedings of the Machine Learning and Knowledge Discovery in Databases, 2014, pp. 432–435.

[6] Bhargavi, K.; Babu, B.; Iyengar, S. “Predicting the brand popularity from the brand metadata”, *International Journal of Electrical and Computer Engineering*, vol. 8, Oct 2018, pp. 3523–3535.

[7] Bing, L. “Sentiment analysis and opinion mining”, *Synthesis Lectures on Human Language Technologies*, vol. 5–1, May 2012, pp. 1–167.

[8] Bojko, A. “Informative or misleading? heatmaps deconstructed”. In: Proceedings of the International conference on human-computer interaction, 2009, pp. 30–39.

[9] Campos Filho, A.; Freitas, F.; Gomes, A.; Vitorino, J. “Brandmap: An information visualization platform for brand association in blogosphere”. In: Proceedings of the International Conference on Information Visualisation, 2012, pp. 316–320.

[10] Constantinides, E. “Foundations of social media marketing”, *Procedia - Social and Behavioral Sciences*, vol. 148, Aug 2014, pp. 40–57.

[11] DaVint Lab. “Brandanalysis”. Source: https://github.com/DAVINTLAB/BrandAnalysis, December 2020.

[12] DaVint Lab. “Davint lab github”. Source: https://github.com/DAVINTLAB/InstagramUtils, December 2020.

[13] DaVint Lab. “Tweetutils”. Source: https://github.com/DAVINTLAB/TweetUtils, December 2020.

[14] DaVint Lab. “Youtubeutils”. Source: https://github.com/DAVINTLAB/YouTubeUtils, December 2020.

[15] Federation Internationale de Gymnastique. “Nile wilson”. Source: https://gymnastics.sport/site/athletes/bio_detail.php?id=31576, December 2020.

[16] Fernández-Gómez, E.; Martín-Quevedo, J. “La estrategia de engagement de netflix españa en twitter”, *El profesional de la información*, vol. 27–6, Dec 2018, pp. 1292–1302.

[17] Galesic, M.; Garcia-Retamero, R. “Graph literacy: A cross-cultural comparison”, *Medical Decision Making*, vol. 31–3, Jun 2011, pp. 444–457.

[18] Gao, Y.; Wang, F.; Luan, H.; Chua, T.-S. “Brand data gathering from live social media streams”. In: Proceedings of the International Conference on Multimedia Retrieval, 2014, pp. 169–176.

[19] Gao, Y.; Zhen, Y.; Li, H.; Chua, T.-S. “Filtering of brand-related microblogs using social-smooth multiview embedding”, *IEEE Transactions on Multimedia*, vol. 18–10, Oct 2016, pp. 2115–2126.

[20] Gray, J.; Bounegru, L.; Milan, S.; Ciuccarelli, P. “Ways of seeing data: Toward a critical literacy for data visualizations as research objects and research devices”. In: *Innovative methods in media and communication research*, Springer, 2016, pp. 227–251.

[21] Hecksher, A.; Ebecken, N. “Estudo comparativo de mineração de opiniões em rede varejista”, *Sistemas & Gestão*, vol. 11, May 2017, pp. 423–430.

[22] Hunter, J. “Matplotlib: A 2d graphics environment”, *Computing in Science & Engineering*, vol. 9–3, Jun 2007, pp. 90–95.

[23] Hutto, C.; Gilbert, E. “Vader: A parsimonious rule-based model for sentiment analysis of social media text”. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, 2014, pp. 216–225.

[24] Internet Society. “Policy brief: Network neutrality”. Source: https://bit.ly/30G6my6, December 2020.

[25] Jeon, H.; Ahn, H. “Identification of the factors that affect the user reaction to posts on facebook brand pages”. In: Proceedings of the International Conference on Computer and Computational Sciences, 2015, pp. 203–206.

[26] Kalampokis, E.; Karamanou, A.; Tambouris, E.; Tarabanis, K. “Applying brand equity theory to understand consumer opinion in social media”, *Journal of Universal Computer Science*, vol. 22–5, May 2016, pp. 709–734.

[27] Khan, M.; Khan, S. “Data and information visualization methods, and interactive mechanisms: A survey”, *International Journal of Computer Applications*, vol. 34–1, Nov 2011, pp. 1–14.

[28] Kirtiş, A.; Karahan, F. “To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession”, *Procedia - Social and Behavioral Sciences*, vol. 24, 2011, pp. 260–268.

[29] Kucher, K.; Martins, R.; Paradis, C.; Kerren, A. “Stancevis prime: visual analysis of sentiment and stance in social media texts”, *Journal of Visualization*, vol. 23, Aug 2020, pp. 1015–1034.

[30] Kumar, N.; Ande, G.; Kumar, J. S.; Singh, M. “Toward maximizing the visibility of content in social media brand pages: a temporal analysis”, *Social Network Analysis and Mining*, vol. 8–Nov, Feb 2018, pp. 1–14.

[31] Lakhiwal, A.; Kar, A. “Insights from twitter analytics: modeling social media personality dimensions and impact of breakthrough events”. In: Proceedings of the Conference on e-Business, e-Services and e-Society, 2016, pp. 533–544.

[32] Lakkaraju, H.; Ajmera, J. “Attention prediction on social media brand pages”. In: Proceedings of the ACM International Conference on Information and Knowledge Management, 2011, pp. 2157–2160.

[33] Liu, X.; Xu, A.; Gou, L.; Liu, H.; Akkiraju, R.; Shen, H.-W. “Socialbrands: Visual analysis of public perceptions of brands on social media”. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 2016, pp. 71–80.

[34] Luan, H.; Li, J.; Sun, M.; Chua, T.-S. “The design of a live social observatory system”. In: Proceedings of the International Conference on World Wide Web, 2014, pp. 1025—-1030.

[35] Matosas-López, L.; Romero-Ania, A. “The efficiency of social network services management in organizations. an in-depth analysis applying machine learning algorithms and multiple linear regressions”, *Applied Sciences*, vol. Oct–15, Jul 2020, pp. 1–16.

[36] Mazloom, M.; Rietveld, R.; Rudinac, S.; Worring, M.; van Dolen, W. “Multimodal popularity prediction of brand-related social media posts”. In: Proceedings of the ACM International Conference on Multimedia, 2016, pp. 197–201.

[37] McKinney, W. “Data structures for statistical computing in python”. In: Proceedings of the Python in Science Conference, 2010, pp. 56–61.

[38] Milolidakis, G.; Akoumianakis, D.; Kimble, C. “Digital traces for business intelligence”, *Journal of Enterprise Information Management*, vol. 27–1, Feb 2014, pp. 66–98.

[39] Minot, J.; Arnold, M.; Alshaabi, T.; Danforth, C.; Dodds, P. “Ratioing the president: An exploration of public engagement with obama and trump on twitter”, *PLoS ONE*, vol. 16–4, Apr 2021, pp. 1–22.

[40] MongoDB Inc. “The database for modern applications”. Source: https://www.mongodb.com/, December 2020.

[41] Mueller, A. “Peekaboo”. Source: https://peekaboo-vision.blogspot.com/2012/11/a-wordcloud-in-python.html, December 2020.

[42] Mueller, A. “Word cloud”. Source: http://amueller.github.io/word_cloud/, December 2020.

[43] Netflix. “Netflix”. Source: https://netflix.com/, December 2020.

[44] O’Connor, B.; Balasubramanyan, R.; Routledge, B.; Smith, N. “From tweets to polls: Linking text sentiment to public opinion time series”. In: Proceedings of the International AAAI conference on weblogs and social media, 2010, pp. 122–129.

[45] Overgoor, G.; Mazloom, M.; Worring, M.; Rietveld, R.; van Dolen, W. “A spatio-temporal category representation for brand popularity prediction”. In: Proceedings of the ACM International Conference on Multimedia Retrieval, 2017, pp. 233–241.

[46] Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. “Systematic mapping studies in software engineering”. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, 2008, pp. 1–10.

[47] Plotly. “Plotly”. Source: https://plotly.com/, December 2020.

[48] Project Jupyter. “Jupyter notebook”. Source: https://jupyter.org/, December 2020.

[49] Python Software Foundation. “Pypi”. Source: https://pypi.org/, December 2020.

[50] Python Software Foundation. “Python”. Source: https://www.python.org/, December 2020.

[51] Qi, S.; Wang, F.; Wang, X.; Wei, J.; Zhao, H. “Live multimedia brand-related data identification in microblog”, *Neurocomputing*, vol. 158, Jun 2015, pp. 225–233.

[52] Singh, S.; Chauhan, A.; Dhir, S. “Analyzing the startup ecosystem of india: a twitter analytics perspective”, *Journal of Advances in Management Research*, vol. 17–2, Nov 2019, pp. 262–281.

[53] Teixeira, C.; Kurtz, G.; Leuck, L.; Sanvido, P.; Scherer, J.; Tietzmann, R.; Manssour, I.; Silveira, M. “Polls, plans and tweets: an analysis of the candidates’ discourses during the 2018 brazilian presidential election”. In: Proceedings of the Annual International Conference on Digital Government Research, 2019, pp. 439–444.

[54] The Matplotlib Development team. “Matplotlib”. Source: https://matplotlib.org/, December 2020.

[55] The Pandas Development Team. “Pandas”. Source: https://pandas.pydata.org/, December 2020.

[56] The Verge. “British gymnastic”. Source: https://www.british-gymnastics.org/gymnast-profiles/245901/nile-wilson, December 2020.

[57] The Verge. “Good omens protesters demand show be removed from completely wrong company”. Source: https://www.theverge.com/2019/6/20/18693159/good-omens-petition-amazon-netflix-neil-gaiman/, December 2020.

[58] Twitter. “Consuming streaming data”. Source: https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data, December 2020.

[59] Ward, M. O.; Grinstein, G.; Keim, D. “Interactive data visualization: foundations, techniques, and applications”. CRC Press, 2010, 486p.

[60] Waskom, M. “Seaborn: Statistical data visualization”. Source: https://seaborn.pydata.org/, December 2020.

[61] Waskom, M. “Seaborn: statistical data visualization”, *Journal of Open Source Software*, vol. 6–60, Apr 2021, pp. 1–4.

[62] Wohlin, C. “Guidelines for snowballing in systematic literature studies and a replication in software engineering”. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering, 2014, pp. 1–10.

[63] Wong, P.; Thomas, J. “Visual analytics”, *IEEE Comput. Graph. Appl.*, vol. 24, Sep 2004, pp. 20–21.

[64] Xu, J.; Tao, Y.; Lin, H. “Semantic word cloud generation based on word embeddings”. In: Proceedings of the IEEE Pacific Visualization Symposium, 2016, pp. 239–243.

[65] Zhao, S.; Yao, H.; Zhao, S.; Jiang, X.; Jiang, X. “Multi-modal microblog classification via multi-task learning”, *Multimedia Tools and Applications*, vol. 75, Nov 2016, pp. 8921–8938.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br