UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

ANALYSIS AND CHARACTERIZATION OF SOCIAL NETWORKS IN THE FORMULA 1 DOMAIN

JESÚS MIGUEL GARCÍA SÁNCHEZ

JUNIO 2023

TRABAJO DE FIN DE GRADO

Título:	Análisis y caracterización de redes sociales en el ámbito de la Fórmula 1
Título (inglés):	Analysis and characterization of social networks in the Formula 1 domain
Autor:	Jesús Miguel García Sánchez
Tutor:	Carlos A. Iglesias Fernández
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

ANALYSIS AND CHARACTERIZATION OF SOCIAL NETWORKS IN THE FORMULA 1 DOMAIN

Jesús Miguel García Sánchez

Junio 2023

Resumen

La cantidad de información que generan las sociedades modernas está en continuo crecimiento. Los datos que podemos extraer acerca de una noticia, un usuario o una tendencia son prácticamente infinitas. Hoy en día muchas son las empresas que conscientes de este suceso, aprovechan el poder de las herramientas de análisis obteniendo datos claves acerca del comportamiento de las personas.

Mi proyecto personal, la plataforma de noticias de Fórmula 1 llamada "JaramaFan", cuenta con un gran seguimiento en redes sociales, en especial en la red de Twitter, por tanto, utilizaremos los datos de esta cuenta para llevar a cabo las investigaciones de este proyecto. Nos enfocaremos en el análisis de redes mediante la teoría de grafos para crear un mapa de la red, el cual será estudiado posteriormente. Obtendremos las principales comunidades de la red global, en este caso, la red de la comunidad de Fórmula 1 y realizaremos análisis de las métricas de popularidad de las cuentas involucradas. Además, emplearemos el Procesamiento del Lenguaje Natural (PLN) para analizar las emociones y sentimientos de los tweets de estas cuentas.

El objetivo final del trabajo es correlacionar los datos obtenidos, tanto en el análisis de métricas de popularidad como en el análisis de métricas de sentimiento. Esto permitirá obtener conclusiones para comprender mejor las redes sociales. Al final del análisis, sabremos cómo las emociones influyen en la popularidad de los tweets, cómo se forman las comunidades en Twitter y cómo varían las métricas de popularidad entre las diferentes cuentas.

Palabras clave: Teoría de grafos, Twitter, Análisis red social, Análisis de Sentimientos, Fórmula 1.

Abstract

The amount of information generated by modern societies is constantly growing. The data that can be extracted about news, users, or trends is practically infinite. Nowadays, many companies are aware of this phenomenon and leverage the power of analytical tools to obtain key data about people's behavior.

In my personal project, the Formula 1 news platform called "JaramaFan" has a large following on social media, especially on the Twitter platform. Therefore, we will use the data from this account to carry out the research for this project. We will focus on network analysis using graph theory to create a network map, which will be studied subsequently. We will identify the main communities within the global network, specifically the Formula 1 community, and analyze the popularity metrics for the involved accounts. Additionally, we will employ Natural Language Processing (NLP) to analyze the emotions and sentiments expressed in the tweets from these accounts.

The ultimate goal of this work is to correlate the data obtained, both in terms of popularity metrics and sentiment analysis. This will provide insight for a better understanding of social networks. At the end of the analysis, we will understand how emotions influence tweet popularity, how communities are formed on Twitter, and how popularity metrics vary among different accounts.

Keywords: Graph theory, Twitter, Social network analysis, Sentiment analysis, Formula 1.

Agradecimientos

Me gustaría expresar mi gratitud a mi tutor, Carlos Ángel Iglesias, por su orientación y apoyo durante la realización de este Trabajo de Fin de Grado. Su experiencia y valiosos conocimientos han contribuido en gran medida al éxito de este proyecto. Estoy sinceramente agradecido por su tiempo, dedicación y constante apoyo.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

As of early 2023, the world population has surpassed 8.01 billion, with over 57% living in urban areas. Around 68% of the population uses mobile phones, with 168 million new users added in the past 12 months. These numbers assume 4.76 billion users of social networks worldwide, representing just under 60% of the total population, with a net addition of 137 million new users in the past year, resulting in an annual growth rate of 3% [3]. Currently, analyzing and detecting communities on social networks is a crucial job for the future, as it allows one to understand the relationships between them and how they relate. This knowledge can have important implications in various fields, from medicine and economics to marketing and health. For example, sentiment analysis of tweets can be used to understand how communities respond to a particular topic or to determine why certain accounts have a greater impact than others.

In the case of businesses, social networks have become a powerful tool of communication, allowing them to reach a wider audience and establish closer relationships with their customers. Social media also offers a unique opportunity for businesses to promote their products and services, post news and offers, and stay in touch with their customers. Ad-

ditionally, the advertising industry has found an effective medium on social networks to reach their target audience. A new figure, the influencer, has emerged who can influence their followers and become a valuable brand endorser. The project aims to study a social network related to the subject: Formula 1, using different methods of Social Network Analysis (SNA) and Sentiment Analysis [4].

In addition, a study of the different graph theory analysis techniques to detect communities will be carried out, ultimately establishing a correlation between both studies to achieve a broader analysis from a different approach. Through SNA techniques we will analyze relationships between elements to determine the importance of the actors, how they have grouped themselves into subcommunities, the flow of information, what is key in transmitting information, etc. In the development of the following work, we intend to use the different methods of both, SNA and graph theory to carry out an in-depth study of social networks using the Twitter platform (an ideal platform for analyzing the interaction of different users based on their "tweets"). The information collected from this social network will be stored in different datasets to study them. At the same time, we will obtain the network map with the different relationships between users.

1.2 Project Goals

The objectives to be achieved in the project will be the following:

- Analysis of a network based on graph theory.
- Analyze communities through the community detection method.
- Detection of the most important nodes in the network.
- Pre-process the collected information by cleaning it for later analysis
- Analysis of the different popularity metrics of an account.
- Analysis of sentiments metrics of an account.
- Study of the correlation of popularity metrics and sentiment analysis.
- Draw conclusions from the data obtained in correlation analysis.
- Apply the same analysis to a specific case study.

1.3 Structure of this document

In this section, we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1 Puts the work in context for the reader and presents the objectives of it. It gives a brief overview of how information will be handled in the project.

Chapter 2 Provides information on the different technologies that have helped to carry out this project. There is information on what they are for, where they come from, and, more briefly, how they will be used.

Chapter 3 Deals with how the data was obtained for this work and how they are divided. It also shows the distribution of these data in space and time.

Chapter 4 Details all the analyzes that have been made to the different datasets, the results obtained and the comparisons and conclusions drawn from them.

Chapter 5 Gives a more concrete vision of the F1 community analyzing data in separate case of study.

Chapter 6 Discusses the conclusions drawn from this project, the goals achieved, the problems faced and their respective solutions, and suggestions for future work

CHAPTER 2

Enabling Technologies

This chapter is dedicated to illustrating the different technologies that have been utilized to collect and analyze information. It includes several libraries, programs, and visual interfaces.

2.1 Tweepy

Tweepy [11] is a popular Python library that is used to access Twitter Application Programming Interface (API) [12]. It provides an easy way to interact with the Twitter platform, allowing developers to search for tweets, retrieve user information, and post tweets. The choice to use Tweepy [11] was mainly due to the ease of performing the authentication process to obtain information on Twitter API [12] and its different methods to manage a large amount of data in a simple way. To access the Twitter API, we must obtain permissions from "Twitter platform for developers" [1], using the corresponding keys and tokens to authenticate requests. For the collection of tweets, we have used Representational State Transfer (REST) API, and the responses are obtained in Java Script Object Notation (JSON) format.

[1] https://developer.twitter.com/en

2.2 Technologies for sentiment analysis

This section introduces the different technologies to analyze sentiments through Natural Language Processing (NLP) [10]. The technologies used in this project employ a combination of linguistics, statistics, and machine learning algorithms to analyze and extract meaning from human language. In this section, we will explain the different technologies for language recognition that we have used to perform sentiment analysis of collected tweets.

2.2.1 NLTK

Natural Language Toolkit (NLTK) [1] is a set of libraries and programs for symbolic and statistical NLP for the Python programming language. The NLTK library [1] is very useful for sentiment detection because it offers several text classification models that can be used to train and predict the polarity of texts. It uses a Naive Bayes classification model using supervised machine learning to obtain a set of sentiment features of the set analyzed.

2.2.2 Vader

Valence Aware Dictionary for Sentiment Reasoning (VADER) [2] is a sentiment analysis tool that is sensitive to both the polarity (positive/negative) and the intensity (strength) of the emotion. It is available in the NLTK package and can be applied directly to unlabeled text data. VADER Sentimental analysis [2] is based on a dictionary that maps lexical characteristics to emotion intensities known as sentiment scores. The final sentiment score is obtained by adding the intensity of each word in the text.

2.3 Analysis and Visualization

In this section, we will explain the different technologies to analyze and visualize the data obtained previously.

2.3.1 Pandas

Pandas [9] is an open-source Python library that provides high-performance data analysis tools to manage different types of data structures. It offers extensive functionalities for grouping, merging, and querying data while supporting time series analysis and visualiza-

tion capabilities. With its user-friendly interface and versatile capabilities, Pandas [9] is a powerful tool for managing and analyzing data structures easily and efficiently.

2.3.2 Gephi

Gephi [8] is a visualization application developed in Java language. It is used primarily to visualize, manipulate, and explore networks and graphs from raw node and edge graph data. It is a free and open-source application that can handle large networks by partitioning and sorting the graphs. Gephi [8] can connect to Twitter API, allowing us to filter the information from the tool itself. We can also perform automatic community detection using its partition-based community detection method. Its ease of use, combined with its strong visual component, makes it a suitable tool for correlating the obtained tweets with their network map.

CHAPTER 3

Data Acquisition

3.1 Introduction

This chapter explains how we obtained the different data for this study. As this project is about comparing data between graphs and tweets, it is essential to follow a concrete procedure. Therefore, the data acquisition process shown in Fig. 3.1 has been applied in this project.

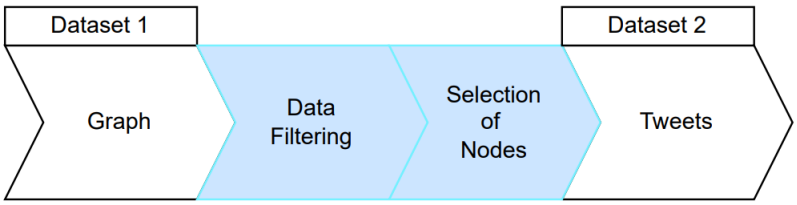

Figure 3.1: Data Acquisition Procedure

We start by obtaining the first Dataset: "Formula 1 Network graph". Here, we have all the information about the accounts and the relationship between them that belong to the F1 community. Next, this data set will be modeled to obtain the different subcommunities in which the Formula 1 community is divided by a specific community detection method.

Finally, we get another dataset from the most influential accounts of each subcommunity included in the previous dataset. In the second dataset, we will focus on collecting tweets instead of accounts to analyze the metrics and sentiments of these tweets.

3.2 Dataset 1. F1 community graph

The first data set was captured on 24 October 2022. As mentioned in Chapter 2, we have used Gephi [8] to obtain the graph through Twitter API that allows us to capture information in real time. This dataset consists of all accounts that belong to the F1 community given in CSV format and represented visually in Gephi [8]. The dataset has been carried out following these features:

- The graph obtained is an unweighted directed graph.
- Nodes correspond to Twitter accounts.
- The accounts collected can be from any region of the world.
- Edges correspond to interactions between accounts: Like, Comment, or RT.
- We made a query with the words: JaramaFan, F1, Fernando Alonso, Sergio Pérez, and Carlos Sainz, as those are keywords from the JaramaFan account, which is the subject of the study.

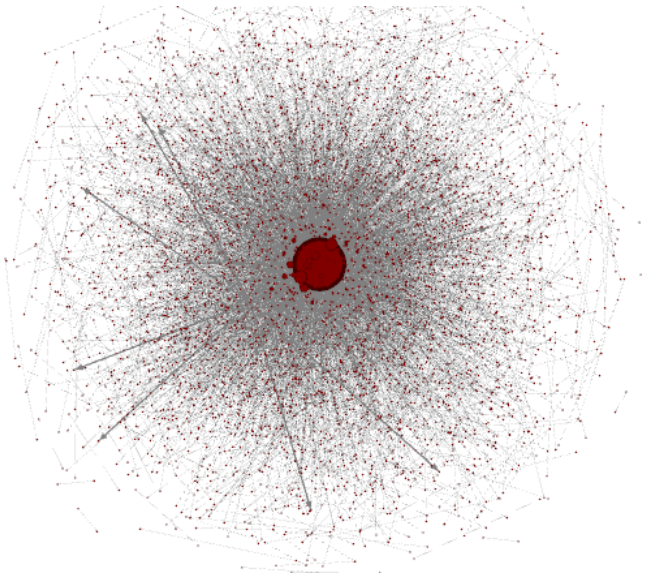

Figure 3.2: Dataset 1: Formula 1 graph

Parameters	Results
Medium Degree	1.927
Diameter	8
Connected nodes	429
Average Path	2,35

Table 3.1: Graph Parameters

We have obtained a total of 4.488 nodes and 8.648 edges, with a medium degree of 2,61. The data obtained in Table 3.1 reflect that most of the nodes have a low interaction. We can see this in the medium degree of the graph or in the average path, where we can see a low number of interactions caused by many accounts that barely interact with others. Filtering the data will be necessary to get a clearer vision of the most critical interactions on the net. On the other hand, as we can see in figure 3.3, we have obtained the initial dataframe in CSV format with the 4,888 accounts that represent the graph nodes. The degree columns represent each account's interaction in the net. We have also collected the provenance of the accounts that interact in the Formula 1 community, described in figure 3.4. The countries with the most significant presence in the community are the United Kingdom, the United States, and Mexico.

	Id	Label	indegree	outdegree	degree	followers_count	location
0	@f1	@f1	1591	11	1602	8705622.0	Worldwide
1	@alo_oficial	@alo_oficial	348	1	349	3251066.0	instagram: fernandoalo_oficial
2	@redbullracing	@redbullracing	313	1	314	4206954.0	Worldwide
3	@schecoperez	@schecoperez	190	1	191	3445023.0	Fuschl am See, Austria
4	@alpinef1team	@alpinef1team	187	1	188	2058365.0	Enstone, England
...	...	...	...	...	...	...	...
4483	@f1keirinberu	@f1keirinberu	0	0	0	12.0	NaN
4484	@picagrossagyn	@picagrossagyn	0	0	0	42.0	Indiana|Illinois
4485	@ohayou_mayonaka	@ohayou_mayonaka	0	0	0	1195.0	London, England
4486	@dime_hacks	@dime_hacks	0	0	0	10526.0	NaN
4487	@thesportynews	@thesportynews	0	0	0	170.0	London, England

Figure 3.3: Dataset 1: List of accounts and relationships

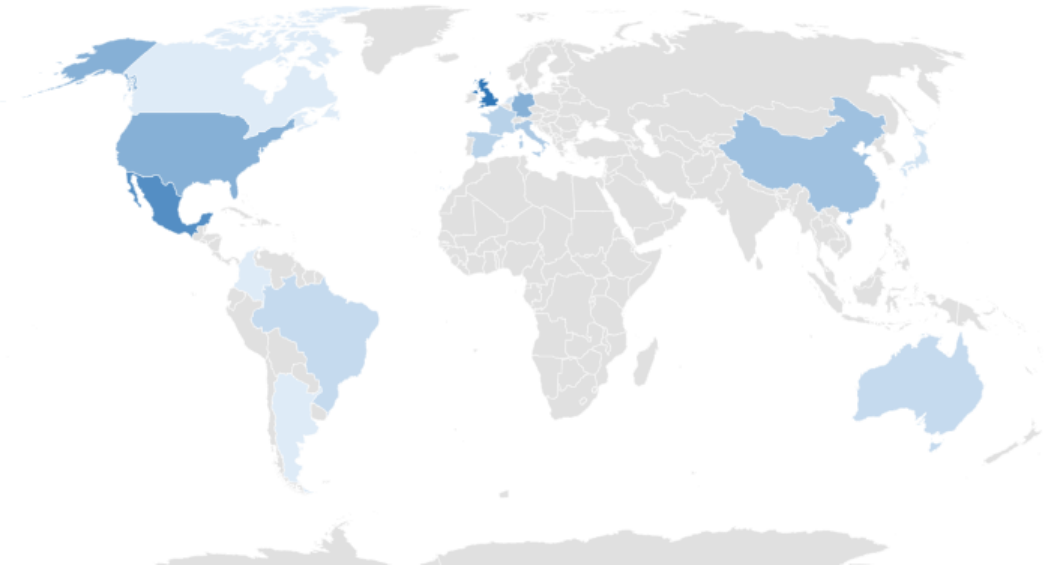

Figure 3.4: Map representing the provenance of the accounts collected

3.2.1 Data Filtering

We have filtered Dataset 1 of unwanted information to obtain accurate information from our graph. The following steps have been followed to carry out the filtering process:

- We have only considered nodes with between 10 and 1602 degrees of input.
- We have applied the Louvain method [6] as a community detection algorithm based on modularity.
- We colored the nodes belonging to the same community with the same color.
- The nodes with the highest number of entries have been painted larger, establishing a 155:25 ratio in them.
- We have selected Ying Fu proportional as a graph topology for this study.

Using the Lovain method [6], we have identified groups of nodes more connected to each other than other nodes in the graph by measuring the division's strength into distinct communities. In Table 3.2, we can see how the main parameters of the graph have been modified after the filtering process. The degree of node has increased due to the elimination of despicable information. However, modularity has decreased, meaning that the nodes have less divisive strength, making the graph more connected.

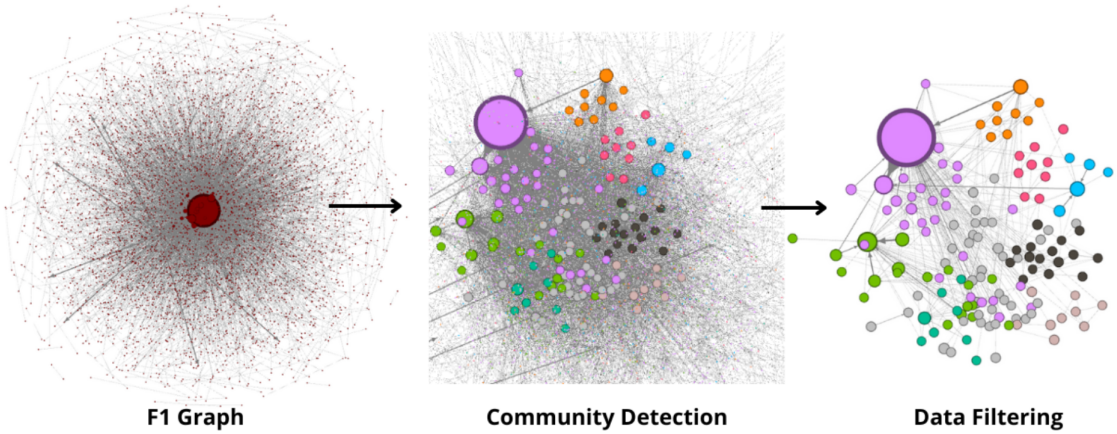

Figure 3.5: Community detection and filtering process

Parameters	Before Filtering	After Filtering
Nodes	4.488	187
Edges	8.648	488
Communities	501	8
Medium Degree	1,93	2,61
Modularity	0,749	0.452
Average Path	2,35	2,182

Table 3.2: Comparison of the graph parameters in the datasets after filtering

We can also see that the number of nodes and links has been considerably reduced to a total of 187 nodes and 488 edges. More than 8.000 interactions have been eliminated because they come from accounts with very little impact in the network. In addition, the number of communities has also been reduced considerably, from the 501 communities we had previously to the 8 we now have. After nodes with little interaction were removed, the average path of the graph decreased, as the remaining nodes had fewer distances in their interactions. This consequence is also reflected in the medium degree, which has increased due to removing nodes with very little interaction. We can see the community detection and elimination process in figure 3.5.

3.2.2 Selection of Nodes

Now, we will talk about how we have conducted the selection process for the nodes of each community. We have used the Louvain community detection method [6] because it is highly efficient, scalable, and designed to maximize modularity. It is very flexible and can be applied to various networks and situations. Furthermore, Gephi allows one to run Louvain's method directly on the generated graph, making community detection very simple and straightforward. Other methods available for Gephi, such as the Girvan-Newman method, have been discarded as they are computationally much more intensive. In the case of Girvan-Newman, it is able to identify communities at different levels of detail, but it is not designed to maximize modularity.

As we can see in Figure 3.6, these are the 8 communities detected in the graph separated by color. With respect to the data obtained, we can observe that communities are organized by language and topic, with emerging communities that arise and disappear once the news is no longer newsworthy. We will choose three accounts in each community with the highest degree computed as the sum of in-degree and out-degree. In community 8, we will pick four accounts instead of three, as the JaramaFan account is the subject of the study. Accounts with less than 1.000 tweets or any relation with the F1 community have been discarded. This will make a total of 25 accounts selected as the most influential in the F1 community, in which we will study their tweets in Chapter 4.

As we can see in Table 3.3, these are the selected accounts in each community as they represent the nodes with the highest degree. We can see huge differences between communities. While in community 1, the selected nodes have a huge node degree in each account, community 3 or 6 has very low interaction. We can also observe that the degree of each node largely depends on the in-degree, while the out-degree yields very low results. This is because accounts with many followers can receive interactions without interacting

Community	Selected Account 1	In-degree	Out-degree	Degree
1	@F1	1591	11	1602
	@RedBull Racing	313	1	314
	@Verstappen	65	0	65
2	@AlpineF1Team	187	1	188
	@HaasF1team	22	2	24
	@MarkBlundell	2	9	11
3	@MotorSport	16	0	16
	@Chrismdelandf1	14	0	14
	@wbuxton	13	0	13
4	@schecoperez	190	1	191
	@skysport	80	1	81
	@checofanpage	59	3	62
5	@hamilton	32	1	33
	@russell	16	0	16
	@mercedesamgf1	31	4	35
6	@mclarenf1	17	0	17
	@autosport	11	0	11
	@danielricciardo	10	0	10
7	@FIA	124	0	124
	@AstonMartinF1	49	1	50
	@TelemetricoF1	12	5	17
8	@FernandoAlonso	348	1	349
	@soymotor	119	0	119
	@crosaleny	106	2	108

Table 3.3: Accounts selected in each community

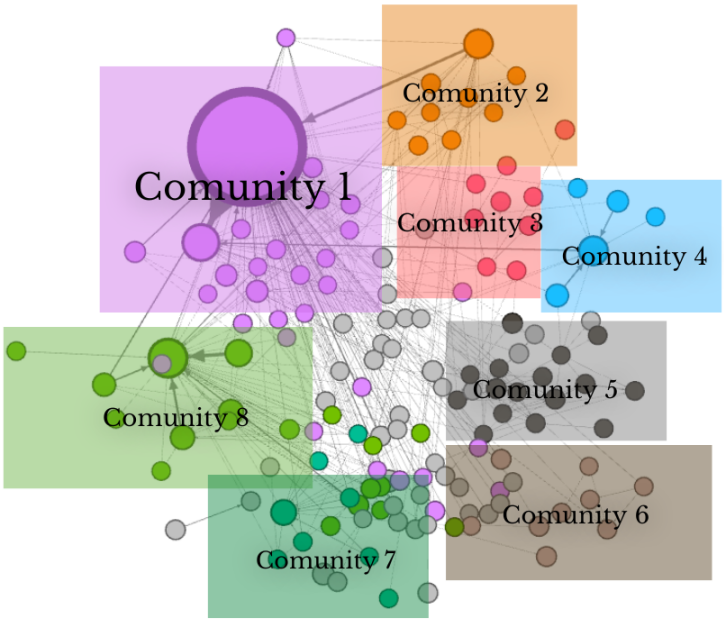

Figure 3.6: Communities in the F1 graph

with other nodes. We will collect the last 3.250 tweets from each account to conduct an analysis study in Sect. 4. On the other hand, it is surprising to see the closeness of some communities that initially seem very different or the remoteness of others. For example, the Latin community interacts more with communities related to Formula 1 news, while the Spanish community interacts more with each other, with Formula 1 or regulatory bodies. Communities 3, 6, and 7 are communities that arise from specific news about the United States Grand Prix. Communities 3 and 6 were formed due to the announcement of drivers' retirement, while community 7 was formed in relation to the incidents of the US race. The remaining communities were established for a longer period of time and differed from each other based on characteristics such as language or different hierarchies.

Community	Community Name	Account Name
1	Main	@F1,@RedBullracingF1 and @MaxVerstappen
2	F1 Teams	@AlpineF1Team, @HaasF1Team and @Markblundellf1
3	Journals	@MotorSport,@Chrismdelandf1 and @Wbuxton
4	Latin	@ChecoPerez,@ChecoPerezFan and @skysportsf1
5	English	MercedesAMGF1@LewisHamilton and @GeorgeRussell
6	Ricciardos's news	@danielricciardo, @MclarenF1Team and @AutoSport
7	Regulatory bodies	@FIA, @Telemetrico and @AstonMartin
8	Spanish	@fernandoalonso, @SoyMotor and @crosaleny

Table 3.4: Description of communities

3.3 Dataset 2. Most popular F1 accounts

In this dataset, we have collected tweets from the most important accounts in the F1 graph. This dataset was obtained from 10 April 2023 to 5 May 2023. F1 community is divided into eight subcommunities, in which we will study their most important nodes. This has resulted in a total of 25 accounts that will be analyzed following these features.

- The information we collect from each tweet is: "User"; "Tweet"; "Followers"; "Location"; "Retweet (RT)"; "Favourites".
- The total number of tweets is limited to each query.
- A maximum of 3.250 tweets will be collected in the 25 accounts.

A total of 77,261 tweets have been collected. The number of tweets collected from each account of the eight subcommunities is presented in Figure 3.7. We have collected the last 3,250 tweets from each account. However, some accounts have not been able to accumulate such a large number of tweets since their creation. Despite it, all accounts have posted more than 1,000 tweets, so we can perform a precise analysis.

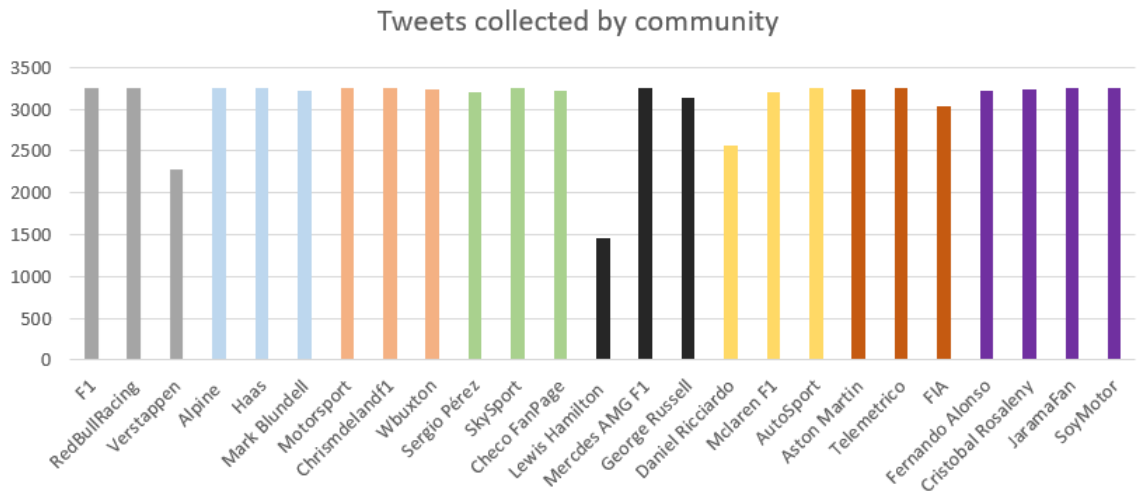

Figure 3.7: Tweets collected in Dataset 2

CHAPTER 4

Data analysis

4.1 Introduction

In this section, we will explain how we have analyzed the information in the selected accounts from Dataset 2. We will talk about the analysis of popularity metrics and sentiments and, finally, we will correlate both analyses to reach a conclusion. It will be necessary to prepare the data set before processing. The diagram shown in 4.1 was followed to carry out the analysis process.

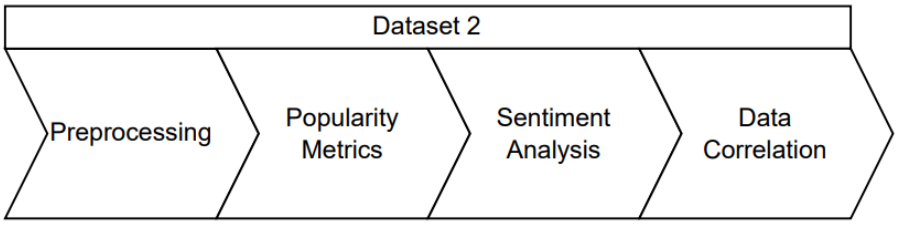

Figure 4.1: Analysis process

4.2 Preprocessing

The aim of this section has been to prepare the data before analyzing the popularity and sentiment metrics. We have included new columns to calculate popularity metrics, which have been chosen based on "Engagement Metrics" [5]. These measures allow us to know how successful a tweet has been in relation to its audience so that we can measure the popularity of each account relative to its community. On the other hand, we have used the NLTK library [1] to obtain a sentiment analysis of the collected tweets. This Python library has allowed us to compute the sentiment of a tweet as a numerical result. Now we will explain how we have calculated the metrics values and how many columns we have used in the study.

- **Popularity Metrics:**
  - **Tweet popularity**: Here, we can obtain the popularity ratio of a specific tweet. We have divided the number of favorites from that tweet into total followers to see the relative impact of that tweet with respect to our community. The formula to calculate is shown in Eq. 4.1.

$$TweetPopularity = \frac{favs \cdot 100}{followers} \tag{4.1}$$

  - **External Popularity**: Following the same line of calculation for the popularity of a tweet, we have now considered retweets that allow us to reach other people who are not our followers. An account with high external popularity is much more likely to increase its community in the short term. The formula to calculate is shown in Eq. 4.2.

$$ExternalPopularity = \frac{favs \cdot RT \cdot 100}{followers} \tag{4.2}$$

  - **Total Interaction**: With this metric, we calculate the sum of RT and Favs in order to obtain the global interaction of that tweet.

- **Sentiment analysis**
  - **POS**: This value indicates the percentage of positivity in the text. It has a value range from 0 to 1, where 0 indicates a total absence of polarity, and 1 indicates maximum polarity, positive in this case.

- **NEG**: This value indicates the percentage of negative text. It has a value range from 0 to 1, where 0 indicates a total absence of polarity, and 1 indicates maximum polarity, negative in this case.
- **NEU**: This value indicates the percentage of neutrality of the text. It has a value range from 0 to 1, where 0 indicates a total absence of polarity, and 1 indicates maximum polarity, neutral in this case.
- **Compound**: The compound value is an aggregated score that indicates the general sentiment of the text in a range of -1 to 1. A value of -1 indicates extremely negative sentiment, while a value of 1 indicates extremely positive sentiment. A value of 0 indicates absolute neutrality.

New columns have been added to the information we already had in Dataset 2. We can see an example of the tweets collected from the F1 account with the new columns in Figure 4.2. After collecting all tweets from the different accounts with added columns, several average values have been calculated to estimate the popularity and sentiment metrics of the different communities. The average values per account and per community have been calculated for the columns: "Fav", "Tweet popularity", "External popularity" and "RT". In addition, the average sentiments obtained in each tweet have been calculated both per account and per community to make a comparison between them. The data provided by the new columns and the average values will be used in the following sections of this chapter.

	User	RT	Fav	Total Interaction	Tweet	Followers	neg	neu	pos	compound	tweet popularity	external popularity
0	F1	430	4923	5353	The Formula 1 community is sending our thought...	9236616	0.000	0.657	0.343	0.8074	0.053299	22.918458
1	F1	376	4780	5156	What's going on here?! □\n\nVideo coming soon!...	9236616	0.000	1.000	0.000	0.0000	0.051751	19.458209
2	F1	56	926	982	Plenty has happened since our trip to Melbourn...	9236616	0.000	1.000	0.000	0.0000	0.010025	0.561418
3	F1	172	2148	2320	Reunited \n\n@PierreGASLY pays a visit to the...	9236616	0.000	0.875	0.125	0.5093	0.023255	3.999906
4	F1	554	9584	10138	Fernando with some driving tips for Nico □\n\...	9236616	0.000	1.000	0.000	0.0000	0.103761	57.483563
...	...	...	...	...	...	...	...	...	...	...	...	...
3245	F1	999	21247	22246	Lewis Seb\n\nThe latest Power Rankings are i...	9236582	0.000	1.000	0.000	0.0000	0.230031	229.800948
3246	F1	52	1584	1636	Despite missing out on the points, Alex Albon ...	9236582	0.000	0.773	0.227	0.6108	0.017149	0.891758
3247	F1	509	15011	15520	We need to agree on a name for "thing above th...	9236582	0.047	0.822	0.131	0.4822	0.162517	82.721065
3248	F1	495	6013	6508	There was drama from the very start on our las...	9236582	0.000	1.000	0.000	0.0000	0.065100	32.224420
3249	F1	72	3264	3336	"I love Texas, I love Austin but that race for...	9236582	0.146	0.610	0.244	0.2711	0.035338	2.544318

Figure 4.2: Preprocessed Dataset 2

4.3 Popularity Metrics

In this section, we analyze the tweets collected from the different accounts of each community. We have analyzed popularity metrics paying special attention to the JaramaFan and

F1 accounts that belong to community numbers 1 and 8. These two accounts have been chosen to observe the differences that exist in popularity metrics between large accounts with millions of followers and smaller accounts.

4.3.1 @Jaramafan and @F1 metrics comparison

First, we selected the @F1 account and got its total interaction to compare it with the average popularity data of tweets in each month. We observe that these two metrics are not related. The highest interactions occur when Formula 1 returns after a break, such as in March, June, and October. However, the popularity of tweets remains relatively constant throughout the year, except for December, when the season has already ended, or March, when the season begins. We can conclude that, in this case, generating more interaction does not necessarily increase popularity.

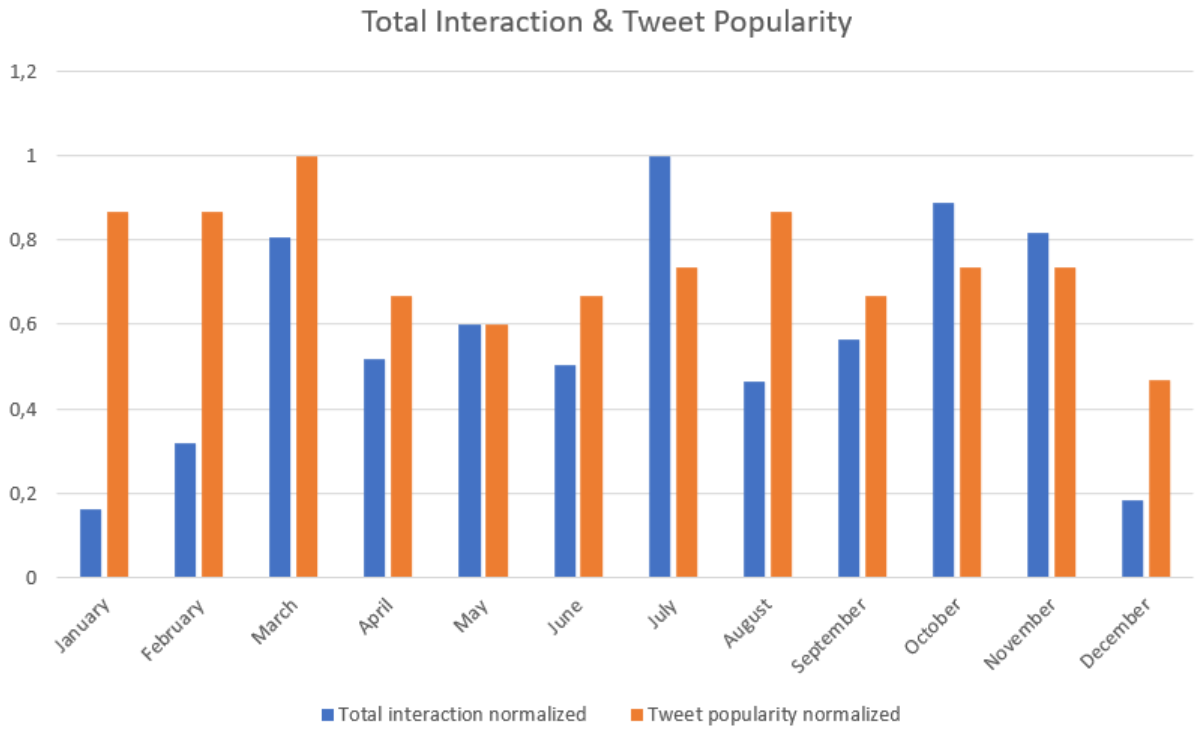

Figure 4.3: @F1 Total Interaction and Tweet popularity

Table 4.1 shows the popularity metrics for both accounts. As expected, the official Formula 1 account has a much higher number of favorites and average RT per tweet and also has much more total interaction. However, the popularity relative to each tweet measured with the "Tweet popularity" metric is higher in JaramaFan. It is interesting to note that a smaller account achieves better engagement with its audience. On the other hand, the "Ex-

ternal Popularity" metric achieves much higher values for the official Formula 1 account. This is because the number of Retweets (RTs) plays an important role in computing that value. A higher account will always have a better RT average, meaning better popularity in external accounts.

Popularity Metrics	JaramaFan	F1
Fav Average	97.64	8902.17
RT Average	9.67	709.21
Total Interaction	329.060	29.245.958
Tweet Popularity	0.73	0.10
External Popularity	7.04	72.71

Table 4.1: @F1 and @JaramaFan Metrics comparison in the last 3250 tweets

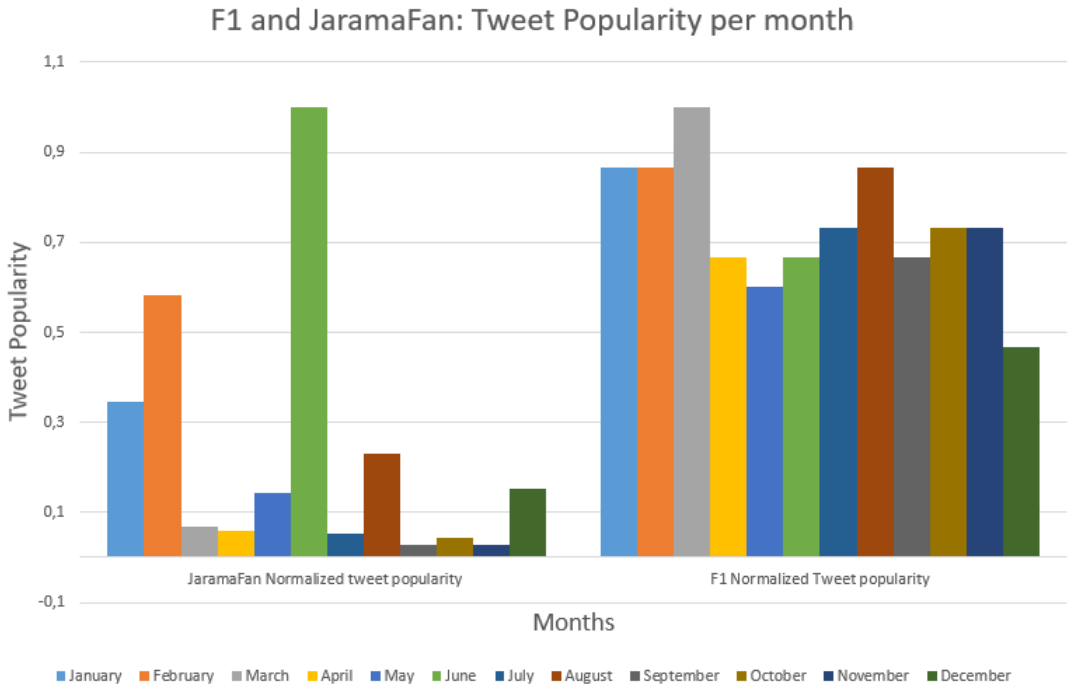

Figure 4.4: Tweet Popularity per month

In figure 4.4, we can observe the popularity of tweets in different months of the year, both calculated in normalized values. The JaramaFan account shows much larger variations than the official Formula 1 account. This suggests that accounts with a large number of

followers stabilize parameters such as "Tweet popularity", while smaller accounts rely on specific moments or news for their popularity. Formula 1 account achieves its highest popularity at the beginning of the season, in March, and its lowest popularity at the end of the season, in December. We therefore observe a predictable graph with minimal variations; from the maximum to the minimum value, showing a maximum difference of fifty percent. Therefore, after analysis, we can say that a higher number of followers provides a more stable popularity.

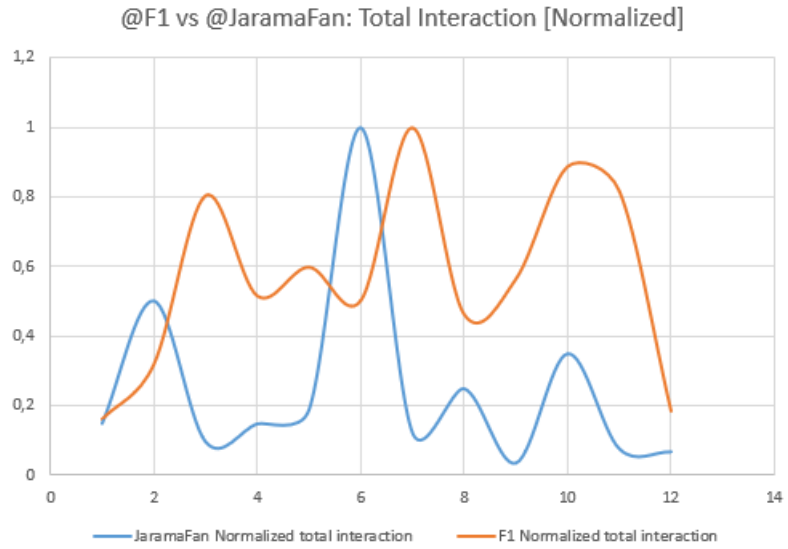

Figure 4.5: F1 and JaramaFan total interaction in normalized values

Finally, in Figure 4.5, we can observe the total interaction between the JaramaFan account and the F1 account in normalized values. First, we can see that the interactions of both accounts follow the same trend; however, it is very interesting to see that most of the JaramaFan tweets are in the right quadrant, resulting in a left-skewed distribution [7]. Accounts with a smaller number of followers seem to notice the change in interactions for a specific event earlier than accounts with a larger number of followers. "Total interactions" are much higher in larger accounts, but, using normalized data, we can see the trend changes. It is very interesting to see that, in this case, trend changes occur earlier in smaller communities, resulting in larger communities echoing the same trend days later.

4.3.2 Popularity metrics by Community

Next, we will discuss the results obtained from the popularity metrics by community. Through the data analysis process, we have observed how nationality and the number of followers can influence the popularity of tweets. Additionally, we have investigated whether

there are differences between communities created by a specific theme or news item and those that remain consistent over time. Based on the results, we have found that communities 4, 5, and 8 have the highest engagement with their followers. These communities are the Latin, English, and Spanish communities, which are communities that remain fixed over time. Other communities that are created as a result of a specific event have much lower popularity metrics, such as communities 3 or 6. Finally, communities 1, 2, and 7, all of them communities belonging to regulatory bodies of official organizations, are the ones with the lowest degree of popularity among their followers.

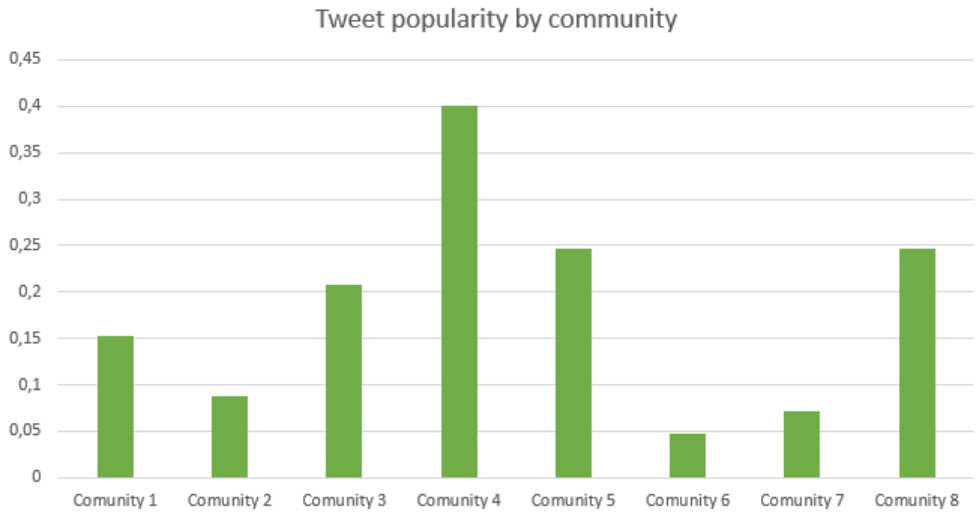

Figure 4.6: F1 Metrics

Type	Community	Tweet Popularity	External Popularity
Language	4,5 and 8	0.297	563.431
Specific New or event	3 and 6	0.126	84.396
Official and regulatory	1, 2 and 7	0.104	186.470

Table 4.2: Popularity metrics for each type of community

In table 4.2, we can observe the three types of community according to their creation with their respective popularity metrics on average. It is interesting to note that the communities that are quickly formed around an event achieve high metric values in a short period of time, positioning themselves even above other communities that remain fixed over time and have a larger number of followers, such as communities 1, 2 and 7, which are the official and regulatory type. Although we have a higher "Tweet popularity" metric in

community type number 2 formed by a new or an event, we can see in the "External popularity" metric that these communities do not have as many followers, which means that their "External popularity" decreases to the detriment of the rest of the communities. The power of retweeting in this metric means that even though recently created communities have a lot of participation and engagement, they cannot reach a large external audience, as other communities with more followers do.

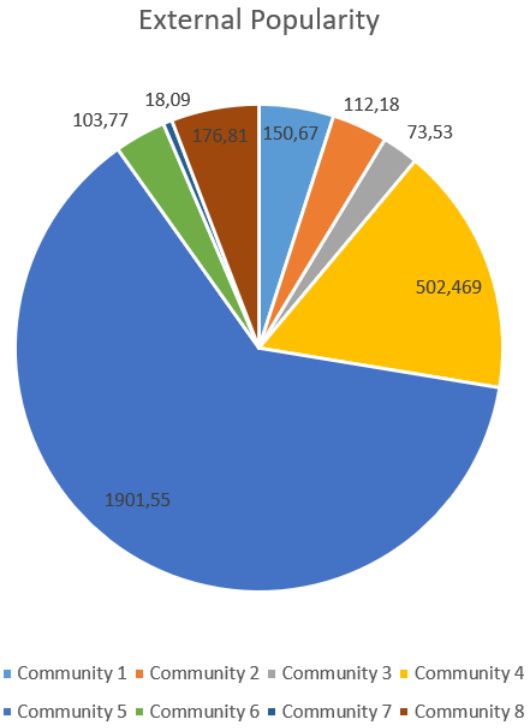

Figure 4.7: External Popularity by community

4.4 Sentiment Analysis

In this section, we will comment on the results obtained after applying sentiment analysis to the accounts of the different communities. As in the previous section, we have paid special attention to the JaramaFan and @F1 accounts to see the differences that can arise between large and small accounts. Finally, we have calculated the average sentiment metric for each account in order to make a comparison between communities and thus be able to see the differences between them.

4.4.1 @Jaramafan and @F1 comparison

First of all, we have compared the JaramaFan and F1 accounts. In the table 4.3, we can see an example of the tweets collected by the F1 account, in which we can see the tweets and their associated positive, negative, neutral, and general sentiments.

Tweet Message	Neg	Neu	Pos	Compound
We've brought in @DanielRicciardo to share some tips.	0.0	0.809	0.191	0.6369
The Mercedes driver runs wide and cuts the corner	0.081	0.919	0.0	-0.296
Drive to survive 5 is almost there	0.0	0.722	0.278	0.6597

Table 4.3: Examples of @F1 sentiment analysis

On the other hand, here we can see an extract of JaramaFan's tweets in Table 4.4 where the sentiment level of the tweets has been calculated. We observe that messages are divided into a component of positivity or negativity mixed with neutrality. It is interesting to see that totally neutral tweets achieve a "Compound" of zero on those messages. We also observe that if a tweet has a positive value, then its negative value is 0, and vice versa.

Tweet Message	Neg	Neu	Pos	Compound
La FIA es una vergüenza!	0.629	0.371	0.0	-0.526
A 3 horas del comienzo de la carrera esta es la parrilla	0.0	1.0	0.0	0.0
Se hizo justicia	0.0	0.37	0.63	0.526

Table 4.4: Examples of @JaramaFan sentiment analysis

In figure 4.8, we can observe the comparison between F1 and JaramaFan regarding the sentiment values obtained for the last 3,250 tweets of each account. In the first stage, we observe that despite having communities with very different numbers of followers, the relationship between positive, negative, and neutral percentages is practically identical. The negativity is slightly higher in the case of JaramaFan, but these values are almost negligible.

We have calculated the average values of the sentiment metrics for each of the accounts and then calculated the average values for each community. As we can see in Table 4.5, we have represented the average values of each of the parameters that measure the sentiment of the tweet. The values represented in green are the highest values of that parameter,

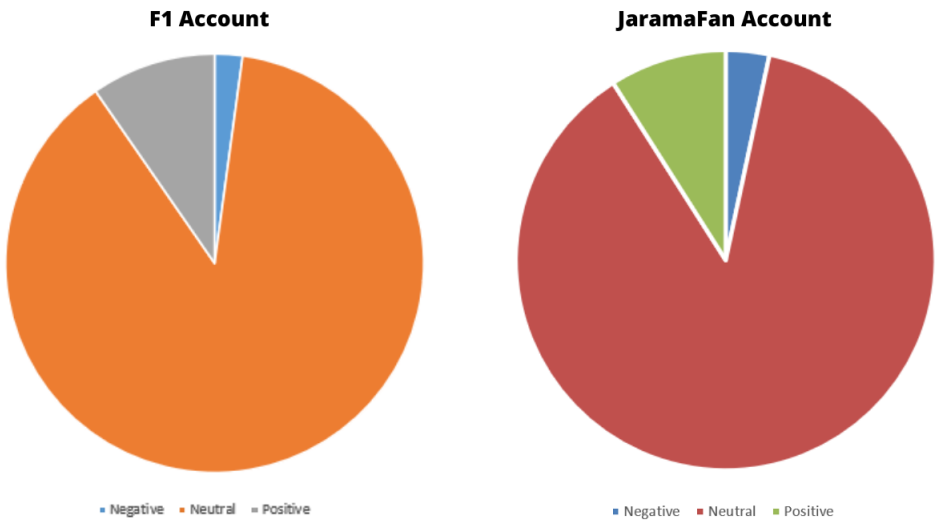

Figure 4.8: Comparison between JaramaFan and F1 sentiment analysis

while the red ones are the lowest. Community 5, the English community, stands out as having the highest negative sentiment, while Community 7, the community belonging to the regulatory bodies, has little negativity. We can see that these two communities are both the most extreme, being the most positive in the case of community 7 and the most negative in the case of community 5. However, despite not having such a high negativity ratio, the most negative community overall is community number 3, the one belonging to newspapers and media. It should be noted that all the communities have very similar values and that the compound of all of them is greater than 0, which means that, on average, they all have a positive sentiment.

Sentiment Analysis	Neg	Neu	Pos	Compound
Community 1	0.021	0.854	0.123	0.243
Community 2	0.025	0.862	0.107	0.199
Community 3	0.044	0.860	0.096	0.153
Community 4	0.014	0.849	0.146	0.253
Community 5	0.087	0.825	0.152	0.301
Community 6	0.027	0.843	0.126	0.240
Community 7	0.012	0.900	0.087	0.200
Community 8	0.030	0.867	0.102	0.169

Table 4.5: Examples of @JaramaFan sentiment analysis

4.5 Correlation

Now, as a final part of the study, we have performed the correlation between popularity metrics and sentiment metrics in all the tweets collected, belonging to Dataset 2 obtained in Chapter 3. In this section, we look for a relationship between these two metrics that can tell us whether the sentiment of tweets is reflected in higher or lower popularity. We have carried out a graphic analysis of the results obtained. We have analyzed in detail the relationship of the variables using Pearson's coefficient [14], we have analyzed the differences between positive and negative tweets in terms of popularity. Finally, we have made a series of hypotheses. We plotted, for each of the accounts, the compound of the sentiments of their tweets on the x-axis, with a range from -1, extremely negative, to 1, extremely positive. On the other hand, we have plotted the "Tweet popularity" metric on the y-axis. In Figure 4.9, we can see an example of this for the Formula 1 account. We observe how the dots representing the collected tweets are distributed along all the sentiments on the x-axis. In this case, we observe how the tweets with the highest popularity are achieved for neutral sentiment values, which means that the tweets that achieve the highest success relative to their audience are the tweets whose sentiments are completely neutral. On the other hand, we see that in the case of the Formula 1 account, we have a left-skewed distribution [7], with very few tweets with completely negative sentiments. In the following sections, we will collect the same data for each of the accounts representing the eight communities in such a

way as to obtain a series of hypotheses on the results.

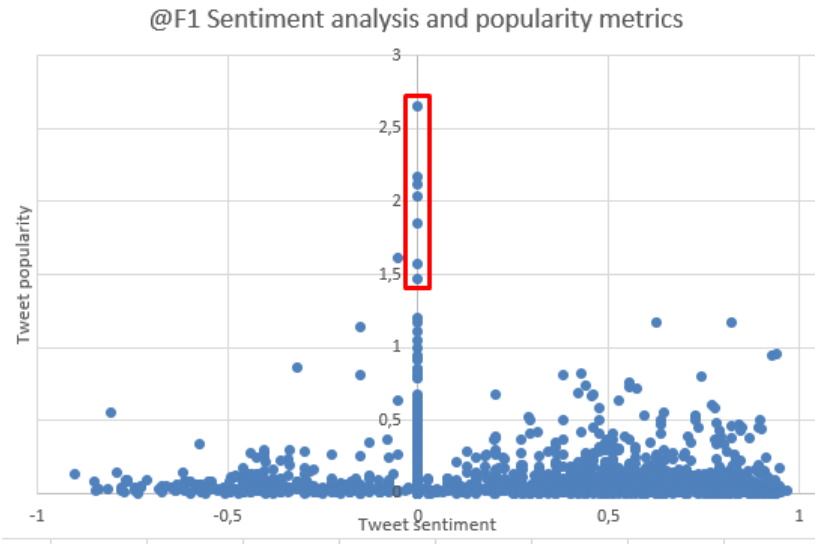

Figure 4.9: @F1 tweet sentiment and popularity

4.5.1 Correlation using Pearson Coefficient

To know exactly whether there is a proportional relationship between the popularity of the tweets and each of the sentiments, we calculated the Pearson coefficient [14] for each of the accounts in each community. The Pearson correlation coefficient, also known as Pearson's r or simply the correlation coefficient, is a measure of the linear relationship between two variables with a range of values from -1 to 1, as it is defined as shown in Eq. 4.3.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4.3}$$

We used the value of this coefficient to obtain the relationship between the following pairs: positive tweets and their popularity, negative tweets and their popularity, and neutral tweets and their popularity. We then averaged the results by the community, obtaining the results represented in 4.6.

In view of the results, we observe that for tweets with neutral sentiments, in all communities except community 4, the Latin community, negative values are obtained in the Pearson coefficient [14]. This means that there is a negative correlation between neutral sentiment and popularity. Thus, we can understand that neutral tweets tend to be less popular. On the other hand, for both positive and negative tweets, we tend to find a positive Pearson coefficient [14]. Although there are communities that do not meet this

Pearson Coefficient	Pos-Pop	Neg-Pop	Neu-Pop
Community 1	0.058	-0.023	-0.044
Community 2	0.058	0.018	-0.061
Community 3	0.005	0.035	-0.024
Community 4	-0.028	0.002	0.028
Community 5	-0.006	0.037	-0.011
Community 6	0.011	-0.015	-0.003
Community 7	-0.027	0.455	-0.001
Community 8	0.026	0.020	-0.032
Total	0.012	0.014	-0.018

Table 4.6: Examples of @JaramaFan sentiment analysis

statement, both positive and negative tweets obtain a positive Pearson coefficient [14] on average. Therefore, we can understand that both positive and negative tweets tend to be more popular.

4.5.2 Graphic Analysis

We have collected the results of the 25 most influential accounts on the network, as we selected in Chapter 3, which means that we have classified 77,261 tweets, establishing a correlation between their popularity metrics and their sentiments. We have represented the same values as we did in the example shown in Figure 4.9, using the x-axis to represent the sentiment of the tweet and the y-axis to represent the popularity of the tweet. The dots correspond to the different tweets of an account. In this section, we have analyzed the results graphically, reaching different conclusions. We have represented the results by the community in such a way that each account has been assigned a color. In Figure 4.10, we have the analysis of the first four communities.

We observe that in all of them, the patterns observed in the previous Figure 4.9, are repeated. The number of dots is slightly shifted to the right, which means that we have a left-skewed distribution [7], with a higher number of tweets in the positive quadrant of the

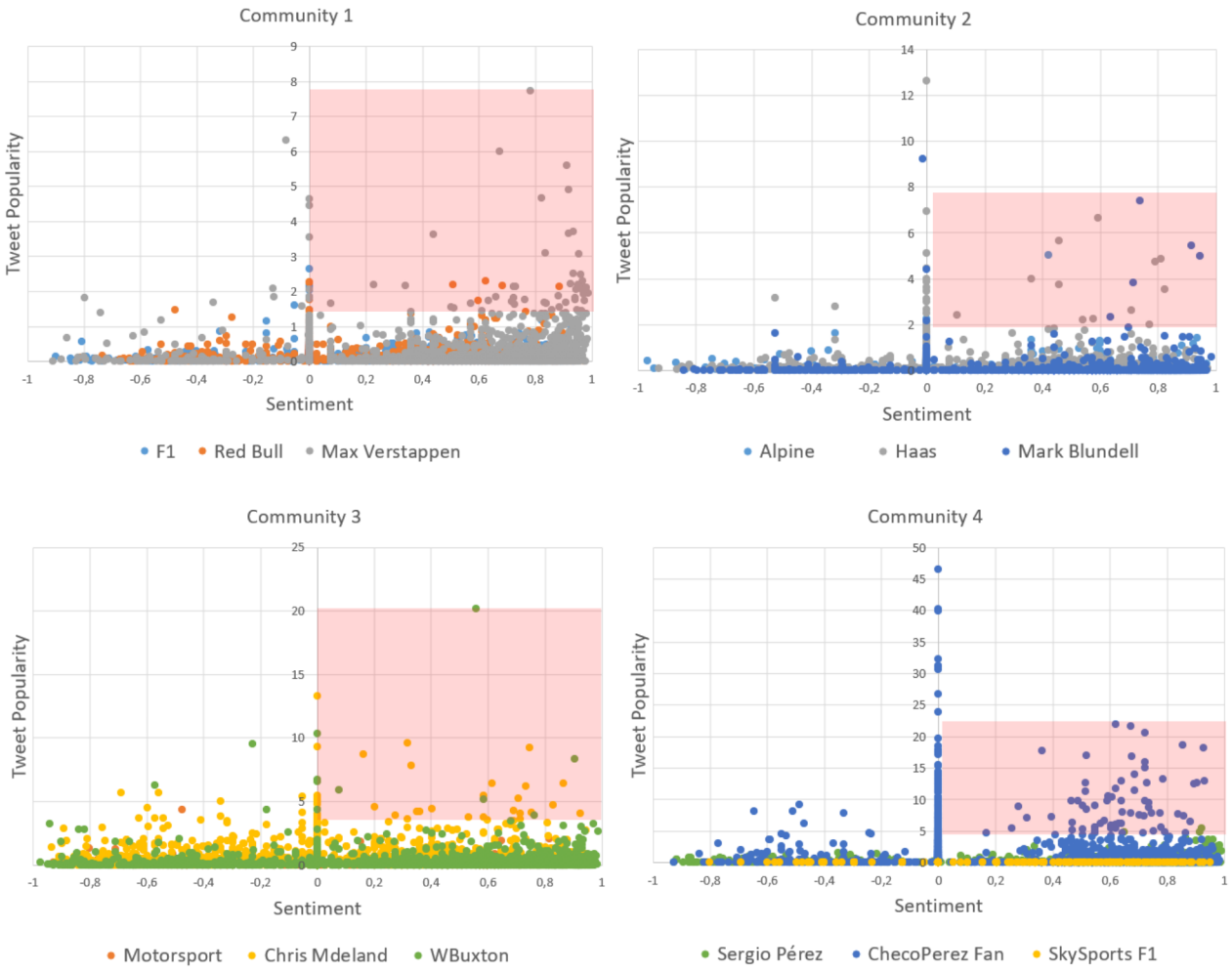

Figure 4.10: Correlation analysis in communities 1 to 4

graphic. On the other hand, neutral values achieve very high popularity scores; however, it cannot be established that neutral tweets achieve greater popularity than tweets with positive or negative sentiments. It will be necessary to use Pearson's coefficient [14] in order to obtain a clear conclusion on the relationship between popularity and the different sentiment metrics: positive, negative, and neutral. Lastly, we have observed a very interesting phenomenon: in all communities, we see that, for a certain threshold of popularity, tweets in the positive quadrant achieve much better results. In other words, positive tweets have a higher chance of obtaining high-popularity metrics. This phenomenon is represented by a red square in 4.10 and in Figure 4.11. For a certain popularity threshold, there is a higher number of tweets in the right quadrant, indicating that it is much more frequent to obtain high values for positive tweets than for negative tweets once we have reached a certain value of popularity.

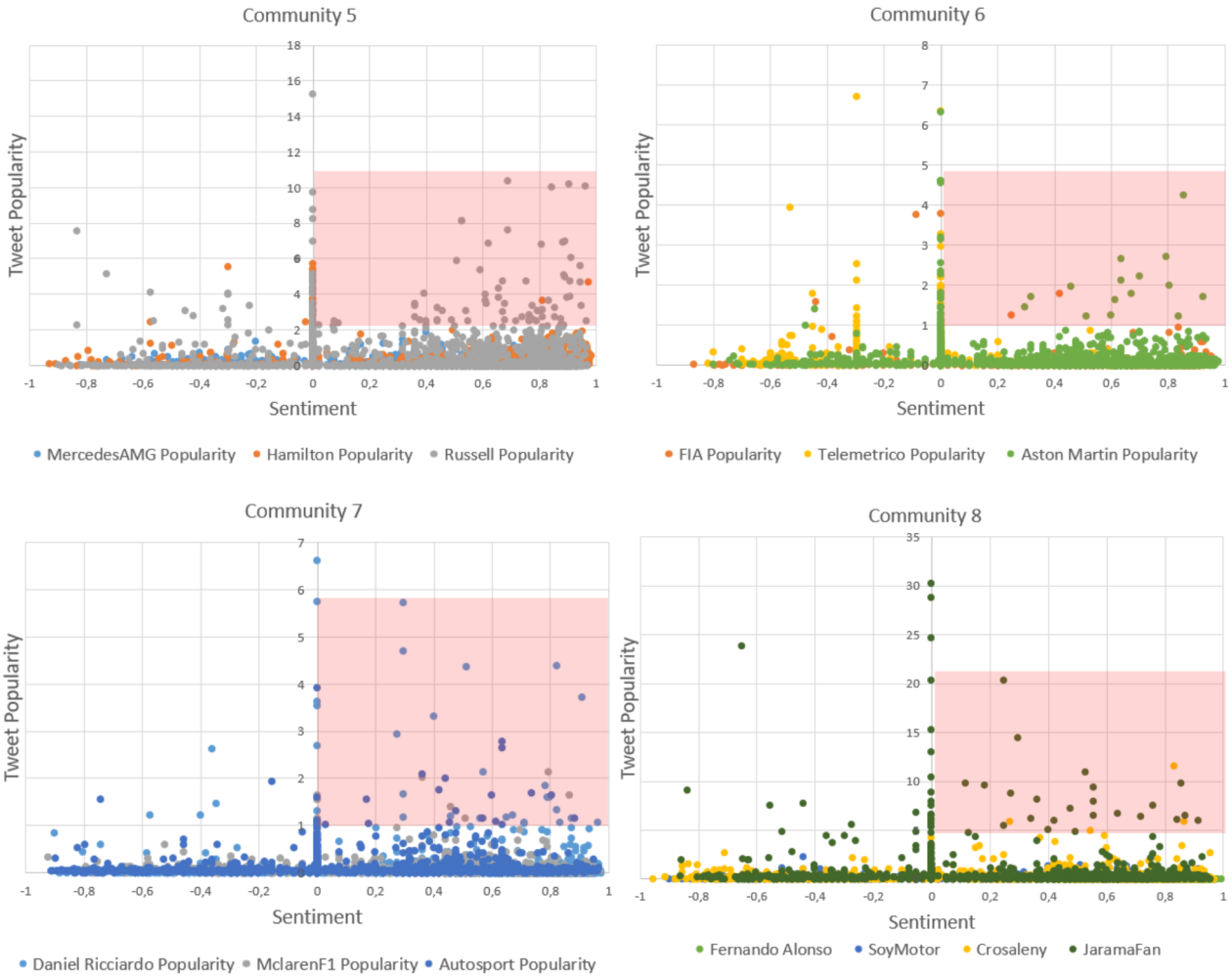

Figure 4.11: Correlation analysis by community II

In Figure 4.11, communities 5, 6, 7, and 8 are analyzed. The first observations hold true in the same way as in the previous communities. A higher number of tweets is observed in the right quadrant than in the left quadrant. On the other hand, the observations from the first four communities hold true the same way here, for a certain popularity threshold, it is much more frequent to find tweets with positive sentiments than with negative sentiments, suggesting that writing positive tweets increases the likelihood of achieving higher popularity. We can conclude the correlation study by stating that after analyzing more than 75,000 tweets, a clear pattern is observed in which tweets with positive sentiments achieve better popularity metrics. However, although we see a clear pattern that positive tweets, on average, are more likely to achieve higher levels of popularity, we must check what average popularity is obtained by separating positive and negative tweets for each account. The analysis of that case has been done in Subsection 4.5.3.

4.5.3 Popularity Comparison between Positive Tweets and Negative Tweets

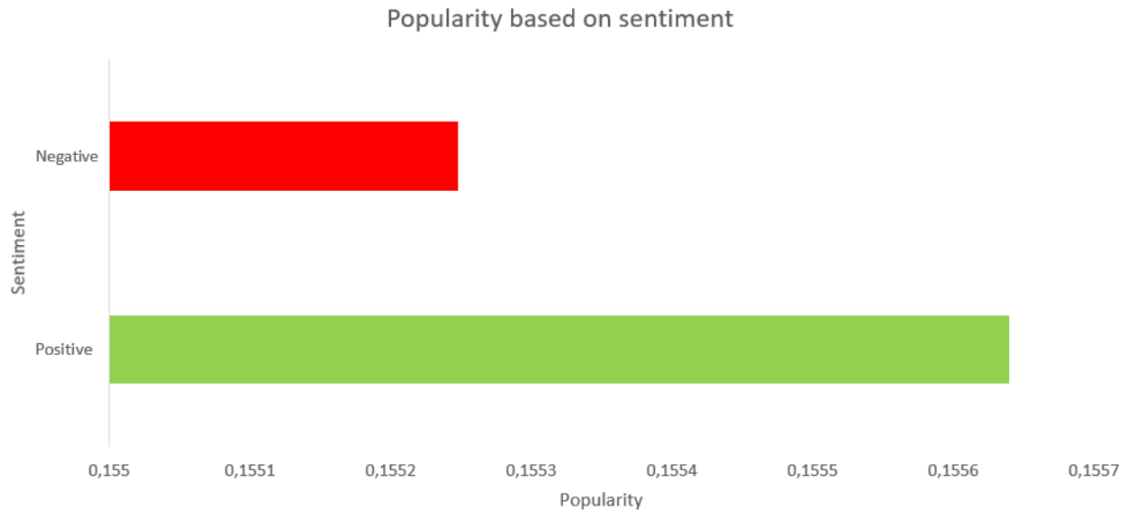

Figure 4.12: Popularity in positive and negative tweets

In this section, we have calculated the average popularity of positive tweets and the average popularity of negative tweets. In relation to what we have studied in subsection 4.5.2, we wanted to find out whether the average popularity of positive tweets is higher than negatives. This study has been carried out for each of the accounts in the global network, and we have obtained the overall average represented in Figure 4.12. We see that positive tweets are more likely to have higher popularity rates although the difference is not excessively

Com.	Account	Popularity in +	Popularity in -	Hypothesis 1 and 2
1	@F1	0.081	0.071	Accomplished
	@RedBull Racing	0.151	0.097	Accomplished
	@Verstappen	0.241	0.227	Accomplished
2	@AlpineF1Team	0.066	0.059	Accomplished
	@HaasF1team	0.161	0.100	Accomplished
	@MarkBlundell	0.038	0.025	Accomplished
3	@MotorSport	0.047	0.042	Accomplished
	@Chrismdelandf1	0.435	0.434	Accomplished
	@wbuxton	0.170	0.190	Not Accomplished
4	@schecoperez	0.130	0.118	Accomplished
	@skysport	0.0003	0.0002	Accomplished
	@checofanpage	0.656	0.588	Accomplished
5	@Hamilton	0.266	0.313	Not Accomplished
	@Russell	0.305	0.347	Not Accomplished
	@mercedesamgf1	0.109	0.135	Not Accomplished
6	@mclarenf1	0.064	0.066	Not Accomplished
	@autosport	0.057	0.034	Accomplished
	@danielricciardo	0.065	0.060	Accomplished
7	@FIA	0.018	0.060	Not Accomplished
	@AstonMartinF1	0.072	0.081	Not Accomplished
	@TelemetricoF1	0.090	0.164	Accomplished
8	@FernandoAlonso	0,0142	0.0165	Not Accomplished
	@soymotor	0.086	0.078	Accomplished
	@crosaleny	0.176	0.144	Accomplished

Table 4.7: Average popularity in positive and negative tweets

large. On the other hand, we have plotted the result of each of the accounts in table 4.7. Here we see that after analyzing the accounts separately, there are some accounts that do not fulfill this hypothesis, as accounts such as @FIA, @TelemetricoF1, @FernandoAlonso, or @AstonMartinF1 obtain better popularity for their negative tweets. We cannot completely affirm that positive tweets obtain better popularity results. However, we cannot deny the results obtained in 4.5.2 that for a certain threshold, there is a higher probability of obtaining tweets with high popularity if the tweet is positive. We will have to investigate other communities to see if this is an isolated case of the Formula 1 community. On the other hand, it would also be necessary to carry out a study with more data, as it is possible that using a greater number of tweets could change the averages, and we could confirm that positive tweets are more popular than negative tweets in each account.

4.5.4 Hypothesis after correlation

After carrying out the correlation study and having analyzed the results in each of the previous sections, we have collected the hypotheses, which we will explain in the following:

- **Hypothesis 1**: After performing the graphic analysis in Subsection 4.5.2, we can say that in view of the results obtained, **there is a specific threshold in the popularity metric at which positive tweets are more likely to be found than negative ones.**
- **Hypothesis 2**: After performing the analysis using the Pearson coefficient [14] in subsection 4.5.1, we have found that **positive and negative tweets tend to be more popular than neutral tweets.**
- **Hypothesis 3**: Finally, by performing the analysis of the popularity comparison between tweets in 4.5.3 we have found that at **positive tweets are, on average, more popular than negative tweets.** This assumption is obtained by averaging over all accounts, although there are individual accounts for which this condition is not met. A study with a larger number of data will be necessary to confirm this.

CHAPTER 5

Case study

5.1 Introduction

In this chapter, we have performed a detailed analysis of one of the communities within the global Formula 1 network. Specifically, we have focused on Community number 8, obtained in Chapter 3, which represents the Spanish community and includes our account, JaramaFan. We have explored the correlation between sentiment analysis and popularity metrics of the most influential nodes in Community 8. Our objective is to determine whether these metrics are related and whether the results align with those obtained in the previous chapter, Data Analysis 4. Furthermore, we have conducted a graph study to examine the interactions between Community 8 and other communities to see the importance of the Spanish community within the global network.

5.2 Community 8: Graph Analysis

In this section, we have calculated the total number of nodes and edges from Community 8 to observe how many interactions they have with the rest of the network. We have used Dataset 1, which represents the initial graph without filtering obtained in Chapter 3

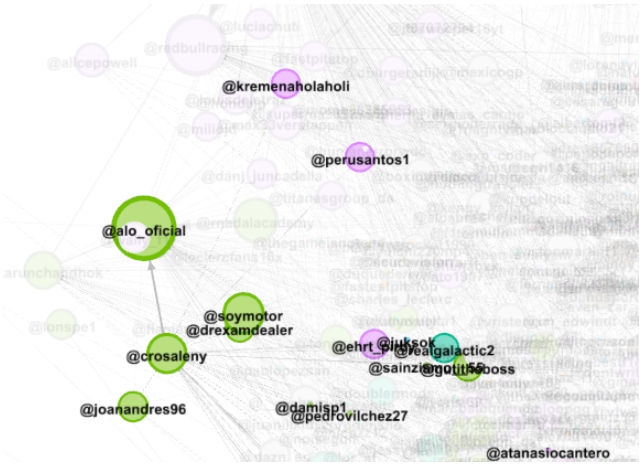

Figure 5.1: Selected nodes from Community 8 interacting with other nodes

	Community 8	Global Network
Nodes interacting	1,104	4,488
Edges	349	8.648
Net diameter	3	8
Average path	1.121	2.350
Most important Node	@FernandoAlonso (349)	@F1 (1602)

Table 5.1: Community 8 parameters

to carry out this study. By calculating this, we have that community 8, represented in Figure 5.1, interacts with 1.104 nodes, which means 24.6% On the other hand, @F1 is the most critical node in the global network, with 1602 connections. Notably, only one of the eight communities comprising the Formula 1 community interacts with one-quarter (1,104) of the total nodes (4,488). We can see that even though there are many independent communities, there is a high level of interaction between all the accounts that make up the Formula 1 community.

5.3 Community 8: Sentiment analysis

In this section, we have performed a tweet sentiment analysis of the most important nodes in Community number 8, selected in Chapter 3. We wanted to check whether sentiments are equally distributed in the community accounts or whether there are differences between them. We can observe in Figure 5.2 that negativity, neutrality, and positivity are not related. Accounts that accumulate many positive tweets can also get negative tweets, so there is no clear trend for an account to be more positive or negative. However, all accounts in Community 8 tend to have high neutrality. As seen in Figure 5.2, @soymotor is the most neutral profile; this may be because it is a business account. However, any of the four most essential nodes have a neutrality close to 0.9 with almost no differences. In terms of negativity and positivity, we can observe more significant differences. @FernandoAlonso is the account with the minor negativity and the highest positivity. This is reflected in the compound value, which is close to 0.3. @crosaleny and @jaramafan follow it, accounts with a more positive relationship between their sentiments. We can see how the positivity metric strongly influences the compound metric. The high neutrality and low negativity mean that the positivity value ultimately determines the total compound metric.

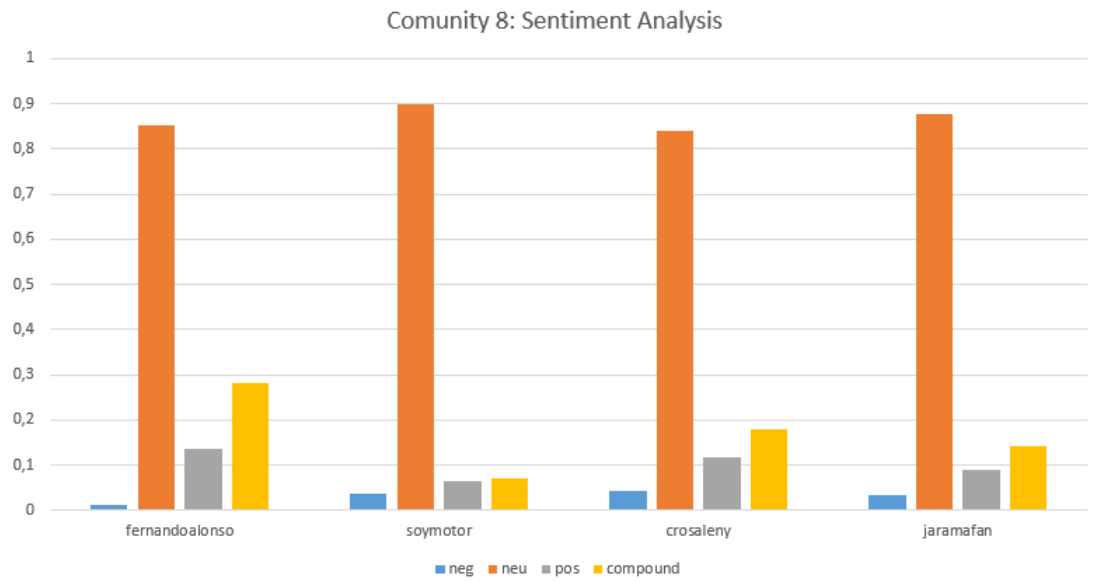

Figure 5.2: Community 8 sentiment analysis

5.4 Popularity Metrics

On the other hand, we have computed all "Popularity Metrics" for each account. As we saw in Chapter 4, we wanted to see the differences in popularity between accounts, in this case, those belonging to Community 8. As we can observe in Table 5.2, we have computed the "Average fav per tweet", the "Average retweet per tweet", and the " External Popularity" on all accounts. As expected, Fernando Alonso reaches the highest "Favs per tweet" and "Retweets per tweet" because it has the most followers. On the other hand, we have computed "Tweet Popularity" for each account. In this case, Jaramafan is the account with the highest "Tweet popularity" about its tweets. Although accounts like Fernando Alonso, with more than 3 million followers, or SoyMotor, with 200,000 followers, have a significant number of likes, the popularity of their tweets is lower than others. As was observed when analyzing communities in Chapter 4, smaller accounts and communities tend to engage more with their audience, as reflected in "Tweet Popularity". However, larger accounts or communities achieve greater "External Popularity" thanks to the importance of RT in this metric, which allows them to reach a larger audience.

	Alonso	VictorAbad	Crosaleny	SoyMotor	JaramaFan
Fav Average	3341.54	1988.44	77.24	167.27	77.77
Rt Average	482.106	123.897	28.324	10.300	29.49
Tweet Popularity	0.095	0.823	0.0983	0.0697	0.8541
External Popularity	45.8418	102.024	2.785	0.7187	30.432

Table 5.2: Community 8: Popularity Metrics

5.5 Correlation

Now we will calculate how much correlation there is between the sentiment of the tweets and their popularity. We have analyzed the most critical nodes in Community 8, selected in Chapter 3, collecting more than 12.000 tweets. We have carried out the same correlation study as in Chapter 4 but this time focused on the selected accounts of Community 8. We will use Pearson's coefficient to measure the correlation between popularity and sentiment metrics, perform a graphical study of the results, and average the filtered popularity results for positive and negative tweets. Finally, we will collect the hypotheses formulated in the study of the global network, mentioned in Chapter 4, and see if they are also fulfilled for this specific case. In Table 5.3, we can see the most successful tweets of the chosen accounts in community 8, where we conducted the study. We see that tweets that obtain the best popularity are related to tweets with a positive compound. As we saw in chapter 4, there seems to be a direct relationship between high popularity and positive sentiments on tweets. As we can see in Table 5.3, we have obtained the three most popular tweets from each account to observe this trend. We have colored those tweets with a positive compound green and those with a negative compound red. Given the results, we can see that, except @SoyMotor, the tweets with the best popularity metrics are obtained with neutral or positive sentiment. It is surprising to note that in the case of JaramaFan, his most popular tweets are neutral. In the following sections, we will analyze the results in depth to get more precise conclusions.

5.5.1 Graphic Analysis

In this section, we have graphically represented the tweets collected by the Community 8 accounts to conclude. We have carried out the same analysis as in chapter 4, where we

	Alonso	Crosaleny	SoyMotor	JaramaFan
Tweet popularity #1	8.010	11.432	2.276	30.174
Compound #1	0	0	-0.439	0
Tweet popularity #2	3.965	5.841	1.429	28.717
Compound #2	0.833	0.866	0	0
Tweet popularity #3	3.922	5.785	1.339	24.601
Compound #3	0	0.269	0.511	0

Table 5.3: Average Sentiment Analysis

correlated the popularity and sentiment of the tweets. Still, this time we focused on each of the accounts that make up Community 8. In figure 5.3, we have represented the four main accounts of community 8 in which we have defined the compound value that measures the sentiment of the tweets on the x-axis with values between -1 and 1. In addition, we have described the popularity of the associated tweets on the y-axis. Given the results, we obtained similar values. The values of the tweets with high popularity seem to concentrate on the right side of the graph, which corresponds to the tweets with positive sentiments. The hypothesis obtained from Chapter 4 related to this chapter was: "**There is a specific threshold in the popularity metric at which positive tweets are more likely to be found than negative ones.**" As was the case in Chapter 4, we have reached the same conclusion for this community. Knowing that we have quantified the number of positive and negative tweets in the quadrant surpassing this threshold. In this way, we wanted to establish a relationship between the threshold and the probability that tweets are in either quadrant. In Table 5.4, we have represented its analysis. The "Max Popularity" feature means that each account's tweet has the highest popularity. We also calculated "Threshold", the popularity value at which we start seeing more tweets in the positive quadrant. After obtaining these values, we counted the tweets that surpassed this threshold in both the positive and negative quadrants.

Observing the results in Table 5.4, we can see that in all accounts, the number of positive tweets that exceed the threshold is greater than the number of negative tweets. It is also interesting to note that the threshold manifests itself around a specific value between accounts. We observe a relationship between the max popularity and the threshold, typically between 4 and 5. In the case of Fernando Alonso, if we were to exclude his most viral tweet,

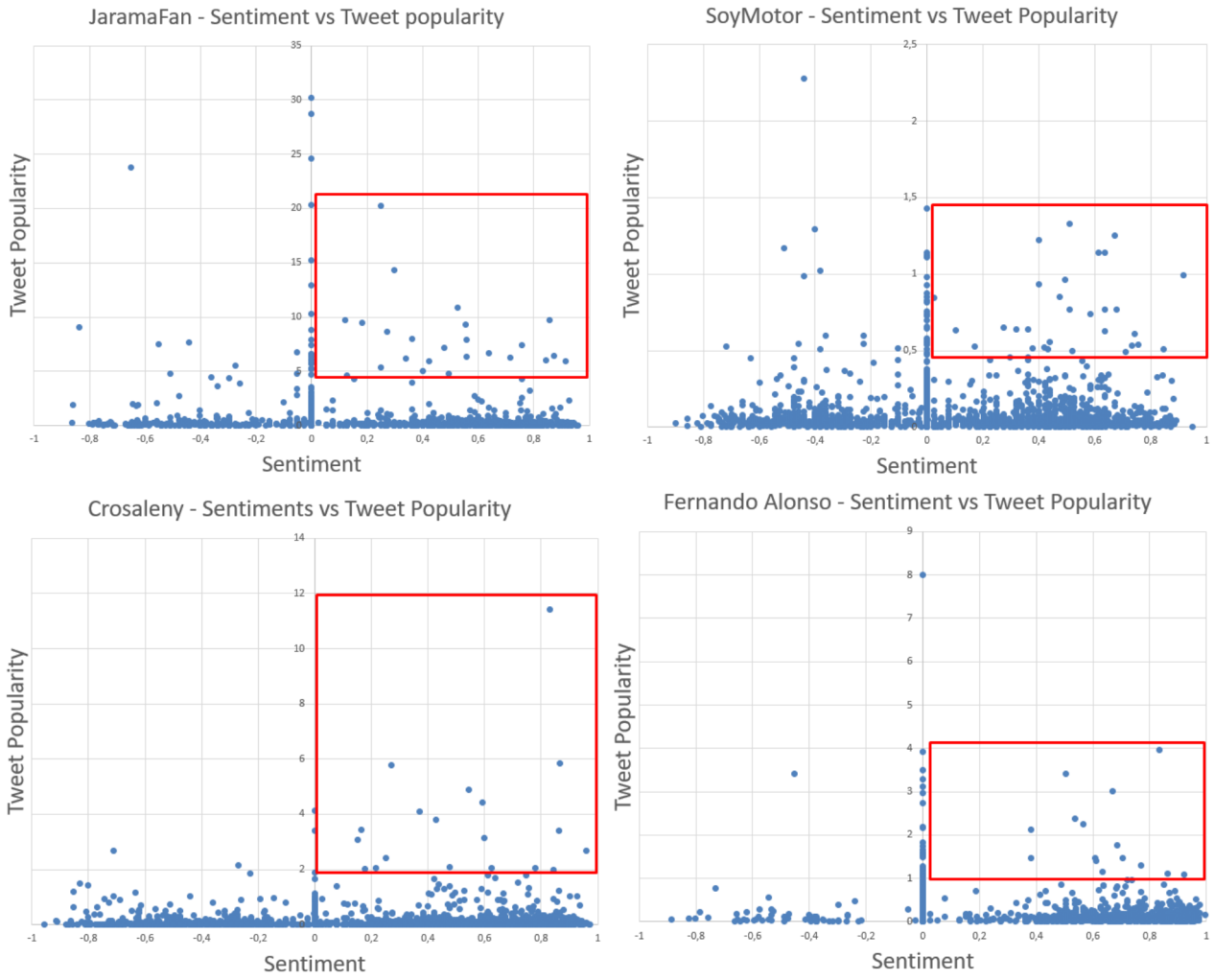

Figure 5.3: Accounts from community eight correlation

	Alonso	Crosaleny	SoyMotor	JaramaFan
Max Popularity	8.010	11.432	2.276	30.174
Threshold	1	2	0.5	5
Number of + tweets	15	19	30	22
Number of - tweets	1	2	11	5

Table 5.4: Threshold from Community 8

the same relationship would hold. With such limited data, it is challenging to draw definitive conclusions. Still, it appears that there is a relationship between maximum popularity, threshold, and the number of positive and negative tweets exceeding the threshold. We can conclude this section by saying that the hypothesis obtained in Section 4 is also valid here.

5.5.2 Pearson Correlation Coefficient

As we did in Chapter 4, to find out if there is a relationship between the popularity variables and the variables that measure sentiment, we have used Pearson's coefficient, which measures between -1 and 1, the degree of positive or negative correlation between two variables. For this analysis, we focus on whether positive or negative tweets are more popular than neutral tweets. In this way, we have related positive and negative tweets to their respective popularity metrics, and the same applies to neutral tweets. We have performed this metric for each of the tweets from the four selected accounts and obtained the results shown in Table 5.5.

Pearson Coefficient	Pos-Pop	Neg-Pop	Neu-Pop
Fernando Alonso	0.019	0.016	-0.023
SoyMotor	0.048	0.014	-0.047
Crosaleny	0.039	0.027	-0.050
JaramaFan	-0.003	0.021	-0.010
Total	0.026	0.020	-0.032

Table 5.5: Pearson's coefficient applied in Community 8

We observe that for all accounts, there is a negative correlation between neutral tweets and popularity. On the other hand, in all accounts except JaramaFan with positive tweets, we observe a positive correlation between positive and negative tweets and their popularity. Furthermore, when averaging the results, we obtain positive values of Pearson's coefficient [14] for positive and negative tweets and a negative one for neutral tweets. Given the results and the hypothesis we obtained in the previous chapter: "**Positive and negative tweets tend to be more popular than neutral tweets**", we can affirm that the hypothesis is also fulfilled in this specific case.

5.5.3 Positive and Negative Tweets

In this section, we try to determine how much popularity positive tweets gain compared to negative tweets. We have carried out the same study as in Chapter 4 but applied it to the accounts that make up community 8. In Figure 5.4, we plotted, for each of the accounts, the average popularity of positive tweets colored green and the average popularity of negative tweets colored red. Furthermore, in Figure 5.5, we have represented the average between the popularity achieved by positive tweets and the popularity gained by negative tweets. The hypothesis we obtained in Chapter 4 for this section was: "**Positive tweets are on average more popular than negative tweets**". Given the results, we arrive at the same conclusion here. In all accounts, except for Fernando Alonso, positive tweets achieve more popularity on average. Also, the popularity of tweets on average obtained in Figure 5.5 adding up all accounts shows this trend. The differences are slight, but the more data used, the more the hypothesized trend appears.

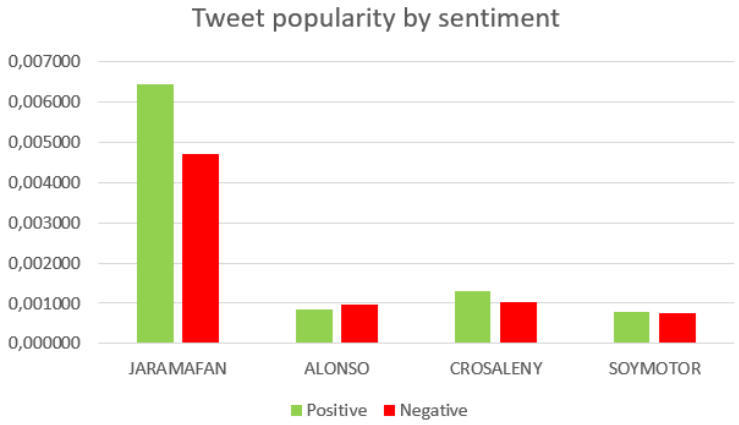

Figure 5.4: Positive tweets and negative tweets with average popularity

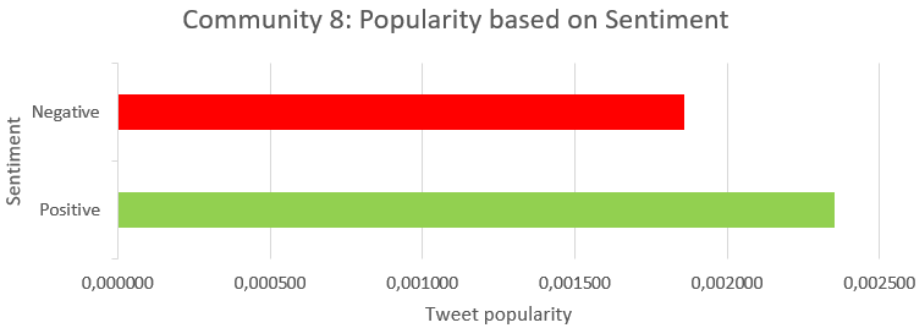

Figure 5.5: Positive tweets and negative tweets with average popularity II

5.5.4 Conclusions

After performing the correlation analysis for the accounts chosen in Community 8, we can confirm the hypotheses formulated in Chapter 4:

- Hypothesis 1: There is a specific threshold in the popularity metric at which positive tweets are more likely to be found than negative ones.
- Hypothesis 2: Positive and negative tweets are more popular than neutral tweets.
- Hypothesis 3: Positive tweets are, on average, more popular than negative tweets.

They are all fulfilled in the same way. Without analyses with other communities or with more data from other accounts, we cannot yet deny any of the hypotheses formulated.

CHAPTER 6

Conclusions and future work

6.1 Introduction

In this chapter, we will describe the conclusions obtained from the analysis of this project, the goals we have achieved, and some limitations that we have faced during the analysis process. In addition, we will provide some suggestions for future work.

6.2 Conclusions

In this section, we present the conclusions that have been drawn throughout all phases of the project.

First, we will focus on Chapter 3: "Data Acquisition". The first conclusions we have drawn here are that the Formula 1 community is very divided, with up to 501 communities, with a relatively low average degree of node entry. After the network filtering process, we reduced the number of communities and redundant nodes to eight communities that accumulated the most interactions. Furthermore, we have seen that nodes with many links tend to be nodes with increased popularity, being the center of a defined community. In

this sense, another conclusion we have drawn is that communities are formed differently. As we reviewed in Chapter 4: "Data Analysis", communities can be organized by geographical location, language, specific news item, or by the fact that they could belong to regulatory accounts. On the other hand, we also find communities that remain fixed over time and others created around a specific news item.

After that, in Chapter 4, we have analyzed more deeply the main nodes of the eight communities that have emerged in the Formula 1 network. We have drawn several conclusions. First, we have concluded that accounts with smaller followers have higher engagement than massive accounts with millions of followers, resulting in higher popularity among their audience. Another important finding is that communities 4 and 8 are the communities that achieve the best popularity metrics. These communities correspond to the Spanish and Latin communities. It is curious to see how the most thriving communities are Spanish-speaking, even ahead of communities in which the primary node is Formula 1 or official accounts of the category. Another conclusion we have found is that smaller accounts show a significant variation in tweet popularity. Unlike larger accounts, which maintain consistent popularity throughout the year, smaller accounts experience fluctuations based on the season or specific news stories. These smaller accounts are more sensitive to changes, as evidenced by trends. When a movement occurs, smaller accounts feel the impact on popularity and follower growth earlier.

Focusing on the analysis of emotions, we have drawn different conclusions. First, all accounts belonging to the Formula 1 social network tend to write tweets with a high neutrality rate. In fact, in all accounts analyzed, the overall average sentiment of tweets varies very little from account to account. On the other hand, all accounts, without exception, write more positive tweets than negative ones. In all the graphs, we saw a clear trend of tweets on the positive side of the chart, indicating that the tweets written by that account are primarily positive.

Finally, as a fundamental part of the project, we have drawn exciting conclusions about the correlation between popularity and sentiment study. Following the study of the relationship of the variables carried out both in chapter 4and in the case study in chapter 5 using Pearson's coefficient [14], we can affirm that positive and negative tweets tend to obtain higher popularity than neutral tweets. On the other hand, we have carried out a graphical study of the results obtained in the popularity and sentiment metrics. In chapter 4 and chapter 5, we concluded that there is a specific threshold in the popularity metric at which it is more likely to find positive tweets. As a final part of the correlation study, we observed that positive tweets are more popular on average. Of course, this was not the case in some

accounts, but when we accumulated the data, the trend was satisfied.

6.3 Achieved goals

This section details the objectives that have been achieved throughout the project.

- **Obtain Formula 1 Network Graph**.

  The first objective was to obtain the Formula 1 network map to perform our analysis. Once we obtained the Twitter API and connected it with the Gephi tool [8], we could visualize the representation of the network and generate the global network data set.

- **Community Detection and Filtering information**.

  Another accomplished objective was related to information filtering. In this project, it was crucial to minimize the number of redundant links to identify the most important communities in the network and focus on the nodes that generate the most impact for our study. We applied the Louvain algorithm [6] to identify several communities.

- **Analysis and of the different datasets**.

  We proceeded with the analysis phase once the information was obtained and processed. Here, we successfully achieved the objectives of analyzing popularity metrics [5] and sentiments analysis [4]of different communities using Python libraries. Additionally, we also examine the popularity metrics studied.

- **Correlation of the popularity metrics and sentiment analysis**

  We have achieved the study's primary objective based on data correlation. After successfully analyzing the collected tweets, we have established a correlation between popularity metrics [5] and tweet sentiments [4].

- **Specific study of metrics in Community 8**.

  Finally, the last goal achieved was to study tweets and their correlation, specifically within the Spanish community in which our account was included.

6.4 Problems Faced

During the different phases of this project, we faced some problems and overcame the difficulties to achieve the goals mentioned above.

- **Learn about new technologies.**

  We faced an initial challenge due to our limited knowledge and experience with the technologies and libraries used in the project. As a result, the project's first phase was dedicated to acquiring a comprehensive understanding of these tools and learning how to utilize them effectively.

- **Limited querying and lack of data**

  The twitter API requests are limited, so we have not been able to store as much data as we would have liked per account. We have solved this lack of data by studying more accounts within the community.

6.5 Future work

Finally, in this section, we will explain some improvements that could be implemented in this project, as well as some future studies that can be carried out.

- **More Features**.

  Increasing the amount of information in the datasets always helps to improve the results, as there is more data to analyze.

- **Global Study**.

  A more comprehensive study could be conducted on the network by gathering nodes from other communities outside the Formula 1 community to analyze their metrics and sentiments. This way, we could obtain more accurate results and thus be able to generalize some of the conclusions we have received for this specific network.

- **Application**.

  Currently, numerous programs analyze your statistics on different social media platforms, but very few programs analyze your results based on competition. Furthermore, now, no programs combine your metrics on various social media platforms with the network map you are on. By having access to Twitter API or any other social media API, it would be possible to develop software that combines all the elements created in this study.

APPENDIX A

Impact of this project

This appendix presents an assessment of the potential social, economic, and environmental impact, as well as the ethical implications, of the project that has been carried out.

A.1 Social Impact

This section will discuss the social impact that this project could have. This project is based on the analysis of social networks exploring the correlation between the popularity and sentiment of the tweets collected. Since the results suggest that positive sentiment is correlated with better popularity, individuals may be motivated to prioritize positive expressions and attitudes in their online interactions. Recognizing the potential benefits of projecting positivity, people may become more aware of the impact their words and actions can have on their online reputation and engagement.

On the other hand, this type of study can help companies in sectors such as advertising or merchandising to know what to offer in each place or what account must be selected to carry out promotions. Knowing the community in which the account we are working with is located can be decisive in determining whether to carry out promotions or not. It also helps identify accounts that have higher engagement despite having fewer followers.

A.2 Economic Impact

This section deals with the possible economic impacts that may arise from this work.

This study could be beneficial for companies engaged in marketing campaigns as it could provide them with more precise metrics about their target audience and effective strategies to reach them, ultimately leading to increased popularity.

From the user's perspective, this study can help them better comprehend their metrics and gain a deeper understanding of the community with which they interact. This understanding can lead to more targeted and cost-effective campaigns, as they can tailor their marketing efforts to directly engage with their desired audience. By reducing costs through more focused campaigns, companies can optimize their resources and potentially improve their economic performance.

A.3 Environmental Impact

This section aims to identify the environmental impact associated with the realization of this project.

The development of our study requires essential equipment, such as computers, servers, and other computer materials, all of which consume energy during operation. This energy consumption can place a significant burden on our electricity networks and contribute to greenhouse gas emissions. Additionally, the cooling systems necessary to maintain optimal equipment operation also require energy.

A.4 Ethical Implications

The correlation study conducted on the collected tweet data, which examines the relationship between popularity metrics and sentiment analysis, raises several important ethical considerations. It is essential to address these implications to ensure responsible and ethical use of data to protect the rights and well-being of the individuals involved.

Researchers should be aware of any potential harm that could result from the findings. Responsible presentation of results should avoid stigmatization, discrimination, or any negative consequences. The aim is to ensure that the study contributes positively without causing harm.

On the other hand, another ethical consideration arises from the use of Twitter data for research purposes, as is the case in this project. Twitter's privacy policy states that users provide their consent to the collection, transfer, and storage of publicly available data. It is important to note that users have the option to modify their account's privacy settings according to their preferences.

APPENDIX B

Economic Budget

B.1 Introduction

B.2 Physical Resources

This section describes the estimated budget required for the hardware part of the project. The budget mentioned here is derived solely from the personal computer used for the entire project. The project did not require a large amount of resources to be carried out. However, having more RAM has helped to perform fast and smooth calculations to obtain the network graph. Here are the specifications of the computer used:

- **CPU:** Intel Core i5-10400f 2.90GHZ x4
- **RAM:** 16 GB
- **DISK:** 500 GB

The approximate cost of this computer is 900 €.

B.3 Human resources

To calculate the cost associated with Human Resources, we have considered that this work has lasted 400 hours and the required profile for tasks such as web scraping, programming, and data analysis is that of an engineer with a minimum annual salary of €24,000 Therefore, the cost associated with the Human Resources section of the project would be approximately €4,615.

B.4 Licenses

This section includes the cost corresponding to the licenses of the software tools necessary for the development and deployment of the system carried out in this project. However, all the software used in this project is open source, so the cost of the software licenses is zero.

B.5 Taxes

If this project is sold to a company interested in its acquisition, the sale must be subject to a tax of 15% of the price of the product.

Bibliography

[1] Abder-Rahman Ali. Presentando el natural language toolkit (nltk). *Tutsplus Tutorials*, May 2017.

[2] Ankthon. Python. sentiment analysis using vader. *GeeksForGeeks*, November 2022.

[3] We are social. The changing world of digital in 2023. *We are social*, January 2023.

[4] Sarvesh Bhatnagar and Nitin Choubey. Making sense of tweets using sentiment analysis on closely related topics. *Social Network Analysis and Mining*, 11(1):44, 2021.

[5] Jenn Chen. Twitter metrics: How & why you should track them. *SproutSocial*, September 2021.

[6] Graph Everywhere. Algoritmo de louvain. *Graph Everywhere*, 2022.

[7] Jim Frost. Skewed distribution: Definition and examples. *Statistics by Jim*, 2022.

[8] Martin Grandjean. Introduction to network analysis and visualization. *MartinGrandJean*, October 2015.

[9] Polovio Onofa. Liberia pandas y python. *IngeniusWorlds*, June 2021.

[10] Unir Revista. ¿qué es nlp y para qué sirve? *Unir Revista*, August 2021.

[11] Tomás Rneboldi. Cómo obtener datos de twitter. hacer todo esto con tweepy (3/5). *Medium Community*, September 2020.

[12] Twitter. Twitter api. *Twitter Platform*, 2.

[13] Twitter. Twitter developer platform. *Twitter*, September 2023.

[14] Wikipedia. Pearson correlation coefficient. *Wikipedia*, 2022.