TESINGIN-PT1.0.1 – Analysis & modeling of psychographic data.
– Thesis / Degree Project –[1]

Felipe Alfonso Gonzalez Lopez
Final degree project/thesis, Graduate Computer Science Engineer, IACC Chile., 2020-2023

Analysis & modeling of psychographic data on the internet using machine learning for organizations, private companies & governments.

Institute of Communications, Arts and Sciences, IACC
Av. Salvador 1318, Metro Santa Isabel, Providencia, Santiago.
Chile.
f.alfonso@res-ear.ch – felipe.alfonso.glz@gmail.com – https://twitter.com/felipealfonsog
https://glzengrg.com - https://freeshell.de/~felipe - https://linkedin.com/in/felipealfonsog

Name Teacher Guide: María Lourdes Geizzelez Luzardo
By: Felipe Alfonso González López.

Computer Science Engineer (Ingº en Informática), Graduated with maximum distinction (7.0 CL / A+ US)

Graduated / Titulado 2023.-

2020-2023

Keywords: Big Data, Data Analysis, Machine Learning, Software, Internet.

Index

Dedication

Dedicated to my dear Mother Gloria López Apablaza, who has always believed in me. To my Father Alfonso González Marquez, who was always an inspiration as well. To my brother Ponchi, to my Aunt Veronica López, my cousin Grisel B. López, and my dog, Aimy.

Abstract

This is a topic of great relevance and trend on the Internet today, since the operational continuity depends on this and sometimes also the strategic type of organizations and governments if there are no tools that allow the development of pieces of software for data analysis. ; which on certain occasions could lead to the point where a company or organization or even a political party can fall into an escalation of failures. These failures can be of a communication type with the same users who connect with the company, organization, etc. All this idea allows users to guide themselves to make certain determinations and decisions so that they influence society, as it is in "Internet society". This is called management and manipulation of masses on the Internet in a colloquial way, although it is rightly psychographic analysis, and thanks to data science, Big Data and Machine Learning or machine learning and also pieces of software that allow processing and deliver a complete analysis of psychographic and behavioral data on Internet users and/or "Sentiment Data", which may be on social networks such as Twitter (which will be used for this research as an example), or in any other form of Internet interaction.

1. The Problem

Today it is complex to find systems that allow managing and controlling a large amount of data and information that flows on the Internet and this is the problem that many organizations face because there are masses that move at high speed and in a flow very big. This large amount of data is usually something that becomes a great dilemma in organizations such as political parties or private companies or even governments. Offering pieces of software that analyze data, and that model all this information that is produced, and is generated in a very gigantic way, requires scaling this data appropriately and applying certain modeling, so that organizations or governments make good decisions.

1.1 Problem Statement

- So that?

To monitor, control and manage the flow of data in the masses on the Internet that may influence the product, image or idea of an organization or government.

- Why?

Why is it very necessary to understand how data works on the Internet today? Simple, this allows positive determinations and decisions to be made for an organization.

- Beneficiaries?

Governments, organizations and private companies in general benefit. It allows you to keep track of communication processes, ideas and their flows, between users and organizations.

- In what way?

Creating pieces of software that allow analysis through data science and machine learning; also all this is to speed up these processes. These pieces of software will allow you to deliver a general and complete analysis of what needs to be visualized.

- Social projection

This allows users to feel valued. It also helps to keep order in organizations and to optimize them, and to have total clarity regarding which determinations and decisions a social group needs to make.

- What does it solve?

Solves problems derived from the construction of ideas and communications towards certain psychosocial groups, user groups on the Internet.

- Which will allow?

Allow a quick and timely solution to problems where a quick solution is required to address certain issues in social groups, people, etc., who are involved in certain media or networks on the Internet.

1.1.1 General Objective

Develop pieces of software that make it possible to facilitate data compression on the Internet for large masses of users.

All this will facilitate the compression of data on the Internet relative to large groups of users, and also allow analysis and modeling in relation to the data science behind the analysis that will be carried out on groups and people on the Internet in general, all available (Users of social networks, etc.). All this in order to deliver precise ideas so that companies, organizations or governments take a defined and clear course of action to position their intentions.

1.1.2 Specific Objectives

Analyze and update the flowchart of the analysis process with the company or organization.
Build certain software for the client and/or pieces of software that work, for example, in a terminal on the server and that allow building statistical models.
Verify the operation of the scripts in the different terminals on the server.

- Analyze the feasibility of the data in relation to the pieces of software that are going to be created to generate a follow-up, and thus to also generate analyzes and comply with the standards that are required at the time.

- Study the technical and economic feasibility for the design of software and/or pieces of software that will eventually be running on a server 24/7 or every certain amount of time under a certain operation carried out by an engineer, and which will allow all requirements to be met. necessary for compliance with the organization.

- Develop software or pieces of software that allow both the company that develops the pieces of software and the client to monitor and monitor the

analysis that will be carried out through Bots, as well as the organization to which the service will be delivered for perform data analysis, etc.

1.1.3 Hypothesis

One of the most important resources on the Internet are users. People have allowed an area known as data science to develop over time, where large amounts of information are analyzed; This, in data engineering, makes it possible to precisely analyze the behavior of users and these become a very important resource, such as in a mine, it can be gold or copper. To perform all this analysis procedures and pieces of software are required. All these will be carried out by a company that delivers these services efficiently and ideally.

There are elements that are always generating changes in users, such as the structures and algorithms in the different social networks or the same algorithms that can be developed in the pieces of software that would allow generating a complete analysis of certain incidents that organizations need to handle. companies, governments and political parties, focusing on the masses on the Internet. When finding abnormal situations, referring to these as data that need to be refocused or redirected, these can be taken as positive situations since they will allow an analysis based on errors and with these it will be generated, by means of an algorithm using mathematics and statistics. , a better utility to generate reports and thus allow a need for a company or government to be resolved. Through these systems, it will be possible to carry out all these processes, allowing us to infer, if necessary, what the review, action and analysis would be, as well as its management or any execution necessary to successfully conclude any action taken by the private company, organization or government.

Definitely:
The fundamental premise is that the system, and throughout its implementation, will deliver data and models to make communication adjustments that affect certain users on the Internet. People or users have allowed a field known as data science to develop over time, where large amounts of information are stored and in turn allow analysis of these. In data engineering, this makes it possible to precisely analyze its behavior, and it becomes an important resource, as was said before rhetorically, like gold or copper in a mine.
To perform all these scans, procedures are needed. All done by a company that delivers it efficiently and adequately.

1.2 Rationale for the Research

A specific and accurate selected research methodology will allow gathering and obtaining the corresponding data for further processing. The result will provide information on the state of the data at the current time as an information system. Ideally, there should be a method through pieces of software that performs all these processes automatically using machine learning.
The main details that will be needed around the respective data analysis and the procedure to be carried out will be followed in a constant, this to give a solution to the problem raised, and that today is an emerging world, in our technological society.

The orientation of this research will focus on finding flaws in the methods how companies relate to their users. The organization is also involved in monitoring future processes. In addition to implementing these procedures and systems, a type of quantitative investigation would be carried out since the results of everything that would be analyzed in relation to the users who are involved with the organization and the nature of the descriptive investigations will be presented. since it describes in detail all the elements that make up the implemented system and that is exposed, within the organization.
In this sense, it is also suggested that in this case the focus of this research is quantitative since it is based on the study of the reality of these processes, quantifying the elements that intervene in the system that is to be implemented and providing a possible solution. On the other hand, this research is of the descriptive / explanatory type because, on the one hand, the dimensions of the problem posed are measured, describing what is being investigated and on the other hand, the phenomenon must be explained where a solution can be offered to the dynamics that are being investigated. raise.

1.2.1 Research Technique

In this project, the research technique will be of a non-experimental type since a study will be carried out, which will be applied to the data with the current problem of the organization, and the design of the system to be implemented, which will consist of observing the procedures currently available to solve the problem.

2. Theoretical Framework

Now, I will refer to the theoretical framework of the project.
The research shows that the application for data analysis must be developed on an operating system such as Mac or Linux, using terminals with Python, under an operating system ideally Linux (Debian), ideally in my view an operating system based on UNIX, since Windows is more robust under my research and analysis, all this since it needs the ability to extend to different parts of the country or the planet, for this references are taken, it is also ideal to have a database server such as MySQL, since ideally the data collected can be saved in it. MySQL has shown throughout its history also a lot of robustness, it is an efficient database engine.
References to these concepts are presented below:

Python [2]: Python is a high-level interpreted programming language whose philosophy emphasizes the readability of its code, it is used to develop applications of all kinds, examples: Instagram, Netflix, Spotify, Panda 3D, among others.2 It is about of a multi-paradigm programming language, since it partially supports object orientation, imperative programming and, to a lesser extent[which?], functional programming. It is an interpreted, dynamic and cross-platform language.
Managed by the Python Software Foundation, it holds an open source license, called the Python Software Foundation License.3 Python consistently ranks as one of the most popular programming languages.

[2]Description taken from Wikipedia - Reference: https://bit.ly/2GuKE4N

MySQL [3]: MySQL is a relational database management system developed under a dual license: General Public License/Commercial License by Oracle Corporation and is considered the world's most popular open source database,12 and one of the most popular in general along with Oracle and Microsoft SQL Server, all for web development environments.
MySQL was initially developed by MySQL AB (a company founded by David Axmark, Allan Larsson and Michael Widenius). MySQL AB was acquired by Sun Microsystems in 2008, and this in turn was purchased by Oracle Corporation in 2010, which had already owned Innobase Oy since 2005, a Finnish company that developed the InnoDB engine for MySQL.
Unlike projects like Apache, where the software is developed by a public community and the code copyright is held by the individual author, MySQL is sponsored by a private company, which owns the copyright to most of the code. This is what makes the previously mentioned dual licensing scheme possible. The database is distributed in several versions, a Community, distributed under the GNU General Public License, version 2, and several Enterprise versions, for those companies that want to incorporate it into proprietary products.
Enterprise versions include additional products or services such as monitoring tools and official technical support. A fork called MariaDB was created in 2009 by some developers (including some original MySQL developers) who were unhappy with the development model and the fact that the same company controls both MySQL and Oracle Database products.

It is developed for the most part in ANSI C and C++.4 It is traditionally considered one of the four components of the LAMP and WAMP development stack.
MySQL is used by many large and popular websites, including Wikipedia, Google (though not for search), Facebook, Twitter, Flickr, and YouTube.

Debian GNU/Linux [4]: Debian GNU/Linux is a free operating system, developed by thousands of volunteers from all over the world, who collaborate through the Internet.
Debian's dedication to free software, its volunteer base, its non-commercial nature, and its open development model set it apart from other GNU operating system distributions. All these aspects and more are collected in the so-called Debian Social Contract. Debian is characterized by not having the latest developments in GNU/Linux, but it does have the most stable operating system possible. This is achieved by means of old packages and libraries, but with many months of testing, ensuring maximum stability for each version that is released by the Debian community.
It was born in 1993, hand in hand with the Debian project, with the idea of creating a GNU system using Linux as the kernel. The Debian Project is the organization responsible for its maintenance today, and also develops GNU systems based on other kernels (Debian GNU/Hurd, Debian GNU/NetBSD, and Debian GNU/kFreeBSD).
One of its main objectives is to separate free software from non-free software into its versions. The development model is independent from companies, created by the users themselves, without depending in any way on commercial needs. Debian does not directly sell its software, but rather makes it available to anyone on the Internet, although it does allow individuals or companies to commercially distribute this software as long as its license is respected.
Debian GNU/Linux can use different installation mechanisms, such as: DVD, CD, USB, and even directly from the network (the latter depends on the speed of the user's network).

3. Methodological Framework

First, the development of a survey is not limited to the definition of the thematic content or its distribution, therefore, it is imperative to carry these contents through a methodology that is oriented towards the achievement of the proposed objectives.
In this research budget, the bases must be the detail of the procedures necessary to obtain the required information, thus being able to structure or solve the problem posed. The research process begins when one asks a question to which one does not know the answer and which must be answered. To do this, such research must be planned, which involves going through a set of preliminary phases, defining the type of study, and outlining the components. In the case of the project in question, the approach to solve the problem raised regarding the creation of pieces of software to monitor the respective models and analysis is first and foremost a "conversation with the users involved", of the workers and the management, to Being able to integrate necessary elements for computer development, then jointly carry out the visual design of the application if necessary, and then technically analyze the programming for its use in development, servers, communication systems, etc. This clearly defines the creation of all development.

3.1 Characteristics of the System to be Developed During the Investigation

They are pieces of software that focus on data analysis, these pieces of software can either work under a terminal, being executed on a Mac OSX server and fundamentally on Debian GNU/Linux.

3.1.1 Methodologies and Fundamentals of Development Raised

The iterative model of the cascading life cycle will be used in the project. This model is intended to reduce the risk that user needs and the final product may arise from misunderstandings during the requirements gathering phase.

It consists of the iteration of several cascading life cycles. At the end of each iteration, the client is delivered an improved or more functional version of any software product offered. In this case, as I have already mentioned before, it corresponds to pieces of software that will work in Terminals of a Debian GNU/Linux server.
The client is the one who, after each iteration, evaluates the product and corrects it or suggests improvements, regarding the delivered results, not the performance that can be

[3]Description taken from Wikipedia - Reference: https://bit.ly/3eh70IK

[4]Description taken from Wikipedia - Reference: https://bit.ly/3RwbF80

delivered itself, but based on certain pieces of software produced.

These iterations will be repeated until obtaining pieces that satisfy any type of need, but above all to proprietary software pieces, since the source code of each of the specific pieces of software is not delivered for this case, it is only delivered as product, the results they will deliver.
Specifically, this methodology gives us the ideal way to work on our system since each piece of software itself is a version in itself. Such as, for example, the production of the GNU/Linux Kernel and its constant development.

This model is generally used in projects where the requirements are not clear to the user, so it is necessary to create different prototypes to present them and obtain the compliance of the client or an investigation. One of the main advantages of this model is that the requirements do not have to be fully defined during development, but can be refined in iterations.

Like other similar models, it has the advantage of carrying out developments in small cycles, which allows better risks, better management of the deliveries of each version in the software pieces.

3.1.2 Type of Investigation

The choice of a specific and precise research methodology will allow collecting and obtaining the corresponding data for subsequent analysis.
The result will provide the information about the state of the data at the current time as system information, ideally there is a method through software that automatically runs all processes using machine learning.
The main details that will be necessary will be around the respective data and the procedure to be followed constantly, this to solve the problem of handling large amounts of data, in what is now an emerging world in our society.
The orientation of this research will focus on research on the relationship methods of companies with their users and research.
The organization is also involved in monitoring processes
In this sense, it is also necessary to specify that in this case, the object of this investigation is quantitative, since it is based on the reality of these processes, quantifying the elements that intervene in the system to be implemented and a possible solution.
On the other hand, this research is of the descriptive-explanatory type because, on the one hand, they measure the dimensions of the problem posed, describing who is being investigated and, on the other hand, the phenomenon must explain where a solution can be offered to the dynamics that arises.

3.1.3 Data Collection

Data collection refers to the systematic approach of collecting and measuring information from various sources to obtain a complete picture in different areas of interest.
Data collection allows an individual or business to answer relevant questions, evaluate outcomes, and better anticipate probabilities and futures.
Accuracy of data collection is essential to ensure study integrity, sound business decisions, and quality.

For example, data may be collected through mobile devices, website visits, social media responses, loyalty programs, and online surveys to learn more about customers. There are different collection methods and techniques that can be useful.
The choice of method depends on the strategy, the type of variable, the desired precision, the collection point, and the skills of the interviewer. Interviews are one of the most common methods. If it is decided to do so, special attention is required in the questions that will be asked, which can also be in a face-to-face interview, by telephone and even by email or via social networks.

- Questionnaires are a useful tool for data collection.
- To get the expected results, they must be carried out carefully.
- This is why before writing it, it is important for the researcher to define the objectives of the research.
- There are two questionnaire formats: questionnaires that are applied when you want to know the opinion of your experiences and feelings on a specific topic.
- On the contrary, in the closed questionnaire, the researchers control what they ask and what they want to know, which may force them to limit the responses of the participants.
- If we prefer on-site observation to get to know your customers, they can be done according to other methodologies.
- Taking into account how the information is found, it will be of great help when analyzing it.
- Being able to measure and report accurate and real data is also very important for good decision making.

3.1.4 Data Collection Instrument Used

Questionnaire with the following questions:
Tell us about your impression and how you feel about the organization.
The first question in the customer or user questionnaire includes all the basic information about them, which consists of:

- *Name of the client or company or users - people.*

Important data of the person (A complete psychographic investigation of the user is carried out, where he lives, what he says, how he says it, etc.)
Contact details, usernames, etc.

- What are your target customers?

This should include all applicable demographics, sentiments expressed, images posted, etc. Understanding the client's audience type will give an idea of what modeling elements to use.

- Do you have a system currently?

You'll need to assess whether there are already any systems that might be functional with the goals to be accomplished in the organization, compare them to the goals of the business to see if they need an overhaul or a complete rebuild.

- What requires more attention?
- The customer's response to this question will help you understand what is most important and what will matter most to them. Defining the purpose, understanding your current weaknesses, and creating a detailed checklist of the direction you will be heading in will help you build a solid foundation for a successful project.
- Why is your organization required systems like the one that will be offered and integrated?

Like the previous question in the client questionnaire, this question helps you understand the weaknesses of your ideas and see what isn't working for the client. It will help you understand the purpose of the new goals. It could be that it just needs to add a different new feature, or it could need to be built on a new platform with different features.

- What characteristics will the ideas have that will be promoted in the users?

This answer should be as detailed as possible. Features include:

- Feelings of closeness towards our ideas
- Influencing specific social groups
- Accumulate data and models on the types of users to influence.

Again, the client or organization is encouraged to consider their audience and the goals of their ideas, when creating the list of necessary features to focus on.

What similar ideas call your attention or what is the group to fight socially on the Internet?
This will show ideas the customer likes, and provide examples of features that might be difficult to describe. It can be especially helpful for the customer to point out the characteristics of your 'competition'.

What is the deadline to start delivering statistical data and models that mean something important for the Bots to start their data collection and analysis task?
This information can be used in the customer questionnaire to determine if and how customer needs can be met before the deadline. A timeline may need to be provided to show what can be done before the deadline and what can be added later and when.
This questionnaire will be applied in meetings with the heads of the organization and departments involved.

3.1.5 Population and Sample

Population refers to the universe, set or totality of elements on which research is carried out or studies are carried out. Sample is a part or subset of elements that are previously selected from a population to carry out a study.
For our research, the population is made up of Internet users, involved in certain ideas that an organization or company represents, which are:

- Users who express ideas of support for the organization or government.
- Users who express ideas of rejection of the organization or government.
- Potential users to influence and attract the organization.

The sample that we will use will be the different types of users on the Internet.

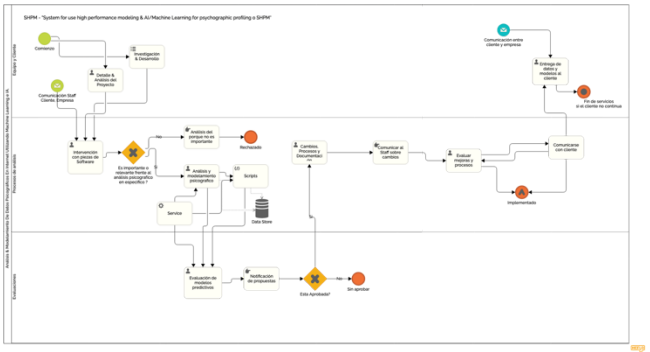
4. Analysis of the Results

Firstly, the technique that will be used for measurement is quantitative observation.
The object of study is: The amount of data to be analyzed in a given amount of time.

For this, a survey will be taken of the workers and additional data will be taken from the headquarters in charge of closing these surveys to collect data either before or after the implementation of the respective services to solve the problem posed.
All of the above is to achieve a complete daily, monthly and annual analysis. In case of presenting any inconvenience, it can be solved immediately since all the corresponding backup systems will be available, in such a way that the registration is efficient and consistent.

4.1 Flowchart of the proposed investigation

Figure 1
Flowchart, for the proposed project.



Note. It is possible to have a complete vision from the following link: https://bit.ly/3EuAL3a

5. Explanation of the flowchart of the proposed investigation

The explanation to the previous flowchart is essentially to give the sample of the beginning of the research processes or of projects and processes, initially based on meetings with the client, about details and analysis of the future project itself. All also in constant communication by email with the team and the client, which will be taken as a resource of vital importance for the analysis and/or evaluation processes, as appropriate. It is important to mention that, constantly, whether there are clients or not, research and development can be carried out on the issues raised.

In the analysis processes section, we will already be facing pieces of software that will be proprietary, and will work under terminals on a server where there will be a database that will ideally also store information regarding statistical data, analysis, etc. This itself is a set of scripts that will make up a service as a whole, specifically.

Then the analyzed will be evaluated unless it has been rejected, the respective analysis of why it was rejected has been carried out. If it was not rejected, we continue with the evaluation of the predictive models; Notifications are generated, which, if approved, go through changes and improvements in their processes and a respective documentation is carried out. All under full communication with the staff, regarding the evaluations of improvements in the versions of the services. Already implemented, with changes and improvements, all the scripts are working, and the client will be informed about the subsequent analysis of these. If the client does not wish to continue with the services, an end of these is determined. But research and development processes can be maintained under the same flow.

5.1 Model, Functionality and Scripts

What this flow will do in its environment, such as pieces of software of different types, is to analyze the psychographic segmentation of Internet users, based on concepts that may be being used, for example, in the social network Twitter (At this stage we plan which segments of users will be within our Target), in which case we are already using a piece of software that generates data around what can be the 'Sentiment Data' of users (here, as indicated, a life cycle of a information system, the analysis process is important).

To analyze results we rely on an algorithm that can determine how positive or negative there are in a generality 'X', compared to certain concepts, ideas or people, etc. These values can anticipate future changes, such as those shown in this graph (This is used to determine in the life cycle of an information system, a correct approach in the design and development stage):

Figure 2
Scheme on Segmentation and Psychographic Factors.

Note. Factors proposed by Consunt. Adapted from "Factores-Psicograficos.gif" [Image], by https://bit.ly/3X0yynk

We also rely on the construction of popularity and subjectivity in certain messages, and through an algorithm, and they are classified to determine within an expression whether or not they are - Bots. That allows you to have a larger filter. It can be of great importance for our client or organization, since in this way they can rely on certain messages, to approach an 'A' or 'B' population and determine their values, likewise the company or our client will make better decisions to get closer to certain mass feelings on the Internet.

Like any piece of software, it is subject to errors, even when we already use Open Source libraries in Python such as TextBlob, among others to streamline processes (At this stage, the scripts that determine certain data, as in this example, go to an integration stage, QA or Testing processes).

Below are some portions of the code from this project, and its results in a terminal, using VSCode, in a Mac OSX environment & under Testing on a Debian GNU/Linux server. (We'll use the 'Virgin Airlines' variable for this case, as an example.)

In this case, our piece of software is in a stage or maintenance phase within the life cycle of an information system.

Figure 3

Called bookstores.

```
import tweepy
from textblob import TextBlob
```

Figure 4

Authorization accepted by the [5]Twitter API.

```
auth = tweepy.OAuthHandler(consumer_key, consumer_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
public_tweets = api.search('virgin airlines')

for tweet in public_tweets:
        print(tweet.text)
        analysis = TextBlob(tweet.text)
        print(analysis.sentiment)
        if analysis.sentiment[0]>0:
                print ('Positivo')
        else:
                print ('Negativo')
        print("")
```

Figure 5 - 6 & 7

Flows of cycles in programming in Python, for evaluations of "Sentiment Data" in Twitter.

```
tweet_list.append(tweet.text)
analysis = TextBlob(tweet.text)
score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
neg = score['neg']
neu = score['neu']
pos = score['pos']
comp = score['compound']
polarity += analysis.sentiment.polarity
```

Figure 6

```
if neg > pos:
    negative_list.append(tweet.text)
    negative += 1

elif pos > neg:
    positive_list.append(tweet.text)
    positive += 1

elif pos == neg:
    neutral_list.append(tweet.text)
    neutral += 1
```

Figure 7

```
positive = percentage(positive, noOfTweet)
negative = percentage(negative, noOfTweet)
neutral = percentage(neutral, noOfTweet)
polarity = percentage(polarity, noOfTweet)
positive = format(positive, '.1f')
negative = format(negative, '.1f')
neutral = format(neutral, '.1f')
```

Figure 8 – 9 – 10 & 11

Examples of the type of output we get in a terminal.

```
CEO of this airlines be sacked now !! How do u expect if other airlines r no allowed to operate in Austr
alia?? So o… https://t.co/5TkxKnNVy1
Sentiment(polarity=-0.125, subjectivity=0.375)
Negativo

RT @stats_feed: World's best economy class airlines in 2022:

AE Emirates
QA Qatar Airways
SG Singapore Airlines
JP ANA All Nippon Airways…
Sentiment(polarity=1.0, subjectivity=0.3)
Positivo
```

Figure 9

```
RT @stats_feed: World's best economy class airlines in 2022:

AE Emirates
QA Qatar Airways
SG Singapore Airlines
JP ANA All Nippon Airways…
Sentiment(polarity=1.0, subjectivity=0.3)
Positivo

RT @RobertCawood2: Where is little Alan, the female pilot who pioneered gender-equality for Qantas is no
w suing the airline for sexual hara…
Sentiment(polarity=0.10416666666666667, subjectivity=0.5)
Positivo
```

Figure 10

```
RT @RobertCawood2: Where is little Alan, the female pilot who pioneered gender-equality for Qantas is no
w suing the airline for sexual hara…
Sentiment(polarity=0.10416666666666667, subjectivity=0.5)
Positivo

RT @stats_feed: World's best economy class airlines in 2022:

AE Emirates
QA Qatar Airways
SG Singapore Airlines
JP ANA All Nippon Airways…
Sentiment(polarity=1.0, subjectivity=0.3)
Positivo
```

Figure 11

```
RT @RobertCawood2: Where is little Alan, the female pilot who pioneered gender-equality for Qantas is no
w suing the airline for sexual hara…
Sentiment(polarity=0.10416666666666667, subjectivity=0.5)
Positivo

Where is little Alan, the female pilot who pioneered gender-equality for Qantas is now suing the airline
 for sexual… https://t.co/3lQ9XJuu02
Sentiment(polarity=-0.09375, subjectivity=0.3333333333333333)
Negativo

RT @DylanDunlevy: @AndyBopinion Truth is we need new airlines.
Qantas national joke, Jetstar always a joke and Virgin flat broke
Sentiment(polarity=0.05568181818181818, subjectivity=0.2897727272727273)
Positivo
```

6. Coding Process, Requirements, Configuration & Tests[6]

The Python programming language is used for the development of the project, whose specifications in any mode have already been detailed above. Python is one of the most popular languages in the world and, in fact, the Netflix recommendation algorithm and the one that controls self-driving cars were created.

In terms of scalability, Python has an advantage over programming languages like R in that it offers a different approach to solving different problems.
In terms of speed, Python also stands out between Matlab and Stata.
Some of the important features of Python are:

- The syntax is quite easy to use and therefore who can learn python in less time.
- Python is also a versatile and easy-to-use language.
- In terms of scalability, Python has an advantage over programming languages like R in that it offers one approach to solving different problems.
- In terms of speed, Python also stands out between Matlab and Stata.
- It has a great library.
- A library or a library is a set of which are linked together.
- It can be used over and over again for programs.

[5]Application programming interface. An API represents the communication capability between software components. Reference: https://bit.ly/3O3ZoqA

[6]It is important to note that, for these cases, we have an account authorized by Twitter, for development.

- It has a very strong community that helps update libraries and frameworks.
- Libraries and frameworks are downloadable and free.

Python is an interpreted programming language, that is, it is first converted to bytecode containing low-level instructions, and then executed by the Python interpreter.
It is cross-platform, which means that once written in Python, it can run on any operating system: Windows, Mac, Linux, etc.

6.1 Categorical Coding

Categorical coding is a technique for coding categorical data.
It's good to keep in mind that categorical data are sets that contain variable labels instead of values. Many machine learning algorithms cannot handle categorical variables. Therefore, it is important to code the data properly to be able to preprocess these variables.
Since you need to fit and evaluate, you need to encode the categorical data and convert any input and output variables to numeric values.
In this way, the model will be able to understand and extract information, generating the desired result.
Categorical data varies based on the number of possible values.
Most categorical variables are nominal.
These variables help categorize and tag attributes.

6.1.1 Most Used Library

Another frontline need in psychographic analysis is to generate visualizations. In this sense, it is impossible to avoid the presence of Matplotlib and Seaborn. Both libraries are widely used for data science, with Matplotlib being the oldest and most popular, and Seaborn being a burgeoning package that is based precisely on the Matplotlib code. Therefore, the use of both libraries is a relevant synergy for data science.

6.1.2 Visual Studio Code for Programming in Python

Visual Studio Code or better known as VSCode is a Microsoft source code editor that can be used on Windows, as well as macOS and Linux. Also, it is an open source editor that is available on GitHub.
It has very interesting features for code development such as syntax highlighting and autocompletion, integration with the Git version control system, and debugging from the editor itself. As with other editors, such as Atom or Sublime, it also supports the ability to install third-party builds that add additional functionality. Installing extensions or add-ons in VSCode is as simple as clicking the corresponding button in the menu or accessing them directly with the Ctrl+Shift+X keyboard shortcut.
This opens a new section on the left side of the program that contains a search engine, so you can search for extensions by name, and a list grouped into three categories: installed, popular, and recommended.

6.1.3 Python And Testing

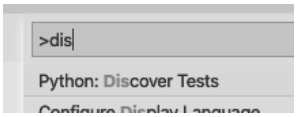
This 'Python' extension, developed by Microsoft, adds many Python features to VSCode, such as autocompletion and code formatting, as mentioned, debugging tools, our own Python code and environment management, among others.
Another consideration to take into account is to activate the test that I will use, to choose between Unittest, Pytest or Nose.
We can easily do this by accessing the configuration by typing Python Testing in the search bar. In my case, as I use Unittest, I mark the corresponding box to activate it or directly in the JSON configuration of VSCode. For this case, the test of a piece of code related to user login to a micro-system in a Terminal will be used.

6.1.4 Running And Debugging Tests

Enable unit testing in VScode.
To enable Unittest in VSCode, we run the Discover test command:

Figure 12

Add-on search in VSCode to find a Testing Add-on.

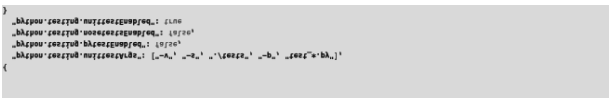


This will prompt us to configure the testing framework used, Pytest or nosetests, Unittest in this example I am using Unittest.

Once configured, the .vscode/settings.json configuration file will be updated like this:

Figure 13

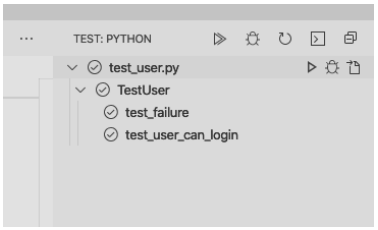
Editing JSON data in the VSCode configuration.



This configuration defines the Unittest arguments used and defines which Python testing framework is used. The tests are located in the './tests' folder and the template used for the test files is test_*. Once we have Unittest and it is configured with VSCode, we can run the test from the Testing tab.

Figure 14

Example in Testing, from VSCode, passing the tests as an example.



Or we can click on the top of each test or where we should now find the option to perform a Run Test | Debug Test.

Figure 15

Functions that will serve as an easy example, to understand how Unittest works.

```
from unittest import TestCase

✓ Run Test | ✓ Debug Test
class TestUser(TestCase):
    """User Test Case"""

    ✓ Run Test | ✓ Debug Test
    def test_failure(self):
        """Example of test failure."""
        self.assertEqual(1, 1)

    ✓ Run Test | ✓ Debug Test
    def test_user_can_login(self):
        """Test that the user can login."""
        self.assertTrue(True)
```

Using Debug Test now, it is possible to easily perform Quick Testing and Debugging on the code, as I mentioned before, here it was done with the example of testing the code of a script for a login-user micro-system in a Terminal.

Figure 16

In the following code we highlight the line where it was performed - Debug Test.

```
        self.assertEqual(1, 1)

    Run Test | Debug Test
    def test_user_can_login(self):
        """Test that the user can login."""
        self.assertTrue(True)

```

7. Code Listing, Scripts

Next, I will list the central sequences of one of the scripts or piece of software, where a psychographic analysis of positive, neutral or negative feelings of a search is also carried out and of a quantity of data to be analyzed depending on how the piece is interacted with. of software that runs in a Terminal, and that generates two graphs directly with Python, in addition to delivering, as in the script mentioned here and demonstrated in a few minutes, certain and specific data. The graph to be evaluated is of the circular type, which determines and shows data in percentages. And the following one is rather of practical use to display analyzed data in a Cloud form. These operations are performed with the Pythom library: Matplotlib, WordCloud, Nltk, among others.

In this case, the concept of extreme sports will be analyzed, within Chile – 'Extreme Sports'.

Figure 17 & 18

Again, the exercise of using the necessary libraries for our exercise is carried out.

```
# Import -Libraries-

from textblob import TextBlob
import sys
import tweepy
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import nltk
import pycountry
import re
import string
import mysql.connector as mysql # conector for the MySQL connection
```

Figure 18

```
from wordcloud import WordCloud, STOPWORDS
from PIL import Image
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from langdetect import detect
from nltk.stem import SnowballStemmer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer
from matplotlib import image as mpimg
```

Figure 19

Integration to the Twitter API.

```
# Auth:
consumerKey =
consumerSecret =
accessToken =
accessTokenSecret =

auth = tweepy.OAuthHandler(consumerKey, consumerSecret)
auth.set_access_token(accessToken, accessTokenSecret)
api = tweepy.API(auth)
```

Figure 20

We request the necessary variables.

```
#Analysis

def percentage(part,whole):
    return 100 * float(part)/float(whole)

keyword = input("INGRESAR UNA KEYWORD OR HASHTAG PARA BUSCAR: ")
noOfTweet = int(input ("INGRESAR UNA CANTIDAD DE TWEETS PARA ANALIZAR: "))


tweets = tweepy.Cursor(api.search, q=keyword).items(noOfTweet)
positive  = 0
negative = 0
neutral = 0
polarity = 0
tweet_list = []
neutral_list = []
negative_list = []
positive_list = []
```

Figure 21 -22 & 23

Using the analysis library functions, we perform comparisons.

```
for tweet in tweets:
    #print(tweet.text)
    tweet_list.append(tweet.text)
    analysis = TextBlob(tweet.text)
    score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    comp = score['compound']
    polarity += analysis.sentiment.polarity
```

Figure 22

```
    if neg > pos:
        negative_list.append(tweet.text)
        negative += 1

    elif pos > neg:
        positive_list.append(tweet.text)
        positive += 1

    elif pos == neg:
        neutral_list.append(tweet.text)
        neutral += 1
```

Figure 23

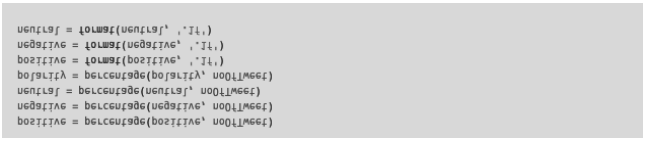


Figure 24

We send to the Terminal, the data analyzed as a result.

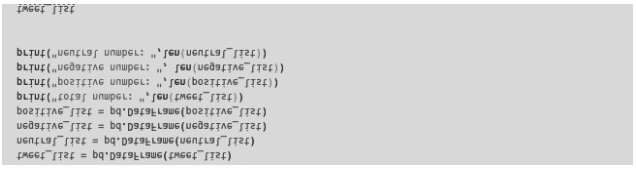


Figure 25

We begin to generate the respective pie chart and it will contain the analysis in percentages.

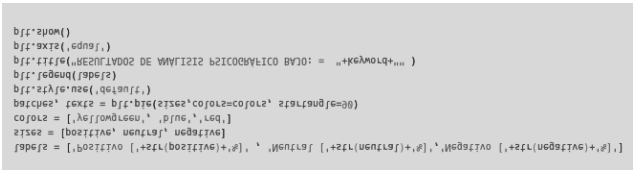
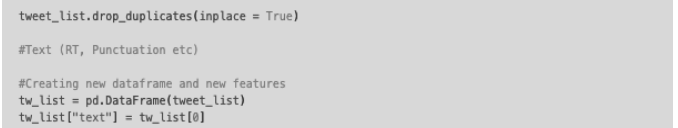


Figure 26

We check for duplicates.



```
tweet_list.drop_duplicates(inplace = True)

#Text (RT, Punctuation etc)

#Creating new dataframe and new features
tw_list = pd.DataFrame(tweet_list)
tw_list["text"] = tw_list[0]
```

Figure 27

We remove certain types of characters that may be uncomfortable in the face of analysis.

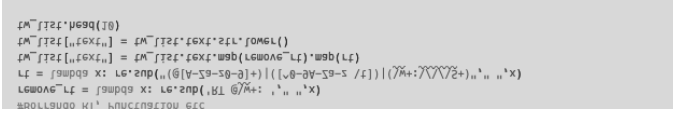


Figure 28

We begin to analyze the polarities of psychographic analysis.

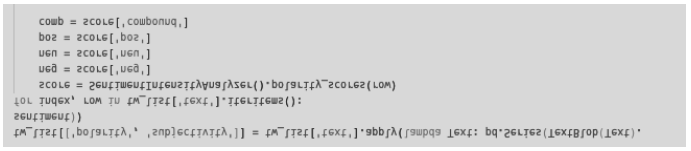
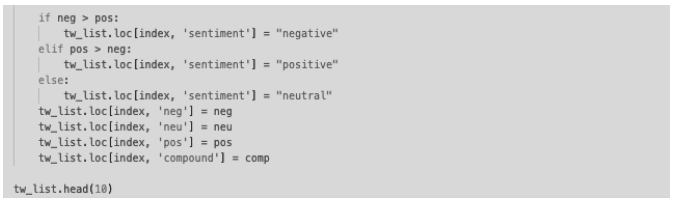


Figure 29

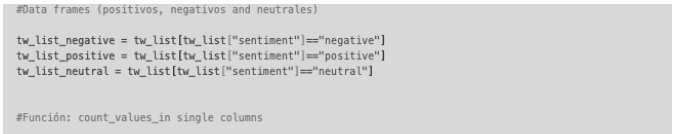
Comparisons are made against whether such a result is higher or lower, as an example.



```
    if neg > pos:
        tw_list.loc[index, 'sentiment'] = "negative"
    elif pos > neg:
        tw_list.loc[index, 'sentiment'] = "positive"
    else:
        tw_list.loc[index, 'sentiment'] = "neutral"
    tw_list.loc[index, 'neg'] = neg
    tw_list.loc[index, 'neu'] = neu
    tw_list.loc[index, 'pos'] = pos
    tw_list.loc[index, 'compound'] = comp

tw_list.head(10)
```

Figure 30

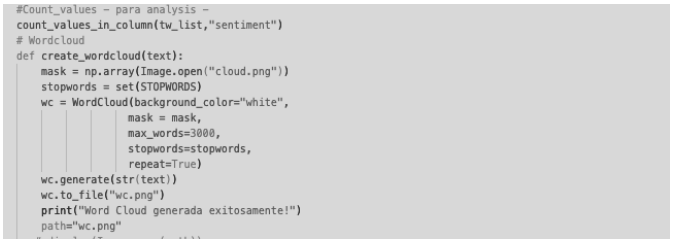
We count, and generate totals and percentages.



```
#Data frames (positivos, negativos and neutrales)

tw_list_negative = tw_list[tw_list["sentiment"]=="negative"]
tw_list_positive = tw_list[tw_list["sentiment"]=="positive"]
tw_list_neutral = tw_list[tw_list["sentiment"]=="neutral"]


#Función: count_values_in single columns

def count_values_in_column(data,feature):
    total=data.loc[:,feature].value_counts(dropna=False)
    percentage=round(data.loc[:,feature].value_counts(dropna=False,normalize=True)*100,2)
    return pd.concat([total,percentage],axis=1,keys=['Total','Percentage'])
```

Figure 31

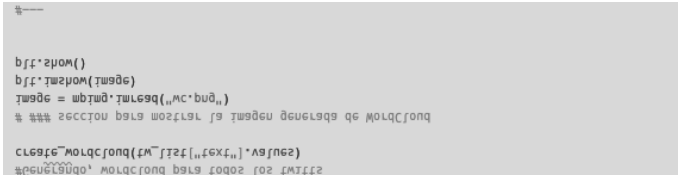
Using the above analysis, we created a graphical 'Cloud' of the most frequently used words.



```
#Count_values - para analysis -
count_values_in_column(tw_list,"sentiment")
# Wordcloud
def create_wordcloud(text):
    mask = np.array(Image.open("cloud.png"))
    stopwords = set(STOPWORDS)
    wc = WordCloud(background_color="white",
                   mask = mask,
                   max_words=3000,
                   stopwords=stopwords,
                   repeat=True)
    wc.generate(str(text))
    wc.to_file("wc.png")
    print("Word Cloud generada exitosamente!")
    path="wc.png"
```

Figure 32

We overwrite the image on wc.png and we can generate a new image and display it.



7.1 Running Pieces of Software (Scripts)

The execution of the script is directly using VSCode.

Figure 33

Using VSCode we carry out the execution of the code, in the Terminal.

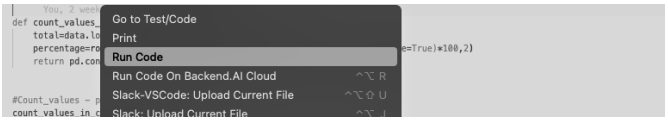


Figure 34

We obtain results according to variables delivered to the microsystem, in the Terminal itself.

Figure 35

The following figure shows the type of pie chart that can be generated, using the above method.

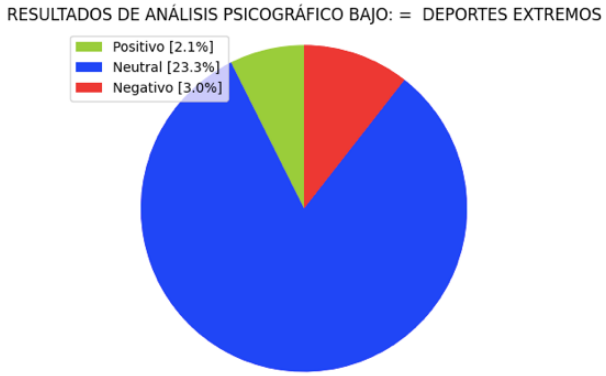


Figure 36

It delivers us through the Terminal a message that a 'Cloud' image has been successfully generated with Python.

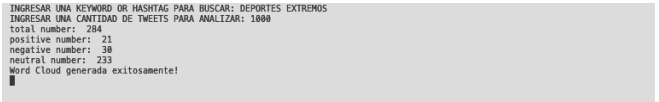


Figure 37

The output is as follows, as a graph:

Note. It is an automatically generated graph using Python. I would like to mention that with NodeJS [7], it is possible to create a mixed system, with Python and NodeJS to display this type of content, with a structure in another system that maintains this data and displays it online.

8. System Tests, Explanation, Exit Criteria and Results

At a high level, software testing is needed to detect bugs in the software and to prove that the software meets the requirements to get the necessary results back to the customer.
This helps the development team to fix bugs to deliver a good quality product.
There are several points in the development process where human error can lead to the software not meeting the necessary requirements. Fundamentally, performance, function, and stress tests will be generated. We will verify that the data provided by the graphs and outputs are correct, in terms of the algorithm used.

Dynamic tests will be carried out, because the pieces and/or piece of software that will be used with the objective of delivering data analysis, is a type of software where the input and output behavior will be verified to be correct, obtaining the expected data. . The functions of the software part(s) are operational. Unit tests will be performed for each piece or pieces of software that will deliver specific data. Then carrying out system tests, checking its load, operability, security, stability, etc. All this because they are fundamentally scripts, or pieces of software/code that will be executed in a Terminal.

For the project, a type of dynamic test will be used, as mentioned, and in this case, of each one, since the main objective of the pieces of software are the outputs or outputs of what is processed (In this case, external data), each piece of software would be considered as independent.

The results obtained during the test are precise since they are extracted from specific means, for sentient

[7]Node.js is a cross-platform, open source, server layer (but not limited to) runtime environment based on the JavaScript programming language, asynchronous, with data I/O in an event-driven architecture. and based on Google's V8 engine. Reference: https://bit.ly/3g2Jpg8 (Wikipedia).

and/or psychographic study analysis, therefore, the results or outputs are at the same time precise. Without further considering that, if there was a bug in the source code, you would not be able to execute the script at all in the Terminal.

Implementing a test service from scratch is a complex and time-consuming task.
In contrasting projects, we see that small, but efficient and relentless steps have been taken towards continuous integration QA service. Steps such as hiring specialized people in the field, implementing tools such as Testlink, test management, SonarQube to assess code quality, Jenkins for continuous integration, or Selenium for testing. Horizons such as Big Data Testing appear in the future of Testing, so the future of Testing is guaranteed.

8.1 Big Data Testing

In Big Data tests, QA engineers verify the successful processing of terabytes of data using one and other supporting components. This requires a high level of testing skills, processing is very fast, and can be of the following types:

- Bach
- Real time
- Interactive

Big Data tests can be divided into three.

Step 1: Validation Stage The Big Data testing stage, also known as the pre-Hadoop stage, involves validation of data from various sources such as relational data sources, blogs, social media, etc., need to be validated to ensure that the correct data is in the system. The source data is compared to the data entered into the Hadoop system to ensure a match. It verifies that the correct data is extracted and in the correct location.
Step 2 - Validation of "MapReduce": In this step, you check the validation of the business logic on each and then validate them after it runs on multiple nodes, ensuring that:

- The MapReduce process works fine.
- The rules of aggregation or segregation of data are in the data.
- “Key Value pairs” can be generated.
- The data is validated after the MapReduce process.

Step 3 - Phase of validation of the results: third and last stage of the Big Data Testing tests, is the process of validation of the results.
Output data files are generated and moved to an enterprise Data Warehouse or any other system based on these requirements.
The activities in this third stage include the following:

- It is verified that the transformation rules are applied correctly.
- Data integrity is verified and successful download to the target system.
- Data is checked for corruption by comparing the target data with the data in the HDFS file system.

8.1.1 Hadoop

Hadoop handles very large volumes of data and a large amount of resources. Therefore, architecture tests are crucial to guarantee the success of a Big Data project like the one presented. A poorly designed or inadequate system may result in poor performance and the system may not meet the requirements.

- The test services must be running on a Hadoop.
- Performance tests include tests for job completion, memory usage, data throughput, and similar system metrics.

9. Testing And Analysis Of Results Ii

To understand what I will detail below, and which directly refers to the tests and analysis in part two of the explanation, it is necessary that you can imply certain classification models.

9.1 Classifiers & Model Examples

Naive Bayes [8]is a simple model that I consider important to mention, among others that exist and can be used for data analysis and classification purposes; this can be used for data classification. In the model, the class ĉ is assigned to a tweet t, as seen in the following formula:

Figure 38

Naive Bayes' text classification model.

$$\hat{c} = \underset{c}{argmax}\ \mathrm{P}(\mathrm{c}|\mathrm{t})$$

$$\mathrm{P}(\mathrm{c}|\mathrm{t}) \propto \mathrm{P}(\mathrm{c}) \prod_{i=1}^{n} \mathrm{P}(f_i|\mathrm{c})$$

Note. Adapted from "twitter_sentimental_analysis_5.jpg" [Image], by pantechsolutions.net - Twitter Sentiment Analysis using Machine Learning on Python. https://bit.ly/3hDjMmy

In the above formula, fi represents the i-th feature out of the total of n features. P(c) and P(fi |c) can be obtained by maximum likelihood estimates.

[8]Explanation about the Naive Bayes model here: https://bit.ly/3UU6eRO

9.1.1 Other Models

maximum entropy
The Maximum Entropy Classifier model is based on the Maximum Entropy Principle. The main idea behind this is to choose the most uniform probabilistic model that maximizes entropy, with given constraints. Unlike Naive Bayes, it does not assume that features are conditionally independent of each other. Thus, we can add features as bigrams without worrying about feature overlap.

Figure 39

Model / Equation of Maximum Entropy. In a binary classification problem like the one we're tackling, it's the same as using Logistic Regression to find a distribution over classes. The model is represented by the following formula:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

Note. Adapted from "Twitter_Sentimental_Analysis_6.jpg" [Image], by pantechsolutions.net - Twitter Sentiment Analysis using Machine Learning on Python. https://bit.ly/3O3Asj1

Random Forest. (Used in what is presented)
Random Forest is a joint learning algorithm for classification and regression data. Random Forest generates a multitude of decision tree classifications based on the aggregate decision of those trees. For a set of tweets x 1 , x 2 , . . . xn and their respective opinion labels y 1 , y 2 , . . . n 'bagging', repeatedly selects a random sample (X b , Y b) with replacement. Each classification tree fb is trained using a different random sample (X b , Y b) where b varies from 1 . . . Finally, a majority vote of the predictions of these B-trees is taken.

9.1.2 Predictive Equation

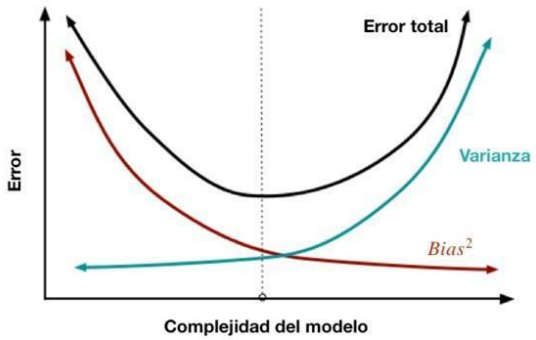
Figure 40

Equation of the predictive model. It is the training algorithm for random 'forests or trees' that applies the general technique of bootstrap aggregating, or bagging, for automatic learning, or machine learning as a tree type.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Note. Adapted from image from Wikipedia, which is not obtainable as -name-, [Image]. From https://bit.ly/3tpg6qX & https://bit.ly/3hGNasd
Figure 41

Complexity of the predictive model. An optimal balance of bias and variance would never overfit or be inappropriate for the model. Therefore, understanding bias and variance is critical to understanding the behavior of prediction models.



Note. Adapted from *"46138669365_b98531b89d_b.jpg" [Image] by https://flickr.com at https://bit.ly/3O0Pm9Q. Bias and Variance in Machine Learning. https://bit.ly/3UVfRzY*

Furthermore, an estimate of the prediction uncertainty can be made as the standard deviation of the predictions of all the individual regression trees at x':

Figure 42

predictive model, general technique of bootstrap aggregating, or bagging, for automatic learning, or machine learning as a tree type.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

Note. Adapted from Adapted from Wikipedia image, which is not obtainable as -name-, [Image]. From Wikipedia at https://bit.ly/3g26QGk.

It is good to mention that there are other models as well, such as the SVM, or super vector machines. The XGBoost, Xgboost is a form of incremental gradient algorithm, which produces a prediction model that is a set of weaker prediction decision trees.

9.1.3 Neural Networks

MLP or Multilayer Perceptron is a class of advanced neural networks, having at least three layers of neurons. Each neuron uses a nonlinear activation function and learns with supervision using a backpropagation algorithm. It works well for complex classification problems, such as sentiment analysis by learning nonlinear models.

9.1.4 Analysis

In this case, through the following piece of software, and as an example, the 'public sentiment' tweets relating to '6 US airlines' will be used as a concept to analyze.
I classified the tweets into their categories i.e. positive, neutral and negative using machine learning techniques in Python. As in a previous example, but under a different concept, and this time with the intention of analysis, using Python as a direct tool to create machine learning, or machine learning using different libraries available in the environment.

Library Import
To run the Python scripts, some libraries are required. As noted below.

Figure 43

Import of libraries.

```
import numpy as np
import pandas as pd
import nltk
import re

import matplotlib.pyplot as plt
import seaborn as sns
```

9.1.5 Importing the Data Set

The data set that will be used to train the machine learning algorithm will use a file available in *.CVS format. This file contains a set of data, such as the user's tweet, the tweet ID, the name of the airline that the tweet text relates to, the count number, etc. You can use the *read_csv() method* of the Pandas library to import the dataset into the piece of software, which will perform the analysis, as shown in the following script:

Figure 44

Data extraction from *.CVS file.

```
dataset.head()
dataset = pd.read_csv(dataset_url, encoding = "latin-8")
dataset_url = "C:\Users\Administrator\Desktop\Tweets.csv"
```

Figure 45

Production **.** The following image shows the first five rows of the data set.

airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text	tweet_created
neutral	1.0000	NaN	NaN	Virgin America	cairdin	0	@VirginAmerica What @dhepburn said.	2015-02-24 11:35:52 -0800
positive	0.3486	NaN	0.0000	Virgin America	jnardino	0	@VirginAmerica plus you've added commercials t...	2015-02-24 11:15:59 -0800
neutral	0.6837	NaN	NaN	Virgin America	yvonnalynn	0	@VirginAmerica I didn't today... Must mean I n...	2015-02-24 11:15:48 -0800
negative	1.0000	Bad Flight	0.7033	Virgin America	jnardino	0	@VirginAmerica it's really aggressive to blast...	2015-02-24 11:15:36 -0800
negative	1.0000	Can't Tell	1.0000	Virgin America	jnardino	0	@VirginAmerica and it's a really big bad thing...	2015-02-24 11:14:45 -0800

data visualization

Next, we will perform data visualization. First, we plot the distribution of positive, negative, and neutral tweets in our data set using a pie chart.

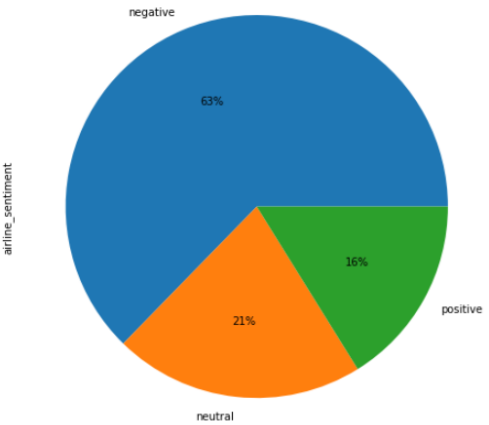
Figure 46

Sort in Python a circular figure as a graph.

```
dataset.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.0f%%')
fig_size = plt.rcParams["figure.figsize"] = [8.0, 8.0]
```

Figure 47

Production. Output of the results as a circular graph.



The result shows that 63% of the general tweets are negative, while 21% and 16% are respectively neutral and positive.

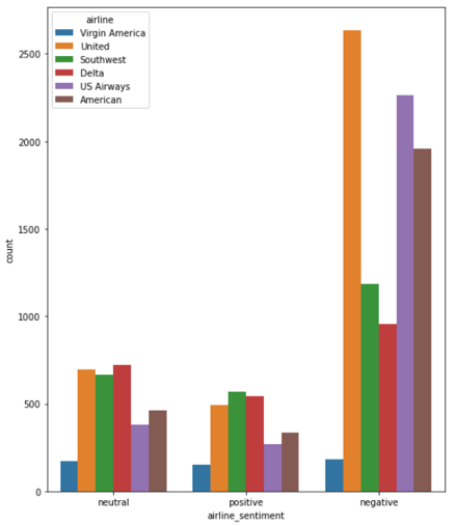
Figure 48

A bar chart is drawn showing the count of negative, positive and neutral tweets for the 6 airlines as shown below:

```
sns.countplot(x='airline_sentiment', data=dataset, hue = 'airline')
```

Figure 49

Production. I generate the column chart , in Python, to segment the sentiment analysis.



The graph above shows that United Airlines has the most negative and neutral tweets, while Airlines has the most positive tweets. Virgin America has the smallest number of positive and neutral tweets. However, the reason could be that Virgin America's overall tweet share is lower than that of the airlines.

9.1.6 Data Preprocessing

Now we need to remove the numbers and certain characters from the tweets. We'll define a function called *text_preprocess()* that accepts text strings and removes all text except the alphabets. Single and double spaces created as a result of digit removal, and special characters are subsequently removed.
The following script is executed to define the *text_preprocesses() function* . The first line of the function removes numbers and special characters. The second line of the function removes all generated uniques and this also represents a result of removing special characters. Finally, the third line of the *text_preprocess() function* removes the double blanks and replaces them with a single space.

Figure 50

I generate a function to process the special characters and numbers.

```
def text_preprocess(sen):

    sen = re.sub('[^a-zA-Z]', ' ', sen)

    sen = re.sub(r"\s+[a-zA-Z]\s+", ' ', sen)

    sen = re.sub(r'\s+', ' ', sen)

    return sen
```

Figure 51

Before we can 'clean' the tweets in this way, we must split the data into features and tags:

```
X = df['text']

y = df['airline_sentiment']
```

Figure 52

Next, we execute a *foreach() loop* that iteratively passes tweets from tweet list X to the *text_preprocess() method* that cleans up the tweet text. The following script is the one that performs the operation:

```
X_tweets = []
messages = list(X)
for mes in messages:
    X_tweets.append(text_preprocess(mes))
```

9.1.7 Text to Number Conversion

Since machine learning algorithms are based on mathematics and mathematics works with numbers, it is necessary to convert text tweets into numerical form.

Figure 53

Although there are several ways to do this, in this case I will use the TfidfVectorizer class *from the sklearn.feature_extraction.text module.* To do this, you can use the *fit_transform() method of the TfidfVectorizer* class shown in the following script:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords

tfidf_vec = TfidfVectorizer (max_features=2000, min_df=20, max_df=0.8, stop_words=stopwords.words('english'))

X= tfidf_vec.fit_transform(X_tweets).toarray()
```

The *max_features attribute* is used to specify the number of most frequently occurring words to convert, which is 5000 in this case. The *min_df attribute* specifies the minimum number of documents in which a word must appear (50). Finally, *max_df* specifies the maximum proportion of documents in which a word should appear, which is 80% in the above script. We also remove stop words like an, is, are, we, at, as they don't provide much information for ranking.

9.1.8 Division of Data into Training and Test Sets

Machine learning algorithms are trained on training sets and evaluated on test sets.

Figure 54

To split the data into training and test sets, you can use the *train_test_split() method of the sklearn.model_selection* module as shown below in this script:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
```

9.1.9 Training Machine Learning Algorithms

Although any sorting algorithm from the list of classifiers can be used here. The Random Forest classifier will be used, since it is the most robust. To use the Random Forest classifier in this case, you can use the RandomForestClassifier class *from sklearn.ensemble in* the scripts given as an example. To train the RandomForestClassifier class on the training set, you must pass the training functions (X_train) and training labels (y_train) to the *fit() method of the RandomForestClassifier class* .

Figure 55

Once the model is trained, predictions can be made by passing the test features *(X_test)* to the predict() method *of the RandomForestClassifier.* class. The following script must be executed to train the *Random Forest classifier* and make predictions.

```
from sklearn.ensemble import RandomForestClassifier

rf_clf = RandomForestClassifier(n_estimators=250, random_state=0)
rf_clf.fit(X_train, y_train)
y_pred = rf_clf.predict(X_test)
```

Algorithm Evaluation

Accuracy, F1, Recall, and Confusion Matrix can be used as metrics to evaluate the performance of a classification algorithm.

Figure 56

To do this in Python, you can use the *sklear.metrics module* to find the values of these metrics as shown in the following script:

```
print(accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

Figure 57

This is the output in the terminal of the results at a general level, based on the proposed algorithm:

```
felipe@Felipes-MacBook-Air PySentApp %
[[2154  116   70]
 [ 376  293   69]
 [ 171   79  332]]
              precision    recall  f1-score   support

    negative       0.80      0.92      0.85      2340
     neutral       0.60      0.40      0.48       738
    positive       0.70      0.57      0.63       582

    accuracy                           0.76      3660
   macro avg       0.70      0.63      0.65      3660
weighted avg       0.74      0.76      0.74      3660

0.7592896174863388
```

The result shows that the algorithm, and this piece of software, can efficiently and accurately classify a tweet as 'positive, negative or neutral' with an overall accuracy of 75.92%.

10. Conclusions and Recommendations

10.1 Subject of Study, Importance

The new marketing trends for companies mark the use of social networks as a means of communication of actions mainly due to their great potential to establish relationships with customers through them.
Social networks also facilitate corporate communication efforts to convey brand values, attract new consumers, disseminate marketing information or obtain measurable results from dissatisfied customers.
According to Pak and Paroubek (2010), social networks have been a useful tool to simplify communication with customers.

It is important to know the purpose of each social group, the format and the target audience and to be on the social networks where the audience and their target expect to find the brand.

10.1.1 General Objective of the Investigation

This research aims to measure the sentiment of users on the social network Twitter in relation to how, based on certain concepts, a psychographic model and its corresponding predictive models can be understood.

To do this, the research is based on data from the Twitter profiles of users, organizations, companies, etc., which were obtained using the Twitter API.
Once the tweets are downloaded, a developed algorithm powered by machine learning, using Python, is applied to split the sample (n=X tweets) into negative, neutral and positive sentiments.

10.1.2 Main Points of the Investigation

Sentiment Analysis with Machine Learning
Sentiment analysis is defined as the process of forming opinions in terms of qualifications, attitudes, and emotions about a particular topic. (Fiorini and Lipsky, 2012).

Sentiment analysis generally serves two purposes, first, expressions of sentiment and definition of sentiment orientation by individuals. (Honeycutt and Herring, 2009; Saura, 2018). Sentiment analysis makes it possible to detect the positive, negative and neutral expression on a specific topic, a product or service, entity, natural person, etc., of a

textual element. (Boyd, 2017; Chunga, 2017).

Debes (2017), indicates that sentiment analysis can refer to approaches and be based on characteristics and automatic tags, conversations, or it can be the case of common tags in a theme or specific events of use, in emoticons or in resources such as lexicons. of sentiment that is with positive, neutral or negative tweets. Fundamental lexicons label words collected in a semantically necessary dimension, called "feeling", "valence" or "semantic orientation" (Saura, Palos-Sanchez / Cerdá, 2017).

Algorithms developed in Python to perform sentiment analysis have predictive power.
The prediction is determined by machine learning. Machine learning is a form of artificial intelligence that trains the virtual machine through data exploration to automate the data analysis process, among other features.

10.1.3 Objectives, Achievements and Formats

After the development of the methodological process that includes the analysis and extraction of data, the psychographic sentiment analysis was carried out in this investigation, as a result of the analysis, it was possible to obtain the average, through the application of the algorithm with automatic learning.

In the table or *output in the terminal, of the tests and results (Part II),* they delivered the results of the analysis process of processed tweets on the sentiment of a determined number of users regarding 6 United States airlines, in addition to the use of the interactions by the companies and the interactions and comments made by the users about them, the categorization made according to sentiment and the average veracity obtained as a result of the machine learning algorithm. As specified above, it was possible to efficiently and accurately classify a tweet as 'positive, negative or neutral' with a total accuracy of 75.92%.

10.1.4 Feasibility And Potential Of The Research And The Project

The main engine of the application used is machine learning, with repeated use and training of average results increases.
As it has been shown, Twitter is configured as the optimal social network for users who can express their feelings, opinions and comments in a specific way, in real time and all over the world.
Twitter has been used as an object of study in multiple periods during the last decade.
In this research, Twitter was used to prioritize the sentiment of the offers, and therefore the quality, of the companies that make up the sample when they publish on said social network.
The results of the research can be used by companies to improve the development of their strategies in social networks and, more specifically, around social Twitter.

Additionally, the search results identify communications and offers via Twitter for others related to the data analyzed, in terms of tone, and what businesses can take advantage of on social media.
This research provides verified data on Twitter actions that can be used for future marketing strategies and will serve as a source of research on the specified topic.

10.1.5 Recommendations for Further Investigation

- Why should it be done?

Without underestimating the value of quantitative KPIs, sentiment and interest in establishing more precisely the audience to which a brand is directed is important, since it allows us to analyze the emotional response of users who have interacted with it. In today's digital environment, people have more tools to express their opinion, share their doubts and express their concerns, whether positive or negative.
That is why sentiment analysis is postulated as an asset that allows obtaining information about the tone in which users speak on social networks, or forums of a certain brand, something that allows people to know what they like and what they like. not.

- What thing should be done?

Sentiment analysis is a process in which machine learning or machine learning is used to find a spot on a keyword or sentiment, and place the information of interest in the process. The feeling can be defined as the result of "a judgment or judgment formed about something, not necessarily facts or knowledge". But with the use of sentiment analysis and data science, opinion or judgment is understood.
It is a subjective evaluation of something based on personal empirical experience.
It is made up of objective facts and partly by emotions. An opinion can be interpreted as a dimension in data on a particular topic.
It is a set of signifiers that, combined, present a vision, that is, a feeling about a particular subject.

Sentiment analysis algorithms
There are two main methods of sentiment analysis.

- *Rules-based approach*

Rule-based sentiment analysis is based on an algorithm with a clearly defined description of an identified sentiment.

- *It includes identification of subjectivity, polarity or theme of opinion.*

The rule-based approach involves a basic natural language processing routine.

Is that how it works:
There are two lists of words.
One of them only includes positive, neutral points and the other negative points.
The algorithm scans the content, finds words that match the criteria.
After that, the algorithm calculates what type of words are most frequent in the content.
If there are more positive words, the text is considered to have a positive polarity.

- Who will it benefit?

Sentiment analysis or psychographic analysis deals with the perception of users around ideas or products, based on an understanding of the market, through the prism of sentiment data that can be found already registered by brands and companies. in social networks or in general on the Internet.
There are many public and private sources of information from which you can obtain information about the quality of

the product by the customer or the quality of the relationship of the user or customer with the product.
To name a few:

- Customer service calls and emails.
- User-generated product reviews.
- Posts, responses or comments on social networks.
- General and special forums.
- Record of interactions with customers.

Sentiment analysis can help companies make sense and add value to and transform the accumulation of unstructured data.
A clearly defined view of what certain customer segments think of the product or the company in general. A deep dive into the state of the market from a consumer perspective.
In any case, it is an influential factor in the formulation and elaboration of the value proposition for a specific audience segment. While at first these activities are relatively easy to do with basic solutions, at some point it becomes logical to use more elaborate tools and extract more sophisticated insights.

- Who will do it?

They are engineers, and fundamentally, Data Scientists, or companies/entrepreneurships dedicated to these specific topics, where they broaden their gaze to different disciplines much more.
A Data Scientist or data scientist is the professional dedicated to the collection, analysis and understanding of large volumes of data and their respective extraction.
They are people who apply their knowledge of statistics and programming to analyze and interpret the data available to companies and extract valuable information.
Organizations hold a wealth of information that, if used correctly, can translate into business benefits. In an increasingly digital environment, taking advantage of the information that companies obtain with their environment is almost essential. Hence the growing need for professionals who can analyze and make sense of all this data, so that it has real value.

What does a Data Scientist do in a company?
The functions of a data scientist may differ from one organization to another, but broadly they include the following:

- Data mining.
- Get all the information you consider useful from various sources.
- The data volume may differ.
- Data cleaning.
- Remove all information that is not irrelevant to prepare the data for processing.
- Data processing.
- Process data applying statistical approaches, analysis, machine learning, predictive models, etc.
- Data visualization.
- Represent data in different ways to make it understandable in the most accurate way.

- Where will it be done?

These technology implementations are carried out in systems controlled and ideally managed by engineers in technology and data analysis companies. This branch is constantly researched, so it is common and very natural for these companies to carry out a lot of research in parallel to the delivery of services, development of technological products for itself and/or consulting.

11. General Conclusions[9]

Undoubtedly, the approach of the problem of a technical investigation supposes an important step for its development, allowing to establish who will be studied and, consequently, which ones will be resolved.
The approach to the problem, as well as the objectives of the research, the starting point of this topic, updating what was done in the previous topic, establishing clearly and concisely the initial direction that was given to the investigation.
This occurs since the investigation part previously constitutes an orderly and coherent reflection that highlights a logical transition of actions and objectives aimed at the detected problem.
A custom software-level product has a positive and desirable impact on the organization where it is implemented.
This happens because it was designed and based on the particular needs of those who need it.
For this impact to be truly positive and generate benefits, its development must be carried out in an orderly manner, with adequate standards and good practices that guarantee compliance with the agreed deadlines, the associated costs and that it works for what it was created for, complying with the requirements the client and the process.

The development life cycle of an information technology application or product is nothing more than a structured and clearly defined sequence of steps to consider when developing a software or information technology product.

Without any doubt, the development of a prototype, capable of the hardware or software of a system, is one in which a

[9]Title Project branch . IACC (2022).

large number of variables must be taken into account.
In addition, it includes a first design stage, an intermediate stage that includes development, and a final stage where the results are evaluated, this last stage in which the interrelation of the parts is tested. On the other hand, flowcharts offer a unique way to visualize and organize complex processes in an easy way, which makes them an excellent tool to improve problem solving, as well as an effective way to share information.
As important as verifying that a user can use the application, it is equally important to verify that the system continues to function correctly when unexpected actions are taken or incorrect data is entered.
Therefore, a good test suite should push the application or software to its limits, not to mention that in the case of automated tests, these are also code, so they deserve detailed consideration as well.
Research is the logical systematization of information used to obtain new knowledge, to discover relevant data or truths related to the facts that are analyzed.
On the other hand, the APA standards contain guidelines that have been universally accepted when documenting the different steps carried out during an investigation.

References.

Balestrini, M. (2006). How the research project is made. 7th edition. Caracas, Venezuela: Associate Consultants.

Restrepo, M. (2008). Production of educational texts. Bogota, Colombia: Editorial Magisterio.

Fernandez, V. (2006). Development of Information Systems a Methodology Based on Modeling. Barcelona, Spain: Editions UPC.

Granados, R. (2014). Development of web applications in the server environment. IFCD0210. Malaga, Spain: IC Editorial.

Sommerville, I. (2005). Software Engineering. Madrid, Spain: Pearson Education.

Barranco, J. (2001). Methodology of structured systems analysis. 2nd edition. Madrid, Spain: Comillas Pontifical University.

Granollers, T.; Vidal, J. and Cañas, J. (2005). Design of user-centered interactive systems. Barcelona, Spain: Editorial UOC.

Lopez, A. (2007). Introduction to program development with Java. Mexico DF, Mexico: National Autonomous University of Mexico.

Mompin, J. (1988). Introduction to bioengineering. Barcelona, Spain: Marcombo.

Alvarez, L. (2009). The materialization of ideas. Realities, needs, opportunities, encounters and disagreements. Postgraduate thesis. Barcelona, Spain: Autonomous University of Barcelona.

Gomez, S. and Moraleda, E. (2020). Approach to software engineering. Ramón Areces University Editorial.

Granados, R. (2015). Deployment and commissioning of software components. IFCT0609. IC Editorial.

Serna, E. (2013). Software Functional Testing: A constant verification process. Medellin, Colombia: Metropolitan Technological Institute.

Barranco, J. (2001). Methodology of structured systems analysis. Madrid, Spain: Comillas Pontifical University of Madrid.

Garriga et al. (2010). Introduction to data analysis. Madrid, Spain: Editorial UNED.
PopulationPyramid.ne (2019) Population density by country. https://bit.ly/3sTwwYl

Requena Serra, Bernart (2014). Bar chart. Universe Formulas. https://bit.ly/2Q1L7PQ

Rodriguez, E. (2005). Investigation methodology. Tabasco, Mexico: Juárez Autonomous University of Tabasco.

Saiz Carvajal, Rosario (2016). Information analysis techniques (summary). https://bit.ly/3zA53ia

Malhotra, N. (2004). Market research: an applied approach. 4th edition. Mexico: Pearson Education.

Munoz, C. (2018). Investigation methodology. Mexico City, Mexico: Editorial Progreso SA

Pardinas, F. (2005). Methodology and techniques of social science research. 38th edition. Mexico City, Mexico: Siglo XXI Publishers SA

American Psychological Association (2002). Publication Style Manual of the American Psychological Association. Mexico: Editorial The Modern Manual. Apa Standards (2019). Apa Standards 2019 Updated. https://normasapa.com/

Rodriguez, E. (2005). Investigation methodology. Mexico: Publications of the Autonomous Juárez University of Tabasco. SANCHEZ, CARLOS (2014). Apa Standards – 7th (Seventh) Edition. Https://Standards-Apa.Org/

Other use and research references

Big Data with Python – Collection, storage and processes. By R. Caballero, E. Martin & A. Riesco. (Personal book).

Programming in C, C++, Java and UML Software Engineering. By Luis Joyannes and Ignacio Zahonero. (Personal book).

Descriptive statistics, probabilities – Inference – Regression models and non-parametric methods. By Pedro Vergara Vera. (Personal book).

BOYD, D. (2007). "Social network sites: Definition, history, and scholarship". Journal of
Computer-Mediated Communication.

BULUT, A. (2015). Lean Marketing: Know who not to advertise to! Electronic Commerce
Research and Applications.

CHUNGA, A., ANDREEVA, P., BENYOUCEF, M., DUANE, A., O'REILLY, P. (2017). Managing an
organization's social media presence: An empirical stages of growth model. International
Journal of Information Management.

YOU MUST, V., SANDEEP, K. AND VINNETT, G. (2017). Predicting information diffusion
Probabilities in social networks: A Bayesian networks-based approach. Journal of Knowledge-
Based Systems.

FAGAN, JC (2014). The Suitability of Web Analytics Key Performance Indicators in the
Academic Library Environment. The Journal of Academic Librarianship.

FILE, KM, AND PRINCE, RA (1993). Evaluating the effectiveness of interactive
marketing. Journal of Services Marketing, 7(3), 49-58. doi:10.1108/08876049310044574

FIORINI, PM, AND LIPSKY, LR (2012). Search marketing traffic and performance models.
Computer Standards and Interfaces.

HERRÁEZ, B., BUSTAMANTE, D. AND SAURA, JR (2017). Information classification on social
networks. Content analysis of e-commerce companies on Twitter.

HONEYCUTT, C., & HERRING, S. C (2009). Beyond microblogging: Conversation and
collaboration via Twitter. In 42nd Hawaii International Conference on System Sciences.

JÄRVINEN, J., AND KARJALUOTO, H. (2015). The use of Web analytics for digital marketing
performance measurement. Industrial Marketing Management.

LEEFLANG, P, VERHOEF, P., DAHSLTRÖM, P. & FREUNDT, T. (2014). Challenges and solutions for marketing in a digital era. European Management Journal.

MATHEWS, S., BIANCHI, C., PERKS, KJ, HEALY, M., AND WICKRAMASEKERA, R. (2016). Internet marketing capabilities and international market growth. International Business Review.

MCDERMOTT, J. (2017, December 03). Black Friday stats: The numbers behind the madness. Retrieved April 18, 2018.

PAK, A., & PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of LREC, 2010. Valletta, Malta.

PALOS-SANCHEZ, P.; SAURA, JR (2018). The Effect of Internet Searches on Afforestation: The Case of a Green Search Engine.

SAURA, JR, PALOS-SÁNCHEZ, P., & CERDÁ SUÁREZ, LM (2017). Understanding the Digital Marketing Environment with KPIs and Web Analytics. Future Internet, 9(4), 76.

SAURA, JR, PALOS-SANCHEZ, PR & RIOS MARTIN, MA (2018). attitudes to environmental factors in the tourism sector expressed in online comments: An exploratory study. International Journal of Environmental Research and Public Health.

TT, KUO, S.-C. HUNG, W.-S. LIN, N. PENG, S.-D. LIN AND WFLIN (2012). Exploiting latent information to predict diffusions of novel topics on social networks, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics.

Algorithmic future. Big data, sensor data and mobile media, Co-authors: Jose Correa and Charles Thraves.

Sort Algorithms and Data structures, by Nchena Linos.

Data Structures and Algorithms with JavaScript, by Wahyu Prayogo. JS & Data Structures.

Programming Problems: Advanced Algorithms (Volume 2) Paperback – February 27, 2013, by Guido Noto La Diega, PhD.

Superintelligence: Paths, Dangers, Strategies Reprint Edition, by Nick Bostrom (Author), PhD.