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

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

Analysis & modeling of psychographic data on the internet using machi
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://t
https://glzengrg.com - https://freeshell.de/ felipe - https://li

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

Computer Science Engineer (Ingº en Informática), Graduated with maxi

Graduated / Titulado 2023. -

2020-2023

Keywords: Big Data, Data Analysis, Machine Learning, Softwar

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 in 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 is done 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 right ly 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 dreamain 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 reputation, 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 those that develop the pieces of software that will allow you to deliver a general and complete analysis that needs to be visualized.

- Social projection

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

- What does it solve?

Solves problems derived from the construction and communications towards certain psychographic user groups on the Internet.

- Who / who is it for?

All who wish to have a quick and timely solution to problems quick solutions required to address certain groups, people, etc., who are involved in networks on the Internet.

1.1.1 General Objective

Develop pieces of software that make it possible to data compression on the Internet for large amounts of All this will facilitate the compression of data on Internet relative to large groups of users, and the analysis and modeling in relation to the data that the analysis that will be carried out on groups in the Internet in general, all available (Users, networks, etc.). All this in order to deliver to companies, organizations or governments that a clear course of action to position their

1.1.2 Specific Objectives

Analyze and update the flow chart of the analysis with the company or organization.

Build certain software for the client and / or software that work, for example, in a terminal and that allow building statistical models

Verify the operation of the scripts in the editor on the server

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

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

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

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

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 by the private company, organization or government.

Definitely:
The fundamental premise is that the system, and through its implementation, will deliver data a[illegible] communication adjustments that affec[illegible] Internet. People or users have allowed an area known as data science to develop over time, where large amounts of information are stored and in turn allow the analysis of [illegible] 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 mine.
To perform all these scans, procedures and [illegible] 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.

[illegible]tion of this research will focus in the methods how companies relate to the organization is also involved in monitorin In addition to implementing these procedu type of quantitative investigation would b the results of everything that would be ana to the users who are involved with the organ nature of the descriptive investigations w since it describes in detail all the elements implemented system and that is exposed, w organization.

In this sense, it is also suggested that in thi of this research is quantitative since it is b of the reality of these processes, quantifyi that intervene in the system that is to be im providing a possible solution. On the other research is of the descriptive/explanator the one hand, the dimensions of the problem measured, describing what is being invest other hand, the phenomenon must be explai solution can be offered to the dynamics tha investigated. raise.

1.2.1 Research Technique

In this project, the research technique wil non-experimental type since a study will b which will be applied to the data with the cu the organization, and the design of the syst implemented, which will consist of observ procedures currently available to solve th

2. Theoretical Framework

Now, I will refer to the theoretical framew
The research shows that the application fo must, be developed on an operating systems Linux, using terminals with Python, under specifically Linux (Debian), ideally in n operating system based on UNIX, since Win [illegible] of my research and analysis, all the ability to extend to different parts of th [illegible], for this, references are taken, it is a database server, such as MySQL, since idea collected can be saved in it. MySQL has sho its history also a lot of robustness, it is an e engine needed. All done
[illegible] these concepts are present

Python[2]: Python is a high-level interpreted pr language whose philosophy emphasizes th code, it is used to develop applications of a examples: Instagram, Netflix, Spotify, Pa others.[2] It is about a multi-paradigm pro language since it partially supports obje imperative programming and, to a lesser ex functional programming. It is an interpret cross-platform language. Managed by the Python Software Foundati open source license, called the Python Soft License.[3] Python consistently ranks as on popular programming languages.

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

MySQL[2]: MySQL is a relational database management system developed under a dual license: Free/Open Source and Proprietary. Free/Open Source License / Commercial License by Oracle Corporation and is considered the world's most popular open source database, 12 and one of the most popular databases used in conjunction with Oracle and Microsoft SQL Server, as well as web development environments.
MySQL was initially developed by MySQL AB (now a subsidiary of Oracle Corporation) 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 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 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 Maria DB 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[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 for 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 in its versions. The de
from companies, created by
depending in any way
Debian does not directly sell its soft
makes it available to anyone on the Interne
allows new individuals or companies to com
distribute this software as long as its licen
GNU / Linux can use different insta
mechanisms, such as: DVD, CD, USB, and ev
from Internet networks (the latter depends on the
or a
).

3. Methodological Framework

First, the development of a survey is not lim
definition of the thematic content or its di
therefore, it is imperative to carry these co
methodology that is oriented towards the a
proposed objectives.
In this research budget, the bases must be t
procedures necessary to obtain the require
thus being able to structure or solve the pro
research process begins when one asks a que
one does not know the answer and which mus
To do this, such research must be planned, w
going through a set of preliminary phases,
of study, and outlining the components. In
projecting question, the approach to solve t
regarding the creation of pieces of softwar
respective models and analysis is first and
"conversation with the users involved", of
the management, to Being able to integrate
elements for computer development, then
the visual design of the application if neces
technically analyze the programming for i
development, servers, communications sys
clearly define the creation of all develop

3.1 Characteristics of the System to be Developed During the Investigation

They are pieces of software that focus on da
these pieces of software can either work un
being executed on a Mac OS X server and fun
Debian GNU/Linux.

3.1.1 Methodologies and Fundamentals of Development Raised

The iterative model of the cascading lifecy
in the project. This model is intended to rec
user needs and the final product may arise f
misunderstandings during the requirem
It consists of the iteration of several casca
At the end of each iteration, the client is de
improved or more functional version of any
product offered. In this case, as I have alre
before, it corresponds to pieces of softwar
Terminals of a Debian GNU / Linux server.
The client is the one who, after each iterati
product and corrects it or suggests simprov
the delivered results, not the performance

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

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

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

These iterations will be repeated until [illegible] satisfy any type of need, but above all to [illegible] software pieces, since the source code of the specific pieces of software is not delivered for th[illegible] delivered as product, the results they w[illegible] Specifically, this methodology gives us [illegible] work on our system since each piece of so[illegible] version in itself. Such as, for example, [illegible] GNU/Linux Kernel and its constant de[illegible]

This model is generally used in projects where the requirements are not clear to the user, so it [illegible] 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 T ype 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 automa[illegible] 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 throug[illegible] website visits, social media responses, loy[illegible] and online surveys to learn more about cust[illegible] different collection methods and techniqu[illegible] useful.
The choice of method depends on the strate[illegible] variable, the desired precision, the collec[illegible] skills of the interviewer. Interviews are o[illegible] common methods. If it is decided to do so, sp[illegible] [illegible] questions that will be ask[illegible] also in a face-to-face interview, by telep[illegible] email or via social networks.

- o Questionnaires are a useful tool for dat[illegible]
- o To get the expected results, they must b[illegible] out carefully.
- o This is why before writing it, it is impor[illegible] researcher to define the objectives of[illegible]
- o There are two questionnaire formats: [illegible] questionnaires that are applied when [illegible] know the opinion of your experiences [illegible] on a specific topic.
- o On the contrary, in the closed question[illegible] researchers control what they ask and [illegible] want to know, which may force them to [illegible] responses of the participants.
- o If [illegible] on-site observation to get t[illegible] customers, they can be done according [illegible] methodologies.
- o Taking into account how the informati[illegible] it will be of great help when analyzing [illegible]
- o Being able to measure and report accur[illegible] data is also very important for good de[illegible] making.

3.1.4 Data Collection Instrument Used

Questionnaire with the following questio[illegible]
Tell us about your impression and how you f[illegible] organization.
The first question in the customer or user q[illegible] includes all the basic information about th[illegible] consists of:

o Name of the client or company or users

Important data of the person (A complete [illegible] investigation of the user is carried out, u[illegible] what he says, how he says it, etc.)
Contact details, usernames, etc.

o What are your target customers?

This should include all applicable demogr[illegible] sentiment expressed, images posted, etc[illegible] the client's audience type will give an idea [illegible] 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:

[U+25AA] Feelings of closeness towards our ideas

[U+25AA] Influencing specific social groups

[U+25AA] Accumulated data and models of 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, examples of features that might be diffi can be especially helpful for the custom characteristics of your 'competition'.

What is the deadline to start delivering models that means something important their data collection and analysis task?
This information can be used in the cust o 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 to on which research is carried out or studies a Sample is a part or subset of elements that a selected from a population to carry out a stu For our research, the population is made up users, involved in certain ideas that an org company represents, which are:

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

The sample that we will use will be the diffe users on the Internet.

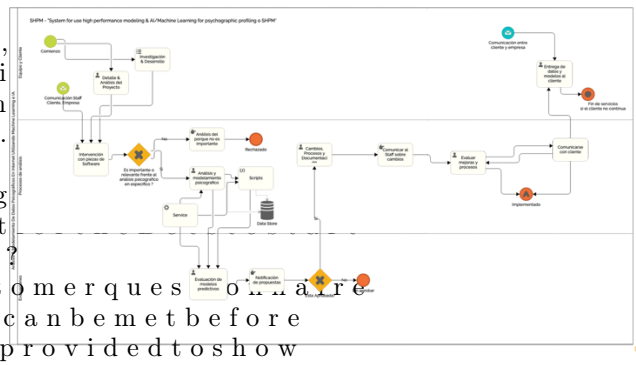
4 Analysis of the Results

First, the technique that will be used for quantitative observation.
The object of study is: The amount of data t a given amount of time.

For this, a survey will be taken of the worke additional data will be taken from the head of closing these surveys to collect data eith the implementation of the respective servi problem posed.
All of the above is to achieve a completed ai annual analysis. In case of presenting any i can be resolved immediately since all the corr backup systems will be available, in such a registration is efficient and consistent.

4.1 Flowchart of the proposed investigation

Figure 1
Flowchart, for the proposed project.



Note. It is advisable to have a complete vis f

5. Explanation of the flowchart of the proposed investigation

The explanation to the previous flowch
give the sample of the beginning of there
or of projects and processes, initially b
with the client, about details and analy
project itself. All also in constant comm
with the team and the client, which will
resource of vital importance for the ana
evaluation processes, as appropriate.
mention that, constantly, whether the
research and development can be carrie
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 the [illegible]
'X', compared to certain concepts, ideas
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 Psychograph

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

We also rely on the construction of popular
subjectivity in certain messages, and thro
and they are classified to determine within
whether or not they are -Bots. That allows y
larger filter. It can be of great importance
organization, since in this way they can rel
messages, to approach an 'A' or 'B' populat
determine their values, likewise the compa
will make better decisions to get closer to c
feelings on the Internet.

Like any piece of software, it is subject to er
we already use Open Source libraries in Pyt
TextBlob, among others to streamline pro
stage, the scripts that determine certain d
example, go to an integration stage, QA or T
processes).

Below are some portions of the code from th
its results in a terminal, using VS Code, in a
environment & under Testing on a Debian G
server. (We'll use the 'Virgin Airlines' va
as an example.)

In this case, our piece of software is in a stag
maintenance phase within the life cycle of a
system.

Figure 3

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

Figure 4

Authorization accepted by 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 the extraction 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

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

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:

🇦🇪 Emirates
🇶🇦 Qatar Airways
🇸🇬 Singapore Airlines
🇯🇵 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:

🇦🇪 Emirates
🇶🇦 Qatar Airways
🇸🇬 Singapore Airlines
🇯🇵 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

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

Positivo
Sentiment(polarity=0.10416666666666667, subjectivity=0.5)
RT @stats_feed: World's best economy class airlines in 2022:

🇦🇪 Emirates
🇶🇦 Qatar Airways
🇸🇬 Singapore Airlines
🇯🇵 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 and models have already been detailed above. Python is one of the most popular languages in the world and, in addition to the recommendation algorithm and the one that self-driving cars were created.

In terms of scalability, Python has an advantage over other 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 there
  can learn python in less time.
- Python is also a versatile and easy-to-
- In terms of scalability, Python has an a
  over programming languages like R in
  one approach to solving different pro
- In terms of speed, Python also stands ou
  MATLAB and Stata.
- It has a great library.
- A library or a library is a set of which are
  together.
- It can be used over and over again for pro

[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

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

Python is an interpreted programming language, that is, it's 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 variables to numeric values.
In this way, the model will be able to understand and act 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 front line 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 package that is based precisely on the Matplotlib code.
Therefore, the use of both libraries is ar
data science.

6.1.2 V isual Studio Code for Programming in Python

Visual Studio Code or better known as VS Code 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 VS Code 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 c
contains a search engine, so you can sear
by name, and a list grouped into three ca
popular, and recommended.

6.1.3 Python And T esting

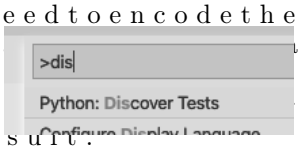
This Python's extension, developed by Mic
Python features to VS Code, such as autoco
code formatting, as mentioned, debugging
Python code and environment management
Another consideration to take into accoun
test that I will use, to choose between Unitt
Nose.
We can easily do this by accessing the confi
ting Python Testing in the search bar. In
Unittest, I mark the corresponding box to a
directly in the JSON configuration of VS C
case the test of a piece of code related to use
micro-system in a Terminal will be used.

6.1.4 Running And Debugging T ests

Enable unit testing in VS code.
To enable Unittest in VS Code, we run the Di
command.

Figure 12

Add-on search in VS Code to find a Testing A

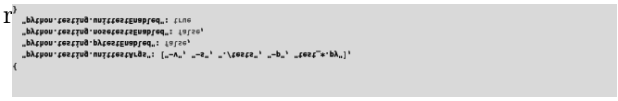


This will prompt us to configure the testing
used, Pytest or nosetests, Unittest in this
Unittest.

Once configured, the .vscode/settings.js
will be updated like this:

Figure 13

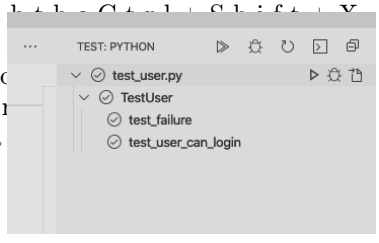
Editing JSON again the VS Code configura



This configuration defines the Unittest ar
defines which Python testing framework is
are located in the ./tests' folder and the te
test files test_*. Once we have Unittest a
configured with VS Code, we can run the tes
Testing tab.

Figure 14

Example in Testing, from VS Code, passing
e



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

Figure 15

Functions that will serve as an easy exa... 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 li... performed - Debug Test.

```
 9

    Run Test | Debug Test
10  def test_user_can_login(self):
11      """Test that the user can login."""
12      self.assertTrue(True)
13
```

7. Code Listing, Scripts

Next, I will list the central sequences of... piece of software, where a psychograph... positive, neutral or negative feelings o... carried out and of a quantity of data to b... depending on how the piece is interacte... that runs in a Terminal, and that genera... directly with Python, in addition to del... 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 spor... within Chile – 'Extreme Sports'.

Figure 17 & 18

Again, the exercise of using the necessary l... 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 pe... c.

```
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
```

Figure22

```
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
```

Figure23

```
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')
```

Figure24

WesendtotheTerminal,thedataanaly

```
tweet_list = pd.DataFrame(tweet_list)
neutral_list = pd.DataFrame(neutral_list)
negative_list = pd.DataFrame(negative_list)
positive_list = pd.DataFrame(positive_list)
print("total number: ",len(tweet_list))
print("positive number: ",len(positive_list))
print("negative number: ", len(negative_list))
print("neutral number: ",len(neutral_list))


tweet_list
```

Figure25

Webegintogeneratetherespectivepie
containtheanalysisinpercentages.

```
labels = ['Positivo ['+str(positive)+'%]' , 'Neutral ['+str(neutral)+'%]','Negativo ['+str(negative)+'%]']
sizes = [positive, neutral, negative]
colors = ['yellowgreen', 'blue','red']
patches, texts = plt.pie(sizes,colors=colors, startangle=90)
plt.style.use('default')
plt.legend(labels)
plt.title("RESULTADOS DE ANÁLISIS PSICOGRÁFICO BAJO: =  "+keyword+"" )
plt.axis('equal')
plt.show()
```

Figure26

Wecheckforduplicates.

```
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]
```

Figure27

Weremovecertaintypesofcharactersthatmaybe
uncomfortableinthefaceofanalysis.

```
#Borrando RT, Punctuation etc
remove_rt = lambda x: re.sub('RT @\w+: '," ",x)
rt = lambda x: re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)"," ",x)
tw_list["text"] = tw_list.text.map(remove_rt).map(rt)
tw_list["text"] = tw_list.text.str.lower()
tw_list.head(10)
```

Figure28

Webegintoanalyzethepolaritiesofpsych
analysis.

```
tw_list[['polarity', 'subjectivity']] = tw_list['text'].apply(lambda Text: pd.Series(TextBlob(Text).
sentiment))
for index, row in tw_list['text'].iteritems():
    score = SentimentIntensityAnalyzer().polarity_scores(row)
    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    comp = score['compound']
```

Figure29

Comparisonsaremadeagainstwhethersuc
higherorlower,asanexample.

```
    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)
```

Figure30

Wecount,andgeneratetotalsandpercenta

```
#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'])
```

Figure31

Usingtheaboveanalysis,wecreatedagrap
themostfrequentlyusedwords.

```
#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"
```

Figure32

Weoverwritetheimageonwc.pngandweca
n

```
#Generando, wordcloud para todos los twitts
create_wordcloud(tw_list["text"].values)

# ### seccion para mostrar la imagen generada de WordCloud
image = mpimg.imread("wc.png")
plt.imshow(image)
plt.show()



#---
```

7.1 Running Pieces of Software (Scripts)

The execution of the script is directly using VS Code.

Figure 33

Using VS Code we carry out the execution
the Terminal.

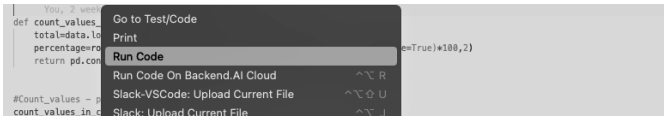


Figure 34

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

Figure 35

The following figure shows the type of p
generated, using the above method.

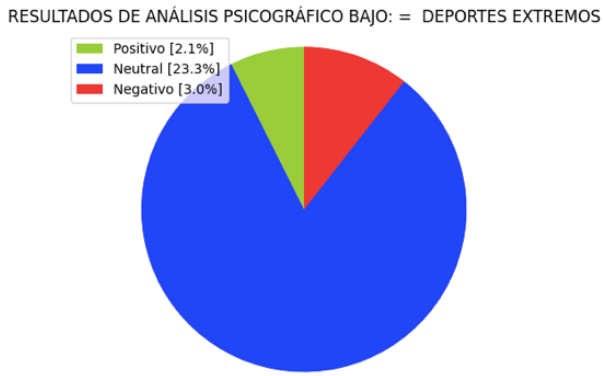


Figure 36

*It delivers us through the Terminal a me
image has been successfully generated with Python.*

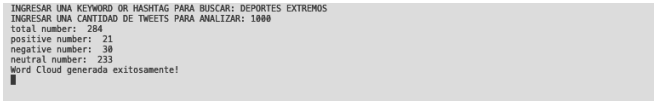
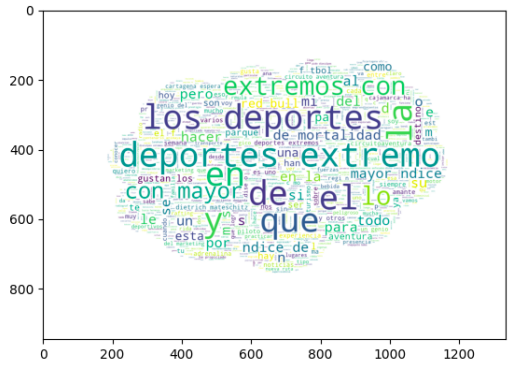


Figure 37

The output is as follows, a graph:



Note. It is an automatically generated gra
would like to mention that it was possible to
create a mixed system, with Python and No
this type of content, with a structure in ano
maintains this data and displays it online.

8. System Tests, Explanation, Exit Criteria and Results

A challenge and of
the software and to prove that the software
requirements to get the necessary results b
customer.
This helps the development team to fix bug
good quality product.
There are several points in the develop men
human error can lead to the software not me
necessary requirements. Fundamentally,
function, and stress tests will be generated
that the data provided by the graphs and ou
in terms of the algorithm used.

Dynamic tests will be carried out, because
piece of software that will be used with the
delivering data analysis, is a type of softw
input and output behavior will be verified
obtaining the expected data.. The functio
part(s) are operational. Unit tests will be
piece or pieces of software that will deliver
The carrying out system tests, checking i
operability, security, stability, etc. All t
fundamentally scripts, or pieces of softwa
be executed in a Terminal.

For the project, a type of dynamic test
as mentioned, and in this case, of each one, s
objective of the pieces of software are the o
of what is processed (In this case, external
of software would be considered as indepen

The results obtained during the testa
since they are extracted from specific mea

[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:

and/orpsychographicstudyanalysis,the outputsareatthesametimeprecise.Withoutfurther consideringthat,iftherewasabuginthesourcecode,you wouldnotbeabletoexecutethescriptatallintheTerminal.

Implementingatestservicefromscratchisacomplexand time-consumingtask.
Incontrastingprojects,weseethatsmall,butefficientand relentlessstepshavebeentakentowardscontinuous integrationQAservice.Stepssuchashiringspecialized peopleinthefield,implementingtoolssuchasTestlink,test management,SonarQubetoassesscodequality,Jenkinsfor continuousintegration,orSeleniumfortesting.Horizons suchasBigDataTestingappearinthefutureofTesting,so thefutureofTestingisguaranteed.

8.1 Big Data T esting

InBigDatatests,QAengineersverifythesuccessful processingofterabytesofdatausingoneandother supportingcomponents.Thisrequiresahighlevelof testing skills,processingisveryfast,andcanbeofthefollowing types:

- oBach
- oRealtime
- oInteractive

BigDatatestscanbedividedintothree:

Step1:ValidationStageTheBigDatatestingstage,also knownasthepre-Hadoopstage,involvesvalidationofdata fromvarioussourcessuchasrelationaldatasources,blogs, socialmedia,etc.,needtobevalidatedtoensurethatthe correctdataisinthesystem.Thesourcedataiscomparedto thedataenteredintotheHadoopsystemtoensureamatch. Itverifiesthatthecorrectdataisextractedandinthecorrect location.
Step2-Validationof"MapReduce":Inthisstep,youcheck thevalidationofthebusinesslogiconeachandthen validatethemafteritrunsonmultiplenodes,ensuringthat:

- oTheMapReduceprocessworksfine.
- oTherulesofaggregationorsegrega inthedata.
- o“KeyValuepairs”canbegenerated.
- oThedataisvalidatedaftertheMapReduceprocess.

Step3-Phaseofvalidationoftheresults:thirdandlast stageoftheBigDataTestingtests,istheprocessof validationoftheresults.
Outputdatafilesaregeneratedandmovedtoanenterprise DataWarehouseoranyothersystembasedonthese requirements.
Theactivitiesinthisthirdstageincludethefollowing:

- oItisverifiedthatthetransformationrulesare appliedcorrectly.
- oDataintegrityisverifiedandsuccessf tothetargetsystem.
- oDataischeckedforcorruptionbycomp targetdatawiththedataintheHDFSf

8.1.1 Hadoop

Hadoophandlesverylargevolumesofdata amountofresources.Therefore,architect toguaranteethesuccessofaBigDataproje presented.Apoorlydesignedorinadequat resultinpoorperformanceandthesystemm requirements.

- oThetestservicesmustberunningonaH
- oPerformancetestsincludetestsforjob memoryusage,datathroughput,and metrics.

9.TestingAndAnalysisOf ResultsIi

TounderstandwhatIwilldetailbelow,and referstothetestsandanalysisinparttwoo itisnecessarythatyoucanimplycertaincl models.

9.1 Classifiers & Model Examples

NaiveBayes[8]isasimplemodelthatIconsiderim mention,amongothersthatexistandcanb analysisandclassificationpurposes;this dataclassification.Inthemodel,theclass tweett,asseeninthefollowingformula:

Figure38

NaiveBayes’textclassificationmodel.

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

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

Note.Adaptedfrom"twitter_sentimenta [Image],bypantechsolutions.net-Twitte AnalysisusingMachineLearningonPytho https://bit.ly/3hDjMmy

Intheaboveformula,firepresentsthei-th totalofnfeatures.P(c)andP(fi|c)canbeo maximumlikelihoodestimates.

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

9.1.1 Other Models

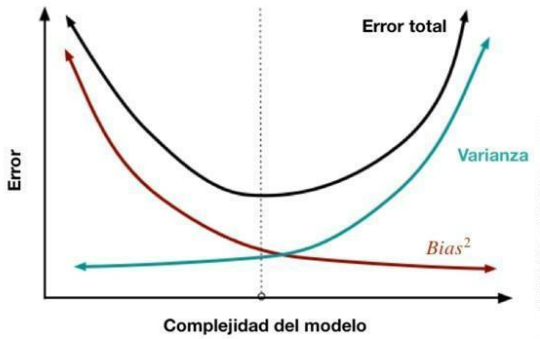


Figure 39

Model/EquationofMaximumEntropy.Inabinary classificationproblemliketheonewe'retackling,it'sthe sameasusingLogisticRegressiontofindadistributionover classes.Themodelisrepresentedbythefollowingformula:

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

Note.Adaptedfrom"Twitter_Sentimental_Analysis_6.jpg" [Image],bypantechsolutions.net-TwitterSentiment AnalysisusingMachineLearningonPython. https://bit.ly/3O3Asj1

RandomForest.(Usedinwhatispresented) RandomForestisajointlearningalgorithmfor classificationandregressiondata.RandomForestgenerates amultitudeofdecisiontreeclassificat aggregatedecisionofthosetrees.Fora 2,...xnandtheirrespectiveopinionla 'bagging',repeatedlyselectsarandomsample(Xb,Yb) withreplacement.Eachclassificationtreefbistrainedusing adifferentrandomsample(Xb,Yb)wherebvariesfrom 1...Finally,amajorityvoteofthepredictionsofthese B-treesistaken.

9.1.2 Predictive Equation

Figure 40

Equationofthepredictivemodel.Itisthetrainingalgorithm forrandom'forestsortrees'thatappliesthegeneral techniqueofbootstrapaggregating,orbagging,for automaticlearning,ormachinelearningasatreetype.

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

Note.Adaptedfromimagefrom Wikipedia,whichisnot obtainableas-name-,[Image].Fromhttps://bit.ly/3tpg6qX &https://bit.ly/3hGNasd

Figure 41

Complexityofthepredictivemodel.Anoptimalbalanceof biasandvariancewouldneveroverfitorbeinappropriate forthemodel.Therefore,understandingbiasandvarianceis criticaltounderstandingthebehaviorofpredictionmodels.

Note.Adaptedfrom"46138669365_b9853 /Image]byhttps://flickr.comathttps:/ Biasand Variance in Machine Learning. https://bit.ly/3UVfRzY

Furthermore,anestimateofthepredictio bemadeasthestandarddeviationofthepre theindividualregressiontreesatx':

Figure 42

predictivemodel,generaltechniqueofboo aggregating,orbagging,forautomaticlea learningasatreetype.

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

Note.AdaptedfromAdaptedfromWikiped isnotobtainableas-name-,[Image].From https://bit.ly/3g26QGk.

Itisgoodtomentionthatthereareothermo suchastheSVM,orsupervectormachines. Xgboostisaformofincrementalgradienta producesapredictionmodelthatisasetofv predictiondecisiontrees.

9.1.3 Neural Networks

MLPorMultilayerPerceptronisaclassofa networks,havingatleastthreelayersofne neuronusesanonlinearactivationfunctio supervisionusingabackpropagationalgo wellforcomplexclassificationproblems, a

9.1.4 Analysis

In this case, through the following piece of software, and as an example, the 'public sentiment' tweets relating to 'US airlines' will be used as a concept to analyze. I classified the tweets into their categories: positive, neutral and negative using machine learning techniques in Python. As in a previous example, but using Python as a concept, and this time with the intention of analysis, using Python as a direct tool to create machin[e] machine learning using different libra[ries] 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 dataset that will be used to train the [...] algorithm will use a file available in *.C[SV] 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_url = "./DM004TweetsAnalysis.csv"
dataset = pd.read_csv(dataset_url, encoding = "utf-8")
dataset.head()
```

Figure 45

Production. The following image shows the first five rows of the dataset.

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. I[...] distribution of positive, negative, and neu[...] data set using a pie chart.

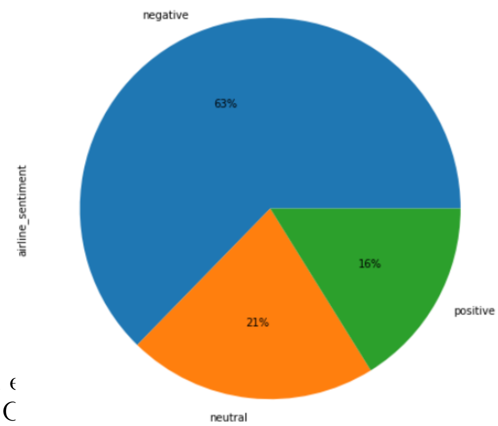
Figure 4.6

Sorting Python in a circular figure as a graph.

```
plt.rcParams["figure.figsize"] = [8,10]
dataset.Type.value_counts().plot(kind='pie', autopct='%1.0f%%')
```

Figure 47

Production. Output of the results as a circ[...]



The result shows that 63% of the general tw[...] negative, while 21% and 16% are respective[...] positive.

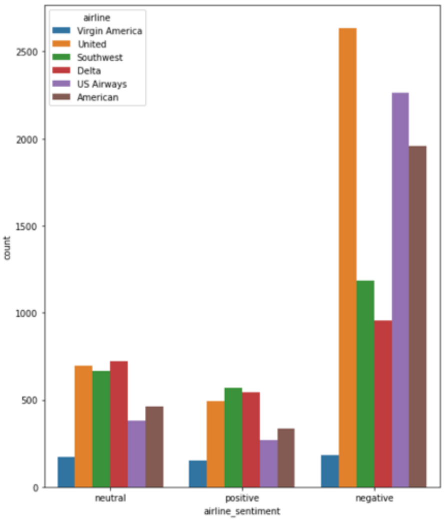
Figure 48

A bar chart is drawn showing the count of ne[...] a

```
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 defi *text_preprocesses() function. The fi* 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 the data into features and tags:

```
y = dataset["airline_sentiment"]

X = dataset["text"]
```

Figure 52

Next, we execute a foreach() loop that itera tweets from tweet list X to the text_prepro that cleans up the tweet text. The followin 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 bas mathematics and mathematics works with necessary to convert text tweets into nume

Figure 53

Although there are several ways to do this, use the TfidfVectorizer class from the *sklearn.feature_extraction.text modu* use the fit_transform() method of the Tfid shown in the following script:

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

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 spec most frequently occurring words to conver in this case. The min_df attribute specifie number of documents in which a word must a Finally, max_df specifies the maximum pr documents in which a word should appear, w the above script. We also remove stop word we, at, as they don't provide much informat

9.1.8 Division of Data into Training and Test Sets

Machine learning algorithms are trained o e

Figure 54

To split the data into training and tests *train_test_split() method of the sklearn* 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 RandomForest classifier will be used, since it is the most robust. To use the RandomForest 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 RandomForestClassifier 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 sklearn.metrics module to find the values of these metrics as shown in the following script:

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

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 results show that the algorithm, and t software, can efficiently and accurately 'positive, negative or neutral' with an ov 75.92%.

10. Conclusions and Recommendations

10.1 Subject of Study - Importance

The new marketing trends for companies ma social networks as means of communicati mainly due to their great potential to estab with customers through them. Social networks also facilitate corporate effort s to convey brand values, attract new dissemin ate marketing information or obt results from dissatisfied customers. According to Pak and Paroubek (2010), soc *The following script* have been a useful tool to simplify communi customers.

It is important to know the purpose of each s format and the target audience and to be on networks where the audience and their targ the brand.

10.1.1 General Objective of the Investigation

This research aims to measure the sentimen social network Twitter in relation to how, b concepts, a psychographic model and its co predictive models can be understood.

To do this, the research is based on data fro profiles of users, organizations, compani obtained using the Twitter API.
Once the tweets are downloaded, a develop powered by machine learning, using Pytho split the sample (n = X tweets) into negativ positive sentiments.

10.1.2 Main Points of the Investigation

Sentiment Analysis with Machine Learni
Sentiment analysis is defined as the proces opinions in terms of qualifications, attitu about a particular topic. (Fiorini and Lips

Sentiment analysis generally serves two p expressions of sentiment and definition of orientation by individuals. (Honeycuttar Saura, 2018). Sentiment analysis makes it the positive, negative and neutral express t

textualelement.(Boyd,2017;Chunga,2017)

Debes(2017),indicatesthatsentimentanalysiscanreferto approachesandbebasedoncharacteristicsandautomatic tags,conversations,oritcanbethecaseofcommontagsin athemeorspecificeventsofuse,inemoticonsorin resourcessuchaslexicons.ofsentimentthatiswith positive,neutralornegativetweets.Fundamentallexicons labelwordscollectedinasemanticallynecessary dimension,called"feeling","valence"or"semantic orientation"(Saura,Palos-Sanchez/Cerdá,2017).

AlgorithmsdevelopedinPythontoperformsentiment analysishavepredictivepower. Thepredictionisdeterminedbymachinelearning.Machine learningisaformofartificialintelligencethattrainsthe virtualmachinethroughdataexplorationtoautomatethe dataanalysisprocess,amongotherfeatures.

10.1.3 Objectives, Achievements and Formats

Afterthedevelopmentofthemethodologicalprocessthat includestheanalysisandextractionofdata,the psychographicsentimentanalysiswascarriedoutinthis investigation,asaresultoftheanalysis,itwaspossibleto obtaintheaverage,throughtheapplicationofthealgorithm withautomaticlearning.

Inthetableoroutputintheterminal,ofthetestsandresults *(PartII),theydeliveredtheresultsoftheanalysisprocessof* processedtweetsonthesentimentofadeterminednumber ofusersregarding6UnitedStatesairlines,inadditiontothe useoftheinteractionsbythecompaniesandtheinteractions andcommentsmadebytheusersaboutthem,the categorizationmadeaccordingtosentimentandtheaverage veracityobtainedasaresultofthemachinelearning algorithm.Asspecifiedabove,itwaspossibletoefficiently andaccuratelyclassifyatweetas'positive,negativeor neutral'withatotalaccuracyof75.92%.

10.1.4 Feasibility And Potential Of The Research And The Project

Themainengineoftheapplicationusedismachine learning,withrepeateduseandtrainingofaverageresults increases.
Asithasbeenshown,Twitterisconfiguredastheoptimal socialnetworkforuserswhocanexpresstheirfeelings, opinionsandcommentsinaspecificway,inrealtimeand allovertheworld.
Twitterhasbeenusedasanobjectofstudyinmultiple periodsduringthelastdecade.
Inthisresearch,Twitterwasusedtoprioritizethesentiment oftheoffers,andthereforethequality,ofthecompaniesthat makeupthesamplewhentheypublishonsaidsocial network.
Theresultsoftheresearchcanbeusedbycompaniesto improvethedevelopmentoftheirstrategiesinsocial networksand,morespecifically,aroundsocialTwitter.

Additionally,thesearchresultsidentifycommunicationsand offersviaTwitterforothersrelatedtothedataanalyzed,in termsoftone,andwhatbusinessescantakeadvantageofon socialmedia.
ThisresearchprovidesverifieddataonTwitteractionsthat canbeusedforfuturemarketingstrategiesandwillserveas asourceofresearchonthespecifiedtopic.

10.1.5 Recommendations for Further Investigation

oWhyshoulditbedone?

Withoutunderestimatingthevalueofqua
sentimentandinterestinestablishingmor
audiencetowhichabrandisdirectedisimp
allowsustoanalyzetheemotionalrespons
haveinteractedwithit.Intoday'sdigital
peoplehavemoretoolstoexpresstheiropi
doubtsandexpresstheirconcerns,whethe
negative.
Thatiswhysentimentanalysisispostulat
allowsobtaininginformationabouttheto
speakonsocialnetworks,orforumsofacer
somethingthatallowspeopletoknowwhat
whattheylike,not.

oWhatthingsshouldbedone?

Sentimentanalysisisaprocessinwhichma
ormachinelearningisusedtofindaspoton
sentiment,andplacetheinformationofin
process.Thefeelingcanbedefinedasthere
judgmentorjudgmentformedaboutsomet
necessarilyfactsorknowledge".Butwith
sentimentanalysisanddatascience,opini
understood.
Itisasubjectiveevaluationofsomethingb
empiricalexperience.
Itismadeupofobjectivefactsandpartlyby
opinioncanbeinterpretedasadimensioni
particulartopic.
Itisasetofsignifiersthat,combined,pres
is,afeelingaboutaparticularsubject.

Sentimentanalysisalgorithms
Therearetwomainmethodsofsentimentan
[U+25AA] Rules-basedapproach

Rule-basedsentimentanalysisisbasedon
withaclearlydefineddescriptionofanide
sentiment.
[U+25AA] Itincludesidentificationofsubj
themeofopinion.
Therule-basedapproachinvolvesabasic
processingroutine.

Isthathowitworks:
Therearetwolistsofwords.
Oneofthemonlyincludespositive,neutra
othernegativepoints.
Thealgorithmscansthecontent,findswor
criteria.
Afterthat,thealgorithmcalculateswhatt
mostfrequentinthecontent.
Iftherearemorepositivewords,thetextis
haveapositivepolarity.

oWhowillitbenefit?

Sentimentanalysisorpsychographicanal
perceptionofusersaroundideasorproduc
understandingofthemarket,throughthep
datathatcanbefoundalreadyregisteredb
companies.insocialnetworksoringenera
Therearemanypublicandprivatesources
fromwhichyoucanobtaininformationabo

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.

o 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 interpretation 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 customers 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 no
  prepare the data for processing.
- Data processing.
- Process data applying statistica
  analysis, machine learning, predicti
- Data visualization.
- Represent data in different ways
  understandable in the most accurate

o Where will it be done?

These technology implementations are car systems controlled and ideally managed by technology and data analysis companies. T constantly researched, so it is common and these companies to carry out a lot of researc the delivery of services, development of te products for itself and/or consulting.

11. General Conclusions[9]

Undoubtedly, the approach of the problem investigation purposes an important step development, allowing to establish who wi consequently, which ones will be resolved.
The approach to the problem, as well as the c research, the starting point of this topic, u done in the previous topic, establishing cl goals, the initial direction that was giv investigation.
This occurs since the investigation part pr construct, an orderly and coherent reflec logical transition of actions and objectiv detected problem.
A custom software-level product has a posi effect on the processes of the organization wher implemented.
This happens because it was designed and b particular needs of those who need it.
For this impact to be truly positive and gen development must be carried out in an orde adequate standards and good practices tha compliance with the agreed deadlines, the and that it works for what it was created for the requirements the client and the proces

The development life cycle of an informati application or product is nothing more tha clearly defined sequence of steps to consid developing a software or information tech

Without any doubt, the development of a pr of the hardware or software of a system, is o

[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 methodolo Autonomous University of Tabasco.

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

Malhotra, N. (2004). Market research: an appli Mexico: Pearson Education.

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

Pardinas, F. (2005). Methodology and techniqu research. 38th edition. Mexico City, Mexico: Si

American Psychological Association (2002). P the American Psychological Association. Mexi Manual. Apa Standards (2019). Apa Standards 2 https://normasapa.com/

Rodriguez, E. (2005). Investigation methodolo the Autonomous Juárez University of Tabasco. S (2014). Apa Standards – 7th (Seventh) Edition.

Other use and research references

Big Data with Python – Collection, storage and p E. Martin & A. Riesco. (Personal book).

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

Descriptive statistics, probabilities – Inferen non-parametric methods. By Pedro Vergara Ver

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

BULUT, A. (2015). Lean Marketing: Know who no Electronic Commerce Research and Applications.

CHUNGA, A., ANDREEVA, P., BENYOUCEF, M. P. (2017). Managing an organization's social media presence: An empi International Journal of Information Management.

YOUMUST, V., SANDEEP, K. AND VINNETT, G. information diffusion Probabilities in social networks: A Bayesian ne Journal of Knowledge-Based Systems.

FAGAN, J.C. (2014). The Suitability of Web Analy Indicators in the Academic Library Environment. The Journal of

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

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

HERRÁEZ, B., BUSTAMANTE, D. AND SAURA, classification on social networks. Content analysis of e-commerce comp

HONEYCUTT, C., & HERRING, S.C (2009). Beyo Conversation and collaboration via Twitter. In 42nd Hawaii Inter System Sciences.

JÄRVINEN, J., AND KARJALUOTO, H. (2015). for digital marketing performance measurement. Industrial Marketi

LEEFLANG, P, VERHOEF, P., DAHSLTRÖM, P. & FREUNDT, T. (2014). Challengesandsolutions formarketinginadigitalera. EuropeanManagementJournal.

MATHEWS, S., BIANCHI, C., PERKS, KJ, HEALY, M., AND WICKRAMASEKERA, R. (2016). Internetmarketingcapabilitiesandinternationalmarketgrowth. InternationalBusiness Review.

MCDERMOTT, J. (2017, December03). BlackFridaystats: Thenumbers behindthemadness. RetrievedApril18, 2018.

PAK, A., & PAROUBEK, P. Twitterasacorpusforsentimentanalysisand opinionmining. In ProceedingsofLREC, 2010. Valletta, Malta.

PALOS-SANCHEZ, P.; SAURA, JR (2018). TheEffectofInternetSearches onAfforestation: TheCaseofaGreenSearchEngine.

SAURA, JR, PALOS-SÁNCHEZ, P., & CERDÁ SUÁREZ, LM (2017). Understandingthe DigitalMarketingEnvironmentwithKPIsandWebAnalytics. Future Internet, 9(4), 76.

SAURA, JR, PALOS-SANCHEZ, PR & RIOSMARTIN, MA (2018). attitudesto environmentalfactorsinthetourismsectorexpressedinonlinecomments: Anexploratory study. InternationalJournalofEnvironmentalResearchandPublicHealth.

TT, KUO, S.-C. HUNG, W.-S. LIN, N. PENG, S.-D. LINANDWFLIN (2012). Exploitinglatent informationtopredictdiffusionsofnoveltopicsonsocialnetworks, in: Proceedingsofthe 50thAnnualMeetingoftheAssociationforComputationalLinguistics: ShortPapers-Volume 2, AssociationforComputationalLinguistics.

Algorithmicfuture. Bigdata, sensordataandmobilemedia, Co-authors: JoseCorreaandCharlesThraves.

SortAlgorithmsandDatastructures, byNchenaLinos.

DataStructuresandAlgorithmswithJavaScript, byWahyuPrayogo. JS& DataStructures.

ProgrammingProblems: AdvancedAlgorithms(Volume2) Paperback– February27, 2013, byGuidoNotoLaDiega, PhD.

Superintelligence: Paths, Dangers, StrategiesReprintEdition, byNick Bostrom(Author), PhD.