A Study of Data Analysis Techniques. Application to Personal Data Analysis in the Context of Social Networks

Autor: Ignacio de Castro Pérez
Tutor: Juan Carlos Yelmo García
Tribunal:
Juan Carlos Yelmo García
Miguel Ángel de Miguel Cabello
José María del Álamo Ramiro

Fecha de Lectura:
Calificación:
A Study of Data Analysis Techniques. Application to Personal Data Analysis in the Context of Social Networks

Autor: Ignacio de Castro Pérez
Tutor: Juan Carlos Yelmo García
Tribunal:
Juan Carlos Yelmo García
Miguel Ángel de Miguel Cabello
José María del Álamo Ramiro
**Summary**

In recent years, society is experimenting various changes. One of these changes is *datafication*. This term can be defined as the systematic transformation of everyday aspects of people’s lives into computerised data. Every day, every minute and every second, any time that someone uses a digital device, data is being stored somewhere. It may be the content of an email but it may also be the number of steps he or she walks every day or his or her medical record.

The simple storage of data does not give any added value by itself. To extract knowledge from data, and therefore to add value to it, data analysis is needed. Data science along with data analysis is becoming increasingly popular. Nowadays, millions of statistical web APIs can be found online; these APIs offer the possibility to analyse trends or sentiments present in social networks or the Internet in general.

One of the most popular social networks, Twitter, is public. Every message, or tweet, that gets published can be accessed by anyone with an Internet connection around the globe. This makes Twitter an interesting media to analyse social habits or consumer profiles. This is the context in which this work is enclosed.

This work, with the combination of statistical data analysis and content analysis, aims at extracting knowledge from public tweets from Twitter. In particular, it aims at establishing if gender is an influential factor in inter-user relationships in Twitter. To determine it, a database containing nearly 2,000 tweets will be analysed. Firstly, the gender of the users in the database will be determined using web APIs. Secondly hypothesis testing will inform if gender influences users when communicating with other users. Lastly, a statistical model will be built to predict the behaviour of Twitter users in regard to their gender.

**Key words:**

R, RStudio, Twitter, Data science, data analysis
Resumen

En los últimos años la sociedad está experimentando una serie de cambios. Uno de estos cambios es la *datificación* (“datafication” en inglés). Este término puede ser definido como la transformación sistemática de aspectos de la vida cotidiana de las personas en datos procesados por ordenadores. Cada día, a cada minuto y a cada segundo, cada vez que alguien emplea un dispositivo digital, hay datos siendo guardados en algún lugar. Se puede tratar del contenido de un correo electrónico pero también puede ser el número de pasos que esa persona ha caminado o su historial médico.

El simple almacenamiento de datos no proporciona un valor añadido por si solo. Para extraer conocimiento de los datos, y por tanto darles un valor, se requiere del análisis de datos. La ciencia de los datos junto con el análisis de datos se está volviendo cada vez más popular. Hoy en día, se pueden encontrar millones de web APIs estadísticas; estas APIs ofrecen la posibilidad de analizar tendencias o sentimientos presentes en las redes sociales o en internet en general.

Una de las redes sociales más populares, Twitter, es pública. Cada mensaje, o *tweet*, publicado puede ser visto por cualquier persona en el mundo, siempre y cuando posea una conexión a internet. Esto hace de Twitter un medio interesante a la hora de analizar hábitos sociales o perfiles de consumo. Es en este contexto en que se engloba este proyecto.

Este trabajo, combinando el análisis estadístico de datos y el análisis de contenido, trata de extraer conocimiento de *tweets* públicos de Twitter. En particular tratará de establecer si el género es un factor influyente en las relaciones entre usuarios de Twitter. Para ello, se analizará una base de datos que contiene casi 2.000 *tweets*. En primer lugar se determinará el género de los usuarios mediante web APIs. En segundo lugar se empleará el contraste de hipótesis para saber si el género influye en los usuarios a la hora de relacionarse con otros usuarios. Finalmente se construirá un modelo estadístico para predecir el comportamiento de los usuarios de Twitter en relación a su género.

**Palabras clave:**

R, RStudio, Twitter, Ciencia de los datos, análisis de datos
Index

1 Introduction .................................................................................. 7
   1.1 Specific goals ........................................................................ 7

2 State of the art ............................................................................. 8
   2.1 Data Science .......................................................................... 8
      2.1.1 Data Science in Gartner’s Hype Cycle .......................... 9
      2.1.2 Statistical Data Analysis ............................................ 10
         2.1.2.1 Data Collection ................................................ 11
         2.1.2.2 Data Processing .............................................. 11
         2.1.2.3 Data Cleaning ................................................. 11
      2.1.2.4 Exploratory Data Analysis .................................... 11
      2.1.2.5 Hypothesis testing ............................................. 12
         2.1.2.5.1 Pearson’s Chi-Squared Test ............................ 13
      2.1.2.6 Statistical Models (Linear Classification Models) ...... 13
         2.1.2.6.1 Logistic Regression ..................................... 14
      2.1.3 Content Analysis .......................................................... 15
   2.2 Personal and social data ...................................................... 15
      2.2.1 What is it? ........................................................................ 15
      2.2.2 Economic value of personal data .................................. 16
      2.2.3 Social networks services APIs ................................. 17
         2.2.3.1 gettwitterid.com API ...................................... 17
         2.2.3.2 AI-Applied Demographics API ............................ 17
         2.2.3.3 NamSor Origin API ........................................ 17
   2.3 Data Analysis support tools ................................................. 17
      2.3.1 R or Python? ................................................................ 18
      2.3.2 R and RStudio .......................................................... 18

3 Data processing ........................................................................ 20
   3.1 Terminology .......................................................................... 20
   3.2 Starting point .......................................................................... 20
   3.3 Authors characterisation ..................................................... 21
   3.4 Gender determination .......................................................... 22
      3.4.1 Authors’ gender determination .................................... 22
      3.4.2 Retweeters’ and mentioned users’ gender determination .................................................. 23
   3.5 Final processing ..................................................................... 24

4 Data Analysis ........................................................................... 25
   4.1 Exploratory data analysis .................................................... 25
      4.1.1 Gender distribution .................................................. 25
      4.1.2 Retweeters and authors relation .................................. 26
      4.1.3 Mentioned users and authors relation .......................... 28
   4.2 Pearson’s Chi-squared test ................................................. 29
      4.2.1 Retweeters and authors relation .................................. 30
      4.2.2 Mentioned users and authors relation .......................... 30
   4.3 Followers and influence ..................................................... 31
      4.3.1 Retweeters and authors relation .................................. 31
      4.3.2 Mentioned users and authors relation .......................... 32

5 Data Analysis – Revisited ......................................................... 34
   5.1 Exploratory data analysis ..................................................... 34
      5.1.1 Gender distribution .................................................. 34
      5.1.2 Retweeters and authors relation .................................. 35
      5.1.3 Mentioned users and authors relation .......................... 36
   5.2 Pearson’s Chi-squared test ................................................. 37
      5.2.1 Retweeters and authors relation .................................. 37
1 Introduction

Social networks and other online services users provide, intentionally or inadvertently, large amounts of personal information that can be used to determine consumption or social habits, location and contact data, financial solvency profiles, etc. Data can also be added to determine correlations that allow us to deduce trends in business, social impact of television programs, broadcasting of epidemics, etc. To the potential advantage of these investigative techniques we must note the rights of citizens in relation to the protection and control of the use of their personal data.

This work will focus in analysing different techniques for personal data analysis in the context of a social network (Twitter) to determine if gender is an influent factor in inter-user relations.

A relation in Twitter is defined as either mentioning another user or retweeting another user’s tweet. This work will try to infer if gender is influencing users at the time to interact with other users.

1.1 Specific goals

These are the specific goals in this work:

- General study of techniques and methodologies for content and data analysis for its understanding and use. A general framework and a brief taxonomy and characterisation will be done for each of the major techniques and methodologies.
- Study of the techniques used in statistical analysis: collection, analysis and inference.
- Learning of the R language in the RStudio environment for statistical data analysis.
- Learning and use of web APIs to fully characterise social networks users and their personal data.
- Implementation of the framework and tool environment for analysing personal data in social networks. In particular, I will work with a dataset of messages from the social network Twitter to study the relationship between mentioned users, retweeters and authors or tweets to determine if gender is an influencing factor in inter-user relations.
2 State of the art

2.1 Data Science

Data science is one of the emerging fields of the 21st century. It is also a discipline with increasing professional demand. [1]

It can be defined as the intersection of three disciplines: Mathematics and Statistics, Computer Science and Domain Expertise. Graphically:

Data science aims to systematically extract knowledge from data. To do so, data scientists apply data analytics, modelling and inference (which lie within the statistics domain) but also take advantage of the “datafication” that society enjoys nowadays; and this is strictly related to computer science.

Traditionally, in order to carry a statistical study a statistician would have to collect relevant data to that study. In the 21st century society, statisticians do not need to do this data collection; it is automatically done by computers when we do anything. From Google searches to clicking the “Like” button at your best friend’s Facebook account, these actions are electronically recorded. Furthermore, traditional data sets used to be an ensemble of variables – either numerical or categorical. Today, data science analyses a much wider variety of data such as images, Geo location data or raw text – e.g. social networks’ profile pictures or tweets and its localisation.

Therefore, we can think of data science as a means to predict and not only explain. Data analysis is generally used to explain a problem or phenomenon; on the contrary the aim of data science is to extract conclusions that allow us to predict future trends, make decisions and ultimately make money.
All this is possible thanks to the low cost of hard drives and high power of computers. Firstly, during the dot-com bubble [5] hard drives became cheap and so corporations and governments were able to afford storing large quantities of data. Secondly, analysing this data would have required calculations that would have taken months or even years to do by hand but with millions of operations performed per second inside the CPU of a computer these tasks have become within the reach of any standard person that knows how to program.

Finally, data scientists need the knowledge of the domain to which the data belong, e.g. you cannot carry a study on average annual precipitation in the UK without basic notions of meteorology [3] [4] [6].

### 2.1.1 Data Science in Gartner’s Hype Cycle

Gartner [7], an information technology research company, publishes annually a Hype Cycle report. This cycle is an idealised model of a standard technology progress, from ‘Innovation Trigger’ to ‘Plateau of Productivity’ - i.e. the moment when at least 30% of the technology’s target audience has adopted it. Gartner also provides their estimate for the time until this plateau is reached.

This is the most recent Hype Cycle (2014):

![Gartner's 2014 Hype Cycle for Emerging Technologies](image_url)
We can see how data science is about to enter the peak of inflated expectations. According to Gartner, this peak is the moment when first-generation products at a high price will start to appear. Still according to Gartner, data science will still have to suffer a trough of disillusionment before eventually reaching the plateau (a moment when Gartner expects third-generation products with services to exist). The good news is that Gartner estimates that the plateau will be reached between 2 and 5 years, a quick growth if we take into account that all of the technologies that surround data science in the Hype are estimated to take from 5 to 10 years (or even more than 10 in some cases).

2.1.2 Statistical Data Analysis
The unstructured massive recollection of data, known as Big Data is characterised by the 3 V’s: Velocity, Volume and Variety. This data, by itself does not provide any meaning or information. It is only when data analysis is performed over this data that useful information may be extracted. The following sections are aimed at describing the statistical data analysis process.

Statistical data analysis is the process of inspecting, cleaning, visualising, transforming and modelling data with the objective of extracting useful information, making decisions or suggesting conclusions. The process can be separated into several phases which can be iterative – i.e. feedback from later phases can mean enhanced data for the earlier ones.

![Data analysis phases](image-url)


2.1.2.1 Data Collection
Data are collected from a variety of sources that will depend on the nature of the analysis. Therefore, data may be collected by people (e.g. an interview), sensors, computers etc.

Generally, data scientists designing a study/experiment will not have access to the entire population of interest and they will have to sample. For a study/experiment to be meaningful the conclusions extracted from the sample must be generalizable to the whole population of interest; if this is to occur then the sample has to accurately represent this population.

There exist multiple sampling methods but this is out of the scope of this project. [9]

2.1.2.2 Data Processing
Raw data must first be processed before it can be analysed to extract information – e.g. raw data may be unformatted and processing would mean sorting and organising it into columns and rows. [9]

2.1.2.3 Data Cleaning
Once the data is processed, it must be cleaned. Data cleaning involves looking for duplicates, errors or missing values.

This process may include filling up blanks in the data with several techniques or looking for impossible outliers. An important distinction must be made here; statistically significant outliers must never be replaced or deleted, only values that make no sense or are clear errors should be treated this way– e.g. a negative value in an age field.

Data may be cleaned by comparison with a validated data set, with specialised software or by hand. Cleaning may be strict (such as rejecting anything outside defined bounds) or fuzzy (such as correcting errors when possible). [9]

2.1.2.4 Exploratory Data Analysis
Exploratory data analysis (EDA) refers to a series of techniques (mostly visual though not limited to that) used to analyse clean data. It may lead to further cleaning or recollection, making it an iterative process as I mentioned before. They were originally promoted by John Tukey at the Bell Labs to encourage statisticians to explore data. Being graphical in nature, EDA triggered the development of statistical computing languages like S, precursor of R, which is used in this work.

Tukey also promoted the use of what he called “five number summary” –i.e. the maximum, minimum, median, upper and lower quartiles instead of mean and
standard deviation as the latter are not defined for every distribution and the former are more robust to skewed distributions.

EDA, along with the development of S, S-PLUS and R facilitated the work of statisticians in fields like engineering, semiconductors and communications networks (all of which concerned the Bell Labs) and complement the traditional analytic theory of statistical hypotheses and tests. [9]

2.1.2.5 Hypothesis testing

One of the most basic kind of statistical inference is hypothesis testing. A statistical hypothesis can be tested by modelling it using random variables. This is called a hypothesis test and it is usually the first thing a data scientist will try after exploratory data analysis.

To carry a hypothesis test we have to state $H_0$ (null hypothesis) and $H_A$ (alternative hypothesis). Only if there is strong evidence against the null hypothesis will this be rejected, usually the threshold probability is held at 95%; this means that we have to be 95% sure before rejecting the null hypothesis, which is considered to represent the status quo.

If the null hypothesis is rejected we say the result of the test is statistically significant meaning that the observations in the sample are unlikely to have occurred simply due to chance.

These definitions help to understand statistical hypothesis testing:

- P-value: The probability of observing an outcome at least as extreme as the one in the sample, giving that the null hypothesis is true.
- Significance level: This is the threshold for deciding if a test is statistically significant or not. If the p-value is smaller than the significance level then we decide to consider the test statistically significant. It is usually set at 5%.
- Type I error: This is rejecting the null hypothesis when it is true.
- Type II error: This is failing to reject the null hypothesis when the alternative hypothesis is true.

Therefore, a statistical hypothesis test would consist in calculating a p-value for a given sample, comparing it to the chosen significance level and deciding to reject or failing to reject the null hypothesis.

Hypothesis testing relies heavily on the chosen value of significance level. This value is either selected by convention or chosen by the data scientist; in the first it may be judged as mindless or random and in the latter it may be biased.
Therefore, hypothesis testing should never be the only method of testing for the veracity of the alternative hypothesis.

It is also important to understand that the conclusion of the test is only as solid as the sample in which it is based. If the sample is biased then the result of the test will be equally biased. This makes the design of the experiment vital if a solid conclusion is wanted. [10]

2.1.2.5.1 Pearson’s Chi-Squared Test
This test is carried later on in this work. Pearson’s chi-squared test is used to test for independence between categorical variables. The test statistic ($\chi^2$) is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where:
- $\chi^2$ is Pearson’s cumulative test statistic which asymptotically approaches a $\chi^2$ distribution
- $O_i$ is the number of observations in type $i$
- $E_i$ is the theoretical frequency of type $i$
- $n$ is the number of cells in the contingency table

A chi-squared test assesses if the frequency distribution of an event in an observed sample is consistent with a particular theoretical distribution—i.e. the distribution we would expect if the variables of the event were, in fact, independent. More on this topic may be found in reference [11].

2.1.2.6 Statistical Models (Linear Classification Models)
In order to look for relations in the data such as correlations, data scientists apply mathematical formulae or algorithms to the data. Generally:

$$Data = Model + Error$$

A statistical model is made of a series of assumptions regarding the creation of the observed data, therefore representing in an idealised form the data-generating process.

Models may be used to predict future outcomes of an event. For example, we may want to know if a higher grade in course $X$ would also mean a better grade in course $Y$, mathematically speaking: $Y = aX + b + \varepsilon$. We would have to define parameters $a$ and $b$ to minimise the error $\varepsilon$ for a given range of possible grades in course $X$. 
Statistical models are a particular case of mathematical models. The particularity in statistical models is that theses are non-deterministic; thus in a statistical models some variables may not have a specific value but a probability distribution.

Linear classification models are a type of statistical models. This models aim to estimate the probability that an item belongs to a certain category using linear regression, thus they are related to deciding the value of a categorical value. Any model will use some variables in the data set to predict the value of the target variable. These variables are called predictors.

### 2.1.2.6.1 Logistic Regression

The linear classification model used in this work is logistic regression (also known as logit regression). This model is a particular case of linear classification models where the outcome variable is binary. Mathematically, we want to know the probability of success for our variable (we will call this probability \( p \)):

\[
p = c_0 + c_1x_1 + c_2x_2 + \cdots + c_nx_n
\]

The problem with this equation is that, theoretically, the value of \( p \) is unconstrained; we know this is impossible, as probabilities range from 0 to 1. To fix this, we need to choose a function for \( p \) that varies from 0 to 1 for every possible predictor value. The transformation used in logit regression is as follows:

\[
\text{logit}(p) = \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_n
\]

In terms of probabilities:

\[
p = \frac{\exp(\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_n)}{1 + \exp(\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_n)}
\]

Where:
- \( x_1, x_2, \ldots, x_n \) are a set of predictor variables
- \( \beta_1, \beta_2, \ldots, \beta_n \) are the parameter values estimated using logistic regression of \( p \) on \( x_1, x_2, \ldots, x_n \)
- \( \exp(x) \) is equivalent to \( e^x \)

As we can see, \( p \) will never be outside the 0 to 1 range, as desired.

[12] [13] [14]
2.1.3 Content Analysis
In contrast with quantitative data analysis, this field focuses in classifying attitudes, feelings or behaviours. It is not based in the numbers but on the content of what it is analysing. For example, in this project we are working with tweets; statistical data analysis will help us do models or hypothesis tests, qualitative data analysis will help us decide if a user is male or female. Therefore, there is an interpreting human factor that cannot be substituted by mathematics or machines.

To perform content analysis, also known as qualitative data analysis, there must be two or more people rating whatever it is they are analysing. In our example there would be two or more raters deciding if the author of a tweet is male or female.

Therefore, content analysis refers to methods for studying and retrieving useful information from documents; we have to understand that documents in the computer science and telecommunication days has a broader definition than the traditional one. Documents will contain text that comes from communication processes –i.e. types of communication intentionally activated by a sender, using a code sufficiently shared with a receiver. There are 5 types of texts in content analysis:

- **Written text**: books, newspapers, etc.
- **Oral text**: speech, etc.
- **Iconic text**: drawings
- **Visual text**: TV programs, films, etc.
- **Hypertext**: one or more of the above, on the Internet

As there are various raters coding a set of units of analysis there is a need to find a measure of agreement between the raters. One of the most popular measures amongst content analysis is Krippendorff’s alpha. It generalises several other known statistics used in inter-coder agreement (like Cohen’s kappa, which will be used in this work) but it is also different from other statistics that are unsuitable to the particulars of coding data generated for subsequent analysis. [17] [18]

2.2 Personal and social data

2.2.1 What is it?
With the actual datafication that society is experimenting, the digital record of our behaviour is becoming increasingly omnipresent. Nearly every decision that the average citizen makes may and will be recorded in some digital means. Personal data can be considered as the digital record of everything that a person makes and does in the Internet and in the world. There some categories into which we can divide this data:
• **Digital identity:** names, email addresses, physical addresses, phone numbers, social networks profiles information, phone number, etc.

• **Health data:** medical history, prescriptions, etc.

• **Institutional data:** academic, employer and governmental data

Organisations or companies may capture this data in several different ways:

• **Voluntarily:** individuals may voluntarily share their data – e.g. when creating a social network profile.

• **Observed:** data may be *captured* when users navigate online – e.g. when selecting preferences for a Google search or with location data in mobile phones

• **Inferred:** Organisations may infer data using data analysis – e.g. inferring the gender of a social network user based on their texts

The decision of analysing and using personal data is today subject of controversy. Some developed countries, especially European Union members, have strict privacy laws that make it difficult to exploit this personal data. Even so, data continues to be stored, maybe permanently and may (or may not) be analysed in the future with different laws to those that exist today.

[19]

2.2.2 Economic value of personal data

By themselves, technology and data are neutral. It is when we apply data analysis on them that we can both generate great value and enormous harm. Storing and analysing big amounts or personal data can bring personal benefits to the owner of the data. For example, healthcare data can be utilised to research new ways of increasing patient care, ultimately saving lives. The opposite may also happen, in the wrong hands this data can be used to harm the patient.

Big Internet actors like Google or Facebook store huge amounts of personal data from nearly every individual. The analysis of this data can be used to determine consuming trends which can later be sold to other companies. This would result in better-adapted goods and services to the consumer needs. Apparently everyone gains something in this model. The problem today with this is governments’ reluctance to modify privacy-protecting laws.

These controversies are being addressed nowadays by governments, the Organisation for Economic Co-operation and Development (OECD) or the World Economic Forum (WEF).

[20]
2.2.3 Social networks services APIs
API is an abbreviation for Application Programming Interface. An API is a set of routines, protocols and tools used for building software applications. In this work only Web APIs have been used. A Web API is the defined interface through which interactions happen between an enterprise and applications using its assets. An API is typically defined as a set of HTTP request messages, along with a definition of the structure of response messages, which is usually in an Extensible Markup Language (XML) or JavaScript Object Notation (JSON) format.

The next sections describe all of the APIs that have been used in this work. [21]

2.2.3.1 gettwitterid.com API
This API can be found at http://gettwitterid.com. This is a very simple API that characterises Twitter users – i.e. it is called via a GET HTTP request method, providing a Twitter’s username (screen name) and it returns the user’s ID, number of followers, number of status and his or her full name. This API is much simpler than the REST Twitter APIs. The response comes in html format. A use example can be seen in section 3.3.

2.2.3.2 AI-Applied Demographics API
This API can be found at http://ai-applied.nl/demographics-api. It estimates the gender of the writer of any text message, using only the text itself. The API can estimate gender as "male", "female" or "unknown" (for organisations) with around 65% accuracy (note that the data set was pre-processed so that it did not contain any tweets from organisations). The accuracy of the API increases to a 75% if you provide the username or full name of the text’s author. The API is accessed with a JSON-based RPC formatted call via GET HTTP request method. The response comes in JSON format. A use example may be found in section 3.4.1.

2.2.3.3 NamSor Origin API
This API can be found at http://blog.namsor.com/api. It estimates the gender of a user given its first name and surname. It analyses the first name in the cultural context of the surname in order to decide the gender of the user. It estimates the gender as either “female” or “male”. For most Western cultures the average success rate is above 95%. The API is called through a GET HTTP request method and the response is in JSON format. A use example may be found in section 3.4.2.

2.3 Data Analysis support tools
As noted in previous sections, computers and technology as a whole are contributing to the development of data science. Capturing and storing data is increasingly simple and cheap but once the data has been collected something has to be done with it. To extract valuable information from this data we need data analysis and to actually perform this analysis we need tools.
To perform statistical analysis any general-purpose programming language or spreadsheet application can be used. Even so, there is specialised software for statistical computing.

The advantage of using specialised software over a general-purpose language derives from the fact that the specialised software will have functions (or methods) that automatically perform tasks that would have to be manually coded if written in a general-purpose language.

Examples of specialised software include R and SAS.

2.3.1 R or Python?

Two of the most commonly used languages are Python and R. The former is a general-purpose language and the latter, as it has been said, is specialised.

There is split views on what is the most suitable language for data analysis among data scientists. This section is not aimed at addressing that discussion but instead giving some views on the main characteristics of both languages.

Although Python is a general-purpose language, it has a number of platforms (cipy, Numpy or Pandas) that are specifically focused in statistical analysis. The advantage of using Python comes from the fact that it is general-purpose. If a task that involves complicated coding has to be done, using Python may expedite the task. There is also general agreement that Python libraries are not as extensive as R’s.

R’s strength comes from the widespread support of the statistical community. This means that there is an already-written package that will perform complicated mathematical or statistical tasks. That fact will make the programmer save time thinking about how to code a particular task.

2.3.2 R and RStudio

All of the work in this project has been done with R in RStudio.

R is a programming language and software environment for statistical computing and graphics. It is widely used amongst data scientists to develop statistical software and data analysis. It is an implementation of the S language, originally created by John Chambers at the Bell Labs. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. R is free and open-source; it is a GNU Project and is freely available under the GNU General Public Licence.

R’s capabilities are extended through an enormous number of user-created libraries, called packages in the R lingo, that allow specialised statistical techniques, import/export capabilities or data processing techniques. R has a large, active community of users many of which are in academia; additionally mailing lists (http://www.r-project.org/mail.html) provide access
to package authors or users that are experts in the field. When a user creates a package it readily becomes available to the rest of users at the Comprehensive R Archive Network (CRAN) which is an R repository.

*R* is an interpreted language, accessed through a command-line interpreter – e.g. if the user writes $3+2$ at the *R* command window and presses enter, the computer automatically replies with $5$. Like *MATLAB*, *R* supports matrix arithmetic. *R* is an object-oriented language; data structures (objects) include vectors, matrices, data frames and lists. Data frames are the main structure used to store information. They are similar to tables in a relational database.

The availability of these structures facilitates programming as they align closely with the natural form in which data come or is usually arranged.

Finally, graphics and data visualisation is one of *R*’s strengths. One of the principles in the design of *R*, probably derived from its predecessor *S* and Chambers at the Bell Labs, is that visualisation of data through charts and graphs is an essential part of the data analysis process. As a result, *R* includes powerful tools for creating graphics.

*RStudio* is a free, open source integrated development environment (IDE) for *R*. It is the most commonly used IDE. There are two editions, *RStudio Desktop* and *RStudio Server*. The former runs locally as a regular desktop application while the latter may be accessed through a web browser while it is running on a remote Linux server. There are distributions of *RStudio* available for *Microsoft Windows*, *Linux* and *Mac OS X*. This work has been performed in the *Linux* distribution. [22]
3 Data processing

3.1 Terminology
Throughout this document there are references to three groups of people associated to a tweet. These are:

- **Author**: The user who wrote the tweet in the beginning
- **Retweeter**: The user who retweeted a tweet
- **Mentioned**: The user who is mentioned in the tweet
- **Twitter user ID**: This is a unique identifier for Twitter accounts. A user can change its username (screen name by Twitter parlance) but the user ID never changes. This is useful when you are using tweets/user's information from a couple of months ago, as with this work, because some users may have changed their username.

For more information on what are tweets, retweets and mentions see [15].

3.2 Starting point
To start this work I had a database containing public tweets. This database, property of the DIT-UPM department, contains public tweets posted between 5th June 2012 and 15th January 2013. The GET statuses/sample Twitter API was used to collect these tweets. This API returns a random sample stream of about 1% of all the tweets published to Twitter at any given moment.

To create the database, one of these samples was stored every 18 seconds, thus creating a representative sample of the tweets posted between the dates mentioned. The language was detected using automatic means; tweets posted in a language other than English were discarded and replaced by the next in the stream. Only retweets were stored.

When the recollection finished, the dataset comprised 472,645 retweets. As the database was going to be used to research on relationships between users, a random sample of tweets containing mentions was taken. Then a filtering to omit tweets with no useful information was carried. Tweets with inaccessible contents (deleted tweets, broken links, suspended accounts), mentions to non-individual users, sensitive content (as defined by Twitter media policy [Twitter2013a]), non-English language, different kinds of spam (mention abuse, massive follow back requests), duplicates, and self-mentions were replaced with tweets from the buffer dataset. The final sample, which is the one used in this work, has 2,382 tweets. All of them meet the following conditions:

- **Retweet**: As cited, all tweets are retweets
• **Mentions:** All tweets contain at least a mention to another user
• **English language:** All tweets are written in English. Therefore, the conclusion of this work will only be generalizable for the English-speaking world

The useful information is divided into three tables:

• **tweet:** Contains tweets and associated information (date posted, author, etc.)
• **mention:** Contains users mentioned in the tweets contained in *tweet*
• **user:** Contains information of the retweeters and mentioned users. It does not contain information of the original author of the tweet. (The gathering of this information is part of this project)

As the data analysis was performed in R, these three tables were exported as `.csv` files and then imported into RStudio using the `read.csv` function.

### 3.3 Authors characterisation

As it was mentioned in the previous section, the original database did not explicitly contain information from the original author. Authors’ usernames were detected and extracted from the tweet’s raw text using the *sub* function. This is an example of how it was done:

```
RT @BarackObama: How are you and your family celebrating this Independence Day?
```

As we can see, the author’s username is always preceded by the “RT @” substring. This facilitated its identification by automatic means. Once all of the author’s usernames were extracted they were stored in a vector.

The next step was to fully characterise every author - i.e. find his or her user ID, full name, number of followers and number of status. To do so, a web API ([http://gettwitterid.com](http://gettwitterid.com)) instead of the REST Twitter APIs was used. The main reason to use this API instead of Twitter’s REST APIs was its simplicity. To obtain the information needed to characterise the authors a simple GET HTTP request method was needed. This is an example of a call to the API to characterise username *BarackObama*:

```
http://gettwitterid.com/?user_name=barackobama&submit=GET+USER+ID
```

The only part that changes from one user to another is the bold letters. To do this characterisation, a function that automatically makes the call accepting the username as an argument was written. This function, called
getUserInfo, is going to be described in detail; its code may be found at Appendix 1.

As stated, getUserInfo accepts one argument. This argument must be a string object containing a valid Twitter username. The function makes a call to the API using the htmlTreeParse function of the XML package. The answer from the API comes in html form. Paragraphs of the html object are extracted using xpathApply (also from the XML package) and then flattened into a character vector using unlist. This character vector now contains all of the needed information of a user to be fully characterised as described above.

The getUserInfo function was applied to every author in the database. 504 users were not characterised –i.e. the API returned an error message. To ensure these users were not characterised because their profiles no longer existed or had changed names (and not due to an error of the API) the getUserInfo was reapplied to those usernames. After the two rounds 352 out of the 2382 users were not characterised.

3.4 Gender determination

Once the three groups of users were fully characterised the next step was to determine its gender. This section is divided into two parts because two different gender APIs were used: one for authors and another for retweeters and mentioned users.

3.4.1 Authors’ gender determination

To determine the gender of this group the Demographics API of AI-Applied (http://ai-applied.nl/demographics-api) was used. This API determines the gender of a user by analysing text written by this user and optionally its name or username. The API is accessed with a JSON-based RPC formatted call via GET HTTP request method. A function (AIquery) to write the calls with the correct format was written (it may be found at Appendix 1). To make the calls, the text used was the author’s tweet stored in the database and its username. This is an example of the JSON-formatted call and answer from the API:
AI-Applied API call for username BarackObama

{"status": 1, "id": null, "response": {"data": [{"gender": "male", "age": "41-55", "confidence_age": 0.665433400632437, "id": 1, "confidence_gender": 0.6307980900337483}, "description": "OK: Call processed.", "success": true, "quota": {"reset_time": "Never (Pre-paid credits)", "remaining_credits": 1845}}}

AI-Applied API answer for username BarackObama

The JSON answers were stored in a vector and the gender was extracted using the sub function. The gender of 153 users was not successfully determined and was therefore done by hand by the author of this work.

3.4.2 Retweeters’ and mentioned users’ gender determination
Due to the nature of the data available for retweeters and mentioned users (no tweets or other texts written by them) the AI-Applied API was not suitable for the determination of its gender. NamSor Origin API (http://blog.namsor.com/api) was used. This API is called through a GET HTTP request method. To determine gender the API analyses the given name in the cultural context of the surname —i.e. Andrea Martorella is most likely an Italian name for a male whilst Andrea García will most likely be a Spanish name for a female. This is an example of the JSON formatted call and answer for the NamSor Origin API:

http://api.namsor.com/onomastics/api/json/gendre/Barack/Obama

NamSor Origin API call for username BarackObama

{"scale":-0.5,"gender":"male","firstName":"Barack","lastName":"Obama","id":"1434190672935"}

NamSor Origin API answer for username BarackObama
To make correctly formatted calls a function (namsorQuery) was written, it may be found at Appendix 1. The names used in this API were the rtee.fullName and mentionee.fullName fields of the database.

These fields contain the names that Twitter users chose as their full name, in contrast with their username or screen name. In practice, some users decide not to input their actual names in this field and instead they write something different. To detect this, a simple algorithm that separated full names with two words from full names with one or more than two words was written. Names with exactly two words were considered true full names whilst names with one or more than two words were processed manually. The manual process consisted in accessing the Twitter profile and deciding the gender based on photos and/or the user’s tweets.

3.5 Final processing

After characterising every user in the database and having determined his or her gender all the information had to be put together in one data frame to facilitate statistical data analysis. Each row in that data frame should therefore contain the following information:

- **Tweet**: raw text of the tweet and a unique numerical identifier (tweet.ID)
- **Author**: the author of the tweet and his data (gender, user ID, full name, username, number of followers and number of tweets)
- **Retweeter**: the retweeter and his data (gender, user ID, full name, username, number of followers and number of tweets)
- ** Mentioned**: the mentioned user and his data (gender, user ID, full name, username, number of followers and number of tweets)

All this information was put together using the merge function with the tweet’s ID as the common column. The final processed data frame was called general.df1 and these are its columns:

| "tweet.ID" | "tweet.text" | "author.ID" |
| "author.username" | "author.fullName" | "author.gender" |
| "author.followers" | "author.status" | "rtee.ID" |
| "rtee.username" | "rtee.fullName" | "rtee.gender" |
| "rtee.followers" | "rtee.status" | "mentionee.ID" |
| "mentionee.username" | "mentionee.fullName" | "mentionee.gender" |
| "mentionee.followers" | "mentionee.status" |

Note: rtee in the code stands for retweeter and mentionee stands for mentioned user. This final data frame contains 1802 entries.
4 Data Analysis
The aim of this section is to determine if there is a relation between the gender of the author of a tweet and its retweeter and/or mentioned user –i.e. if mentioning a user or retweeting a tweet is influenced by gender.

4.1 Exploratory data analysis
The first step that was taken was to analyse the data in the form of graphs to have a better idea of how the sample data look like.

4.1.1 Gender distribution
First of all, we look at how gender is distributed throughout the three groups:

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>926</td>
<td>876</td>
</tr>
<tr>
<td>Retweeter</td>
<td>901</td>
<td>901</td>
</tr>
<tr>
<td>Mentioned</td>
<td>646</td>
<td>1156</td>
</tr>
</tbody>
</table>

A significant difference can be seen on mentions. At first glance it appears that more females are mentioning males than vice versa, having into account that there are more female authors but more male mentioned users. On the contrary, this does not appear to happen with retweeters which seem much more equally distributed. It is more obvious if seen in a relative frequency bar plot:

Relative distribution by gender
Anyway, these are just estimates as we cannot ascertain anything without seeing the actual relation between gender of the author and that of the retweeter or mentioned user.

4.1.2 Retweeters and authors relation
The above bar plot only informs us on the number of males and females per group but it does not give any information on the relation between the gender of the author and that of the retweeter per each tweet. To find this data we need to represent the number of times we have a female author with female retweeter, female author with male retweeter, etc. The following plot shows that information:

Relation between the author’s gender of a tweet and its retweeter’s gender
In the above plot “f-m” stands for female author and male retweeter, “m-f” male author and female retweeter and so on. We see in the plot that people tend to retweet people from their same gender; there are more males retweeting males and females retweeting females than people retweeting authors from the opposite gender. In any case we cannot yet ascertain if this is due to chance because numbers are quite similar - i.e. there are no big differences between the height of the columns in the bar plot. This last statement is easily understood with the following plot:

As we expected, percentages are similar. We can see how the percentages of retweeting someone from your own gender and someone opposite do not vary between genders - i.e. males retweet males in the same approximate proportion than female retweet other females (there is a small 0.2% difference). This plot does suggest that gender of the author and gender of the retweeter may not be independent but we cannot be certain, it is just a hypothesis.
The next step will be to test this hypothesis using Pearson’s chi-squared test; this will be done in section 4.2.

4.1.3 Mentioned users and authors relation
If we go back to what we saw in section 4.1.1 we would expect that a lot of female authors are mentioning male users. To find if this is true and to be able to make further considerations we want a bar plot of the relation between the author’s gender of a tweet and the gender of the mentioned user:

In the above plot “f-m” stands for female author and male mentioned user, “m-f” male author and female mentioned user and so on.
We can observe how authors behave differently than we would expect if we thought that their behaviour was going to be similar to that of the retweeters. Here we see that, apparently, independent from the author’s gender people tend
to mention male users more than female ones. To confirm this we need a segmented frequency bar plot:

![Bar plot showing mentioned user gender by author's gender](image)

It starts to become clear that both female and male authors tend to mention male users more from the data in the sample.

4.2 Pearson’s Chi-squared test

After having explored the data in section 4.1 we have seen that gender seems to be an influencing factor when interacting with other users in Twitter. In statistical terms, we are hypothesising that the gender of the author of a tweet and that of the mentioned user and/or retweeter are not independent. To assess this, we are going to perform a hypothesis test using the Chi-squared distribution.
4.2.1 Retweeters and authors relation
This is the contingency table for this relation:

<table>
<thead>
<tr>
<th>Author</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweeter</td>
<td>Male</td>
<td>467</td>
<td>434</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>409</td>
<td>492</td>
</tr>
<tr>
<td>Total</td>
<td>876</td>
<td>926</td>
<td>1802</td>
</tr>
</tbody>
</table>

Thus, the null hypothesis ($H_0$) will state that the gender of the author of a tweet and the gender of its retweeter are independent whilst the alternative hypothesis ($H_A$) states that gender of the author and retweeter are not independent:

$$H_0 = \text{Gender of author and retweeter are independent}$$
$$H_A = \text{Gender of author and retweeter are not independent}$$

The `chisq.test` function is used to perform Pearson’s Chi-squared test with Yates’ correction for continuity yielding the following results:

$$\chi^2 = 7.2175; \; df = 1; \; p - value = 0.007219$$

These results mean that if $H_0$ were true, then there would be a 0.007219 probability of having results at least as extreme as those in our sample. With such a small p-value we can reject the null hypothesis (even at the 99% significance level). We can therefore be reasonably sure that gender influences users when retweeting other users.

4.2.2 Mentioned users and authors relation
This is the contingency table for this relation:

<table>
<thead>
<tr>
<th>Author</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentioned</td>
<td>Male</td>
<td>590</td>
<td>566</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>286</td>
<td>360</td>
</tr>
<tr>
<td>Total</td>
<td>876</td>
<td>926</td>
<td>1802</td>
</tr>
</tbody>
</table>

Similarly to section 4.1.1, the null hypothesis ($H_0$) states that the gender of the author of a tweet and the gender of the mentioned user in the tweet are independent. The alternative hypothesis states ($H_A$) that the gender of the author of a tweet and the gender of the mentioned user in the tweet are not independent:
$H_0 = \text{Gender of author and mentioned user are independent}$

$H_A = \text{Gender of author and mentioned user are not independent}$

The `chisq.test` function yields the following results:

$$\chi^2 = 7.3251; \; df = 1; \; p - value = 0.0068$$

Again, the same interpretation prevails: if $H_0$ were true, then there would be a 0.0068 probability of having results at least as extreme as the ones in our sample. We can therefore be quite confident to reject the null hypothesis in favour of the alternative.

### 4.3 Followers and influence

We could argue that the number of followers that a user has could influence mentions or retweets - i.e. someone with a large quantity of followers is more likely to be mentioned or retweeted because he is famous than for his gender alone.

#### 4.3.1 Retweeters and authors relation

In this case, the number of followers that an author has could influence potential retweeters. We look at the authors’ followers distribution:

![Author's followers density plot (Logarithmic transform)](image)
In the above plot “f-m” stands for female author and male retweeter, “m-f” male author and female retweeter and so on. As the actual distribution of followers is extremely right-skewed the logarithm of followers has been plotted instead. We can see that this variable may be influencing retweeters. Followers are not equally distributed amongst the four groups represented in the above plot. In the case of female retweeting males (m-f graph) we see a clear peak at the right of the main one. This can mean that female users may have retweeted some male authors because of their large quantity of followers. This peak is not as notable in the case of female retweeters and female authors (f-f graph). Another noteworthy feature is that male retweeters also retweeted other males (m-m graph) with large quantity of followers though this is more equally distributed than with female retweeters. This did not happen when retweeting female users. We could hypothesise that males are more likely to be retweeted based on large quantities of followers than females but that study is out of the scope of this work.

4.3.2 Mentioned users and authors relation
In this relation the importance of followers may influence mentions. If this was to be true, the relevant number here is the number of mentioned users followers and not authors’ followers. This is the mentioned users’ followers distribution, divided by gender:
We can see how the behaviour of male and female authors is very similar: density plots m-m and f-m have analogous shapes and the same happens with m-f and f-f. Thus, we could hypothesise mean that mentioning a user in tweeter does not depend on their number of followers; but again, this is out of the scope of the present project.
5  Data Analysis – Revisited
We have seen that when users have a large number of followers (and ultimately fame) it may influence other users to interact with them – e.g. a user will most likely mention user Barack Obama because he is famous and he is the president of the US than solely based on his gender. Therefore, it may be interesting to subset among our sample to extract only those users that are not famous. We will define Twitter fame as having more than 1,000 users. This is an informal threshold used nowadays in Internet sites and blogs to define users as something more than the average user.

5.1  Exploratory data analysis
As with section 4 we will first perform some kind of visual analysis to see if we can spot interesting trends amongst our data.

5.1.1  Gender distribution
This is how gender is distributed in our subset:

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>182</td>
<td>155</td>
</tr>
<tr>
<td>Retweeter</td>
<td>179</td>
<td>158</td>
</tr>
<tr>
<td>Mentioned</td>
<td>134</td>
<td>203</td>
</tr>
</tbody>
</table>

We continue to see a significant difference in mentioned users, something that already happened with the larger sample. On the contrary, there seems to be some more difference between male and female users in the other two groups. We can see this more clearly with a relative frequency bar plot:
As we expected, the difference between male and female mentioned users has reduced (by 4%). On the contrary, for retweeters and authors the difference has changed by 3% approximately. We have to go further and look at the relations if we want to be sure that there are significant changes.

5.1.2 Retweeters and authors relation

With the data we have until now, we could think that retweets are very equally distributed amongst author’s gender (after all, the number of authors and retweeters is very similarly distributed). The following plot shows the importance of looking at the actual relations and not only the distribution by group:

![Retweeter gender by author’s gender](image)

We can see that contrary to the hypothesis that gender would be equally distributed, females were retweeted in a higher proportion by female users. This effect is much less extreme in the case of male authors, whose tweets were more equally distributed amongst the gender of retweeters.

If we compare this plot to the one on section 4.1.2 we see that again, there is only a significant change in the distribution of female authors where a greater
percentage of their tweets were retweeted by female users, male author gender distribution remaining quite similar.

5.1.3 Mentioned users and authors relation

We would expect to see a lot of female users mentioning male users like in section 4 here. We could even hypothesise that the distribution will have even more female authors mentioning male users as the actual authors distribution has a higher percentage of female authors than the one in section 4. This is how a segmented frequency bar plot for the relation between mentioned users and authors looks like:

![Segmented frequency bar plot](image)

Contrary to our initial hypothesis, there are much less female authors mentioning male users than vice versa. If we compare this plot with the one in section 4.1.3 we see that proportion of gender in mentioned users by male author has remained approximately the same whilst the proportion in female authors has changed. Now this proportion is much closer to the expected one if we supposed that mentioning a user is independent of gender (theoretical expected distribution).
5.2 Pearson’s Chi-squared test
Similarly to section 4.2, gender seems to be influencing users when the time to interact with other users comes. Also, we can be relatively sure that the possible followers factor has been erased in this subset. We are going to perform another Chi-squared test to check if our hypothesis is right or wrong.

5.2.1 Retweeters and authors relation
This is the contingency table for this relation:

<table>
<thead>
<tr>
<th>Retweeter</th>
<th>Author</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Male</td>
<td>84</td>
<td>74</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>71</td>
<td>108</td>
<td>179</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>155</td>
<td>182</td>
<td>337</td>
</tr>
</tbody>
</table>

Hypotheses are identical to those in 4.2.1:

\[ H_0 = \text{Gender of author and retweeter are independent} \]
\[ H_A = \text{Gender of author and retweeter are not independent} \]

Pearson’s Chi-squared test with Yates’ correction yields the following result:

\[ \chi^2 = 5.6258; \ df = 1; \ p-value = 0.0177 \]

This p-value, although small, is approximately 2.5 times bigger than the p-value in section 4.2.1 (0.007219). As we saw in the exploratory data analysis, the distribution of users is much closer to the theoretical expected distribution. This bigger p-value confirms this idea. This p-value would mean that we would fail to reject the null hypothesis at the 99% significance level but we would reject it at the 95%. As this is a very subjective issue, further analysis will have be done in the next sections to see if we can determine with more certainty if gender is an influential factor when retweeting another user.

5.2.2 Mentioned users and authors relation
This is the contingency table for this relation:

<table>
<thead>
<tr>
<th>Mentioned</th>
<th>Author</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Male</td>
<td>106</td>
<td>97</td>
<td>203</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>49</td>
<td>85</td>
<td>134</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>155</td>
<td>182</td>
<td>337</td>
</tr>
</tbody>
</table>
Again, hypotheses are the same as in section 4.2.2:

\[ H_0 = \text{Gender of author and mentioned user are independent} \]
\[ H_A = \text{Gender of author and mentioned user are not independent} \]

The `chisq.test` function yields the following results:

\[ \chi^2 = 7.341; \; df = 1; \; p-value = 0.00674 \]

If we compare this value to the one in section 4.2.2 (0.0068) we can see they are very similar. Thus, the interpretation remains the same, either at the 99% or 95% significance level the null hypothesis would be rejected and we would conclude that gender, indeed, influences users when the mention other Twitter users. Nevertheless, we cannot be certain which gender (if not both) is causing the dependency. Thus, further analysis will be carried out in the next sections to try to find this.

### 5.3 Binomial test

The advantage of the Binomial test over Pearson’s Chi-squared test is that the former can give us information on which gender is causing our variables to be independent. Thus, retweeters’ gender may be equally distributed for female authors but not for male authors or vice versa. In the following sections we will perform binomial tests to try to find more information about our data.

#### 5.3.1 Retweeters and authors relation

As stated in section 5.2.1, the p-value of the Chi-squared test could be considered ambiguous. Furthermore, looking at the graph of section 5.1.2 we could see that the gender of retweeters for female authors was very unequally distributed but this did not happen for male authors. This fact may lead us to think that maybe only the gender distribution of retweeters’ of female authors is causing the p-value in the Chi-square test to be small. To see if this is true we can do two binomial tests, one for male authors and another one for female authors. In this test we will suppose that for each author gender the true probability (p) of having a male (or female) retweeter is 0.5 – i.e. that gender of author and gender of retweeter are in fact independent. Then we will see if the actual gender distribution for retweeters could have occurred just by chance. We will suppose that each retweet is independent from the rest of retweets.

##### 5.3.1.1 Male authors

In Binomial tests we need to define number of successes, number of trials and true probability or probability of success. We will consider male retweets successes and female retweets failures (note that the words failure and success do not have any meaning other than the mathematical). Therefore, the number of successes for this test is equal to the number of male retweeters given that the author of the tweet was a male user (84). The number of trials is the size of the
sample, which is the total number of retweets given its author, was male (155 retweets). Finally, we will consider 0.5 to be the probability of success as this would be the probability of finding a male or female retweeter if gender of authors and retweeters were in fact independent. It is important to note that we will perform a two sided test; the reason being that we are not interested solely on the extremeness of male retweeters but also in the potential extremeness of female users. This is represented by a two-sided test instead of a one-sided test. Thus, we will be testing what is the probability of having 84 male retweeters out of a total of 155 retweeters if gender of author and retweeter were independent. We will use the binom.test function, which yields the following result:

\[ p - value = 0.3351 \]

This is a high p-value, which means that we cannot be certain that the difference in the actual gender distribution of retweeters for male authors with the theoretical expected distribution could be due to chance. Therefore, we would fail to reject the null hypothesis, meaning that male authors are not retweeted based on their gender.

5.3.1.2 Female authors
We are now going to perform the same Binomial test but for female authors. We will continue to define success as male retweets. Therefore, the number of successes in this test is 74, the number of trials is 182 and the true probability is 0.5. The binom.test function yields the following result:

\[ p - value = 0.01422 \]

We can now see that it was the distribution of gender of retweeters in female authors the one that caused the p-value for the Chi-squared test to be small. Again, this value would cause the null hypothesis to be rejected at the 95% significance level but not at the 99%. We can therefore be quite certain, but not completely certain, that retweeting a female author is in fact influenced by gender.

5.3.2 Mentioned users and authors relation
Similarly to section 5.3.1, we are going to use a Binomial test to see if we can determine which of the two genders made the p-value in the Chi-squared test to be so small – i.e. determine which gender, is influenced by the potential mentioned user’s gender at the time to do that mention.

5.3.2.1 Male authors
As in former Binomial tests in this work, we will define success as the mentioned user being male. Therefore, our number of successes is 106, out total number of trials 155 and the probability of success is 0.5. This is the result of the test:
\[ p - value = 5.444 \times 10^{-6} \sim 0 \]

The p-value is so small that we can consider it to be zero. This would mean that in the case of a male author, mentioning another user is always influenced by the potential mentioned user’s gender. If we go back and look at the graph in section 5.1.3, this seems obvious, as the distribution is very different to the theoretical expected distribution.

### 5.3.2.2 Female authors

Now we will test if female authors also are influenced by the potential mentioned user’s gender. As with all other Binomial tests in this work, success is defined as the number of male users. Therefore, the number of successes is 97, the number of trials 182 and the true probability 0.5. These are the results yielded by the `binom.test` function:

\[ p - value = 0.4149 \]

This is the highest of the p-values in this project. We can be quite certain that the difference between the actual gender distribution of mentioned users and the theoretical expected one is due to chance.

We can also conclude that it is male users who make gender of the author of a tweet and gender of the mentioned user in the tweet to be dependent.

### 5.4 Logistic regression (Logit model)

One of the aims of data science and particularly data analysis is to create models that enable us to predict future outcomes of a certain event. In this work, we have been analysing the relation between the genders of a tweet’s author, its retweeter and the mentioned user. All of the previous work has been oriented to determine if there is a relation between those genders or if those gender distributions could be attributed solely to chance.

This section aims a bit further. We will now create a logistic regression model to try to estimate the gender of the author of a tweet given that we know certain other variables. These variables, called predictors, will be those that we have in our database –i.e. the basic public variables that can be extracted and/or inferred using an Internet API. Namely: the author’s, retweeter’s and mentioned user’s number of followers plus the retweeter’s and mentioned user’s gender.

To build the model we will use the `glm` function. The data set used for this model will be the original one containing 1802 tweets.

We will create two subsets from the big data set, a training one and a test one. They will be random samples taken from the data. The training subset will have a size equal to 70% of the original data frame and therefore the test subset will be 30% of the original one. The training subset will be used to build the model and the test one to evaluate how good the model is.
5.4.1 Training subset

The *glm* function yields the following results:

\[ p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5)} \]

Where:

- \( p \) is the probability of being male (0 is female and 1 is male)
- \( \beta_0 = -4.426 \times 10^{-1} \)
- \( \beta_1 = 3.989 \times 10^{-8} \)
- \( \beta_2 = 3.254 \times 10^{-1} \)
- \( \beta_3 = -7.935 \times 10^{-7} \)
- \( \beta_4 = 2.345 \times 10^{-1} \)
- \( \beta_5 = -7.043 \times 10^{-9} \)
- \( x_1 \) is the author’s number of followers
- \( x_2 \) is the retweeter’s gender (0 is female and 1 is male)
- \( x_3 \) is the retweeter’s number of followers
- \( x_4 \) is the mentione user’s gender (0 is female and 1 is male)
- \( x_5 \) is the mentioned user’s number of followers

To help with the interpretation of this model, we plot \( p = \frac{\exp(x)}{1+\exp(x)} \) where 

\[ x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5 \]
As we can see, probability is a monotonically increasing function varying between 0 and 1. When \( x = 0 \) then \( p = 0.5 \), which would mean the gender is unknown. Therefore, any positive value of \( x \) will make the model predict a male and any negative value will be a female prediction. If we look at the \( \beta_i \) coefficients, we can see that those related to gender (\( \beta_2 \) and \( \beta_4 \)) are positive. This means that when either \( x_2 \) or \( x_4 \) are 1, which corresponds to a male retweeter/mentioned user (remember they can only be 0 or 1) the odds of the author’s gender being male increase. Likewise, when either \( x_2 \) or \( x_4 \) are 0 the odds of being female increase. This means that the model is predicting that interacting with females is most likely done by a female and interacting with males is most likely done by a male.

Another noticeable feature of this model is the difference in orders of magnitude between the \( \beta_i \) coefficients associated to a gender variable and those associated to a number of followers variable. The former are at least 6 orders of magnitude greater than the latter.

The \( \beta_i \) coefficients associated to number of followers variables are all negative, this means that when a mentioned user or retweeter has a big number of followers, the model is predicting that the author will most likely be a female.

### 5.4.2 Test subset

Now we have built our model, we have to test it with data that it has never seen. This means we are going to use the model to predict the author’s gender of a subset of tweets with the test subset. This is the contingency table for the model’s predictions:

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Actual</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>Total</td>
</tr>
<tr>
<td>Female</td>
<td>153</td>
<td>121</td>
<td>274</td>
</tr>
<tr>
<td>Male</td>
<td>82</td>
<td>95</td>
<td>177</td>
</tr>
<tr>
<td>Total</td>
<td>235</td>
<td>216</td>
<td>451</td>
</tr>
</tbody>
</table>

The accuracy of the prediction, defined as correct predictions over total predictions is:

\[
\text{accuracy} = \frac{153 + 95}{451} = 0.5498891 \sim 55\% 
\]

Another measure we can use to evaluate the model is Cohen’s kappa coefficient [16]. This statistic measures inter-rater agreement; it varies from 0 to 1, 0 being no agreement and 1 perfect agreement. If we suppose the model to be one rater and the actual gender the other rater, then we can calculate the coefficient as:
\[ \kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \]

Where:
\( Pr(a) \) is the relative observed agreement among raters
\( Pr(e) \) is the hypothetical probability of chance agreement

Therefore, Cohen’s kappa has the following value in our model:
\[ \kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} = 0.2591241 \]

We can see, both by the Cohen’s kappa’s and the accuracy value that the model is not very good. We can see in the contingency table that the model predicted much more females than males. This is could be attributed to the fact that all the \( \beta_i \) coefficients associated to number of followers variables are negative and thus all the tweets containing a user with a large number of followers have been likely classified as having a female author.

Another factor influencing the low accuracy is how the predictions are distributed, this is a box and whisker plot for the gender probabilities for the author’s gender (0 is 100% probability a female and 1 is 100% probability a male):

<table>
<thead>
<tr>
<th>Minimum</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Maximum</th>
<th>Interquartile range</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3363</td>
<td>0.4477</td>
<td>0.4702</td>
<td>0.4855</td>
<td>0.5292</td>
<td>0.9334</td>
<td>0.0815</td>
</tr>
</tbody>
</table>
We can see that the value of the IQR (the box in the plot) is very small. It is centred on $0.47$ which makes sense with the fact that more predictions were females. The majority of values for the prediction are near $0.5$. Even though we have considered that any value above $0.5$ is a male prediction and any value below $0.5$ is a female prediction, when numbers are so close to $0.5$ it means that the prediction is not certain.

It makes sense that the model does not work very well, after all predicting the gender of the author of a tweet simply by knowing the retweeter’s and mentioned user’s gender and their number of followers seems complicated.
6 Conclusions
All of this project’s goals have been amply fulfilled. Firstly, a variety of techniques and methodologies for data and content analysis have been studied and characterised. Secondly, these techniques have been applied to solve a practical problem: the characterisation and analysis of gender in inter-user relationships in a particular social network, Twitter. Thirdly, R in the RStudio environment has been learnt and the aforementioned methodologies and techniques have been implemented and applied to the practical problem using the language. Finally, a series of web APIs have been utilised to characterise the social network users and infer their personal data. We can therefore consider this project a success.

Strictly speaking about the content of this work, we have seen how gender does in fact influence users at the time to interact with another user. We have also seen the number of followers of a user tends to influence the fact that he or she is mentioned. This fact was specially pronounced in the case of female mentioned users. It is important to remember at this point how the tweets were extracted. We only considered English-language tweets and so these conclusions are only valid for the English-speaking world.

Finally, we have seen that it is complicated to predict the gender of a user based solely on the gender of the retweeter and mentioned user in the tweet plus the number of followers of each of the three users involved. Although gender of the three groups involved in a tweet is not independent, as we have seen, going a step further –i.e. predicting the gender of the author of a tweet requires more information than that that was available for this project.

To continue this work, more information should be considered to try to predict more accurately the gender of users. Firstly, timestamps could give information. Secondly, using geo-location information from the tweets to restrict the study to one culture instead of all the English speakers could give more relevant results. Finally, counting with a team (this work has not been done by a team, obviously) to do content analysis to infer the gender of Twitter users would certainly make the conclusions of the data analysis much more relevant as content analysis should, by definition, be done by more than one person.

On the whole, it has been a satisfactory and interesting project that opens the door for future research in the field of personal data, inter-user relations and gender dependency.
7 References


http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm

http://www.ats.ucla.edu/stat/r/dae/logit.htm

http://en.wikipedia.org/wiki/Twitter


https://en.wikipedia.org/wiki/Krippendorff's_alpha

http://onlineqda.hud.ac.uk/Intro_QDA/what_is_qda.php


https://en.wikipedia.org/wiki/Application_programming_interface

8 Appendix 1

This appendix contains relevant code mentioned in other parts of this document

8.1 getUserInfo

```javascript
getUserInfo = function(username) {
  s1 = "http://gettwitterid.com/?user_name=
  s2 = username
  s3 = "&submit=GET+USER+ID"
  query = paste0(s1,s2,s3)
  htmlAnswer = htmlTreeParse(query,encoding = "UTF-8",
                            useInternal = TRUE)
  answer = unlist(xpathApply(htmlAnswer, '//p', xmlValue))
  return(answer)
}
```

8.2 AIquery

```javascript
AIquery = function(tweet,user,id) {
  # Deletion of characters that cause errors in the API
  correctedTweet = str_replace_all(tweet,""," ") # "
  correctedTweet = str_replace_all(correctedTweet,"#"," ") # #
  correctedTweet = str_replace_all(correctedTweet,";"," ") # ;
  correctedTweet = str_replace_all(correctedTweet,"&"," ") # &
  correctedUser = str_replace_all(user,""," ") # \
  correctedUser = str_replace_all(correctedUser,"#"," ") # #
  correctedUser = str_replace_all(correctedUser,";"," ") # ;
  correctedUser = str_replace_all(correctedUser,"&"," ") # &
  correctedUser = str_replace_all(correctedUser,"\\\\"," ") # \
  # Declarations
  quotMarks = ""
  comma = ","
  s1 = "http://api.ai-applied.nl/api/demographics_api/?request={"
  "data":{
  "api_key":"506d21a5ca7b517368333bb6363f76ab30f79894",
  "call":{
  "return_original":false,
  "data":[
  {
  "text":"

  s2 = paste0(quotMarks,correctedTweet,quotMarks,comma)
```
```r
s3 = "\"language_iso\":\"eng\","

s4 = "\"user\":"

s5 = paste0(quotMarks,correctedUser,quotMarks,comma)

s6 = "\"id\":"

s7 = id

s8 = "}

"

query = paste0(s1,s2,s3,s4,s5,s6,s7,s8)

return(query)
```

### 8.3 namsorQuery

```r
namsorQuery = function(name,surname) {
  # Deletion of characters that cause errors in the API
  correctedName = str_replace_all(name,"\""," ") # "
  correctedName = str_replace_all(correctedName,"#","") # 
  correctedName = str_replace_all(correctedName,"","") # 
  correctedName = str_replace_all(correctedName,"","") # 
  correctedSurname = str_replace_all(surname,"\""," ") # "
  correctedSurname = str_replace_all(correctedSurname,"#","") # "
  correctedSurname = str_replace_all(correctedSurname,"","") # "
  correctedSurname = str_replace_all(correctedSurname,"","") # "
  correctedSurname = str_replace_all(correctedSurname,"","") # "
  correctedSurname = str_replace_all(correctedSurname,"","") # "

  # Declarations
  s1 = "http://api.namsor.com/onomastics/api/json/gendre/
  s2 = correctedName
  s3 =="/"
  s4 = correctedSurname
  query = paste0(s1,s2,s3,s4)
  return(query)

```