Summer 5-2019

Supervised Machine Learning Models for Fake News Detection

Gofaas Group

Andrea Lopez

Adelo Vieira

Zafar Ahsan

Farooq Saqib

See next page for additional authors

Author

Gofaas Group, Andrea Lopez, Adelo Vieira, Zafar Ahsan, Farooq Saqib, and Shirley Marinho

CCT COLLEGE

___

Supervised Machine Learning Models for Fake News Detection

by Gofaas group

INDEX:

Executive Summary

Fake News noun [U]

UK /ˌfeɪk ˈnjuːz/ US /ˌfeɪk ˈnuːz/

Definition: <u>false</u> <u>stories</u> that <u>appear</u> to be <u>news</u>, <u>spread</u> on the <u>internet</u> or using other <u>media</u>, usually <u>created</u> to <u>influence</u> <u>political</u> <u>views</u> or as a <u>joke</u>.

Synonyms and related words: True/ Real/ False/ and Unreal

We live in an age where fake news is rife across social media platforms, certain news outlets and streaming sites. We are bombarded with news every second of every day. But how do we know if the news feed we are reading is from a reliable source? This is in effect our objective with this project – to acknowledge that news – fake or otherwise – is not transparent and to explore ways in which online content can be filtered more thoroughly.
In short, we are filtering out the fake.

Before we begin...a word of thanks

We would like to sincerely thank Muhammad Iqbal for his patience encouragement and guidance. We would not be able to accomplish this project without your support. Special thanks to Graham Glanville and Mark Morrissey for their encouraging words and shrewd observations.

Abstract

Fake news or the distribution of disinformation has become one of the most challenging issues in society. News and information are churned out across online websites and platforms in real-time, with little or no way for the viewing public to determine what is real or manufactured. But an awareness of what we are consuming online is becoming apparent and efforts are underway to explore how we separate fake content from genuine and truthful information.

The most challenging part of fake news is determining how to spot it. In technology, there are ways to help us do this. Supervised machine learning helps us to identify in a labelled dataset if a piece of information is fake or not. However, machine learning can be a black-box tool - a device, system or object which can be viewed in terms of its inputs and outputs – that focuses on one aspect of the problem and in doing so, isn't addressing the bigger picture. To solve this issue, it is very important to understand how it works. The process of data pre-processing and the dataset labelling is part of this understanding. It is also worth knowing the algorithms mechanisms in order to choose the best one for the proposed project.

Evaluating machine learning algorithms model is one way to get better results. Changing paths within algorithms is not a bad thing if it is addressing the limitations within. With this project, we have done just this, changing from Sports news detection using Twitter API to labelled datasets and as a result we have an original Gofaas dataset, Gofaas library R package and Gofaas WebApp. Machine Learning is a demanding subject but fascinating at the same time. We hope this modest project helps people to face these challenges and learn from our findings accordingly.

Key words: Fake news, Machine learning, Data Mining, Supervised learning

Chapter 1. Introduction

Initial Proposal

"The term fake news is simultaneously too broad and too narrow"
(Alice E. Marwick in an academic paper titled Why Do People Share Fake News?
A Sociotechnical Model of Media Effects.)

At the beginning of this project two big problems were addressed. One was the *fake news definition* and the second *how to detect it using machine learning*. Fake news subject itself is far more complex than was thought and brings about many challenges. The root of fake news is *disinformation* and as such it is a term that can be stretched in several different directions, because of its proposed content and subsequent connotations. [Chapter 2. Literature Review Fake News] is a good explanation of Fake News and is an indicator of how the term is perceived in the writing of this project.

According to Shu et al (2016) fake news can break the authenticity balance of the news ecosystem because it intentionally persuades consumers to accept biased or false beliefs causing confusion and distrust among people. Fake news is a piece of information that secretly leaks into the communication process in order to deceive and manipulate. The most popular way of spreading it is using social media platforms.

Social media is an immediate and less expensive way to publish, produce, share and consume information when compared with traditional news media (Shu et all, 2016) and it is not by chance that Facebook, YouTube, WhatsApp and Twitter [Figure.1] are the top social network for news.

Across 26 countries
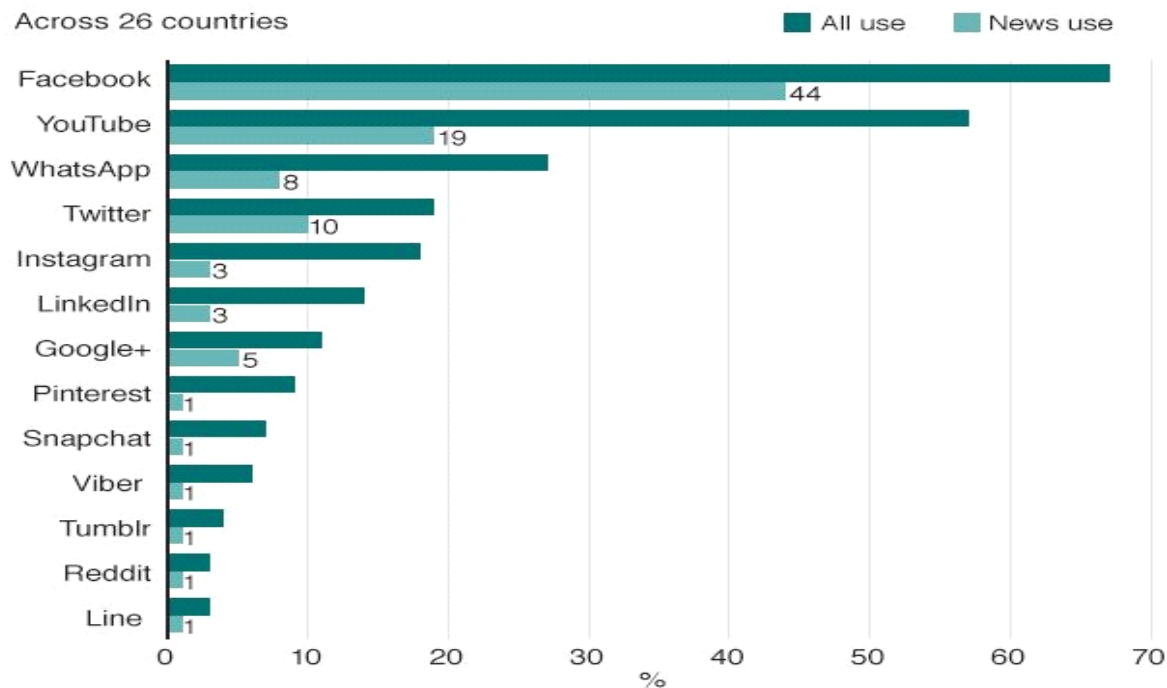
[Figure.1] Top Social Network for news

"Authentic and Fake News are very challenging to separate unless some steps are carefully followed such as tidy data, checking source, checking the date, examining evidence."(Mindtools, 2018). In order to detect fake news using machine learning, the first attempt on our part was data mining extraction from Sport News by using *tweepy* in R language. The reason why Twitter platform was chosen is because the piece of information published is public and accessible (Sistilli, 2015), whereas Facebook and WhatsApp are private.

From February to April 2019, data was collected from Twitter [Figure.27] as carefully described in [Chapter 3. Twitter – Sentimental Analysis]. The data was cleaned, processed and analysed by using Text Mining Techniques. The very first idea was to use Python language to interact with Twitter's API and store the data in an MySQL database following the format at [Figure.2]. However it was replaced by R language.

| News article	Data about users	Comments	Likes	Shares	T/F
text of the news 1	data	data	data	data	T
text of the news 2	data	data	data	data	T
text of the news 3	data	data	data	data	F
text of the news 4	data	data	data	data	T
text of the news 5	data	data	data	data	F
...	...	...	...	...	...

[Figure.2] Structure of dataset.

Sentimental Analysis was the first approach to identify and categorise opinions expressed in the collected data. The results were calculated to show people's feelings about Sports news. The key words: #halamadrid, #fcbacelona, #messi was used to collect tweets from Twitter's API.

The [Figure.3] shows a simple model to detect fake news by word using the same approach of collecting data, and [Figure.4] supposed to be the final prototype of the fake news detector. The idea was to develop an application that works as an extension of the internet browser. In this sense, when the reader accesses the news, the application would display a Pop Up Notification to inform the read if that information was fake or authentic.

[Figure.3] A simple model to detect fake news.
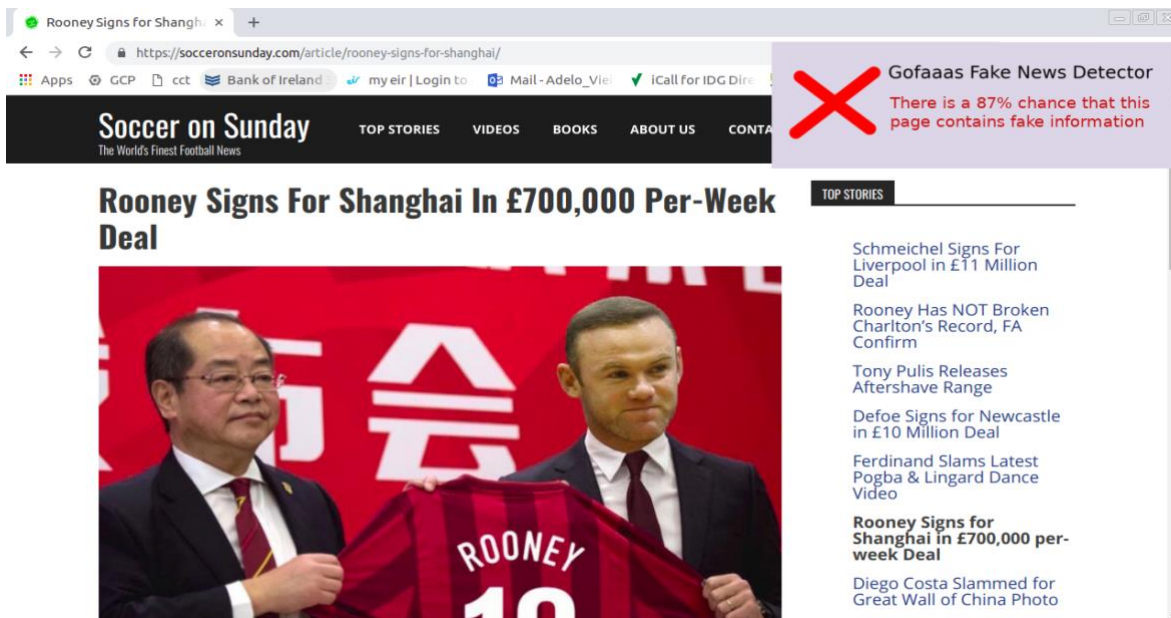
[Figure.4] A simple model to detect fake news.

It turned out that we had to change paths when we faced the second challenge of not being able to use tweets. The pre-processing data would take longer than was thought and had the possibility of resulting different outputs. Other problems included the labelling process, because when the data is unlabelled, supervised learning is not possible (Wikipedia, 2019). Helmstetter and Paulheim (2018) spotted out that manual annotation of tweets as fake or non-fake news is an expensive and tedious endeavour.

The solution was to find binary labelled datasets sourced from Kaggle* and VictoryUniversity. It has a vast repository of datasets and two of them were used to train and test the model. Naïve Bayes, Random Forest and XG-Boosting were the algorithms chosen to be worked on. [Chapter 4. Training Model] describes the datasets, each algorithm their machine learning model implementation and the reason they were chosen.

The results were very likely as the state of dominance of the algorithms. XG-Boosting, Random Forest, Naïve Bayes. However the accuracy rate could not be guaranteed by trusting the datasets labels, they were done by third-parts which does not ensure reliability. As a result, 511 entries dataset were manually compiled, the Gofaas dataset to ensure reliability of the results.

The Chapter 6. Gofaas Web App describes the final result of the project. The application takes a piece of text or a dataset and process the model to display to the user a result of fake or non-fake content. The application uses in the backend of Chapter 5. Gofaas R package which runs the model created.

*Kaggle is an online community of data scientists and machine learners, owned by Google LLC.

Chapter 2. Literature Review Fake News

Historical Foundations

"A lie gets halfway around the world before the truth has a chance to get its pants on"
Winston Churchill

Fake news was named 2017's word of the year, raising tensions between nations, and may lead to regulation of social media. Fake news is not only the most debated socio-political topics of the last three years but also it is seen as one of the greatest threats to democracy, free debate and the Western order. (Telegraph, 2018).

Despite of the fact that Fake News has been widely used in recent years may ensure to be new terminology. Therefore, it is actually quite an old matter, known as "Disinformation". The history of disinformation has been repeating throughout history with the same main goal - to influence and persuade. It was a concept used by governments and powerful individuals as a weapon for millennia. Octavian famously used a campaign of disinformation to aid his victory over Marc Anthony... (Carson, 2018).

There are plenty of examples of false news throughout history, in the 15th Century, more specifically in 1439, when the printing press was invented by Johannes Gutenberg the diffusion of disinformation and misinformation was facilitated through sensationalism accounts of the everyday events, there list of scandals, lies, hoaxes thought history is long. (Darnton, 2017).

In 1522, the election of a Pop was manipulated by writing "*wicked sonnets*" of all the candidates but of Medici, Pietro Aretino was one of the favourites authors of these sonnets. The sonnets were placed at the bust of a figure known as Pasquino. It was located near to the Piazza Navona in Rome where most of fake news about public figures were diffused. *Pasquinade* was the

giving name which developed into a genre of news. After 1755 when the Lisbon Earthquake took place, the church and many European authorities blamed the natural disaster on divine retribution against sinners. Fake news pamphlets alleged that some survivors owed their lives to an apparition of the Virgin Mary. These comments were one of the more complex news stories of the time (Soll, 2016).

It was announced in the 1780s, the capture of a monster in Chile that was allegedly being shipped to Spain. It was called "canard", which could be considered as the successor of the *pasquinade.* Canards were the fake news version in Paris for the next two hundred years.

Going to India New Delhi, in July 1983 a remarkable history appears in the newspaper: *"Aids may invade India. Mystery disease caused by US experiments. It was created as a* biological weapon".[Figure.5]

[Figure.5] New Delhi, July 1983 remarkable history in the newspaper.

The fake history which is still believed to be truth nowadays was spread around countries like Africa, Kenya, Bangladesh, Bulgaria, Cameron, Finland, Pakistan, London and Russia. In March 30th, 1987, the information reached America where it was broadcast over national television. The image of the United States was damaged with a toxic impact on their culture and policies. It was in every day on the back of their minds, fear and anger emerging from the Americans.

According to *The New York Times* (2018), Disinformation Campaign is weapon created by the KGB (Committee for State Security) in the 50's. The main goal of it is to change the perception of reality of every American until the point where no one is able to get sensible conclusions in the interest of defending themselves or their community. The KGB had a department named as Ideological Subversion also known as Active Measure (Russian: активные мероприятия) which was active since the 50s. Their target was to subvert anything with value in the United States. They wanted demoralized the fiber of the nation, destabilize them from inside like a virus.

The KGB spent 85% of its time creating false histories to influence people. They worked according to seven rules. The first one looked for cracks or social division. The second created a big lie and wrapped it with the third rule, a kernel of truth. Fourth, concealing your hand - in other words, ignore the source. The fifth use was to find a useful person to push the message to the population. The sixth rule was to deny everything because people's attention is short. The seventh and final rule was to play the long game – the news could accumulate, and by repetition and longevity as a proposed fact it eventually became the truth.

It was in 1986-87 when the '*square of truth'* turned out to be the most powerful and effective disinformation campaign to be a Fake News Propaganda. (United State Department of State, 1987). The official document

revealed the source and how Fake News was detected. In 1989, the Soviet Union was dissolved and so too was the Active Measures system. Washington, March 27, 2016, the Clinton's email scandal took on a life of its own after a hacking attack. The source was found to be Guccifer 2.0 GRU officer Grizodubovoy Saint Petersburg, Russia, KGB has returned in another form and uses technology as a venue for their most effective weapon.

At the time of 1975 to 1991, Vladimir Putin joined the Community Part of Soviet Union and KGB. To be promoted in the KGB, agents had to spend 25% of their time generating ideas to create false histories. It is thought that during this period disinformation was tested among the Russian population.

In 1998-99 Vladimir Putin became President of Russian.

The creation of the Global English language News channel was in order to promote Trump sympathy in Americans. In 2008, there was the launch the Internet Research Agency. All the pieces came together, the generation of pro-Trump feed and an agency through which to share the information - and Donald Trump successfully win the election of 2016.

Technology as a way of spreading fake news is now much quicker than was thought possible. A process that used to take six years of work now achieved in just six months. With the arrival of the internet in the late 20th century and the rapid evolution of mobile devices, social media has also been growing at exponential rates since the early 2000s transitioning society into a more digital, mobile and social media environment.

As it is shown [Figure.6] since 2009 the number of Internet users worldwide has skyrocketed, internet users increased to 44 million in 1995 and 413 million in the year 2000. Since then the growth of internet users has accelerated and reached 3.4 billion in 2016. The first recognizable social media site, Six Degree was created in 1997.
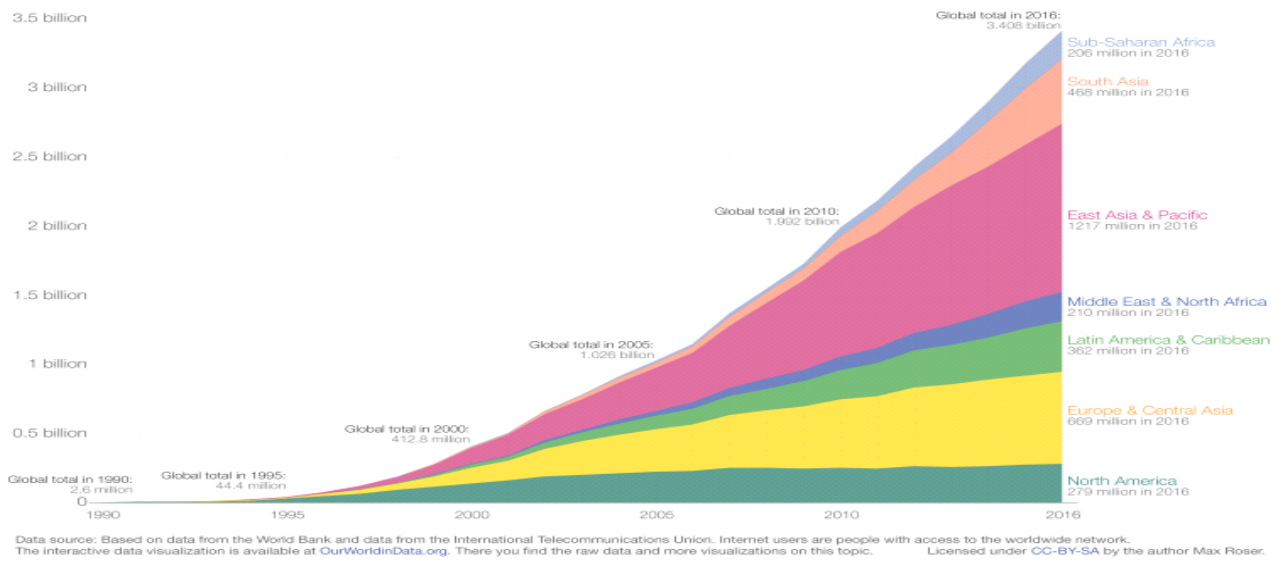
[Figure.6] Internet users by world region since 1990.

It enabled users to upload a profile and make friends with other users. Sites like *MySpace*and *LinkedIn* gained prominence in the early 2000s. *YouTube* came out in 2005—, follow by *Facebook* and *Twitter* in 2006. They both became available to users throughout the world. These sites remain some of the most popular social networks on the Internet as shows in [Figure.7] (Hendricks, 2013).
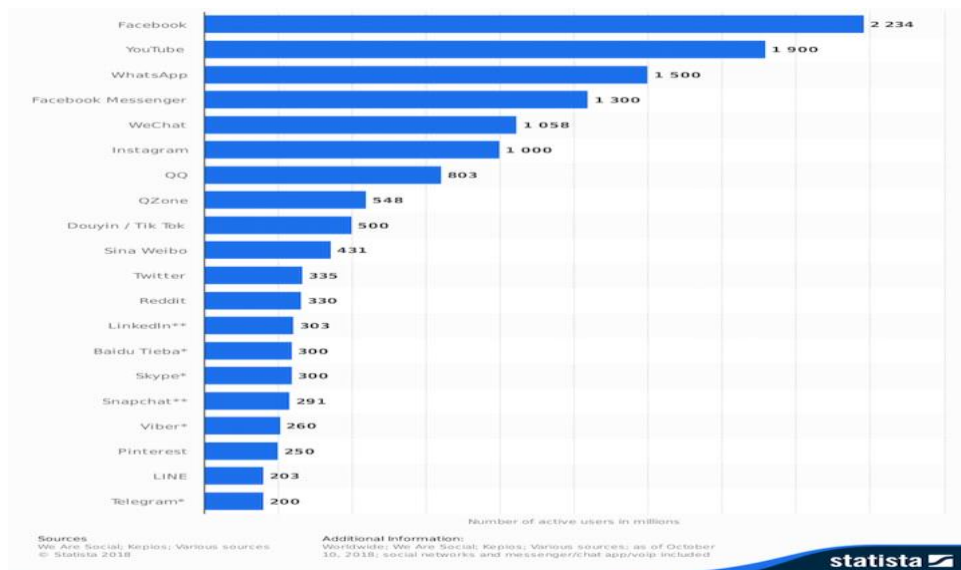
[Figure.7] Most Popular social networks worldwide as of October 2018 ranked by number of active users (in millions).

In this social media age, marked by technological innovation, organisations face new challenges, resulting in an adaptation to this platform-dominated media environment. It is more accessible to find news in social media rather than traditional news organisations. This accessibility is the result to its often timelier and less expensive reach of news and its further comment and share of them.

The chart at [Figure.8] is a report based on a YouGov survey of about 50,000 people across 26 countries: Facebook and other social media outlets have moved beyond being "places of news discovery" to become the place people consume their news, it suggests. (Wakefield, 2016).
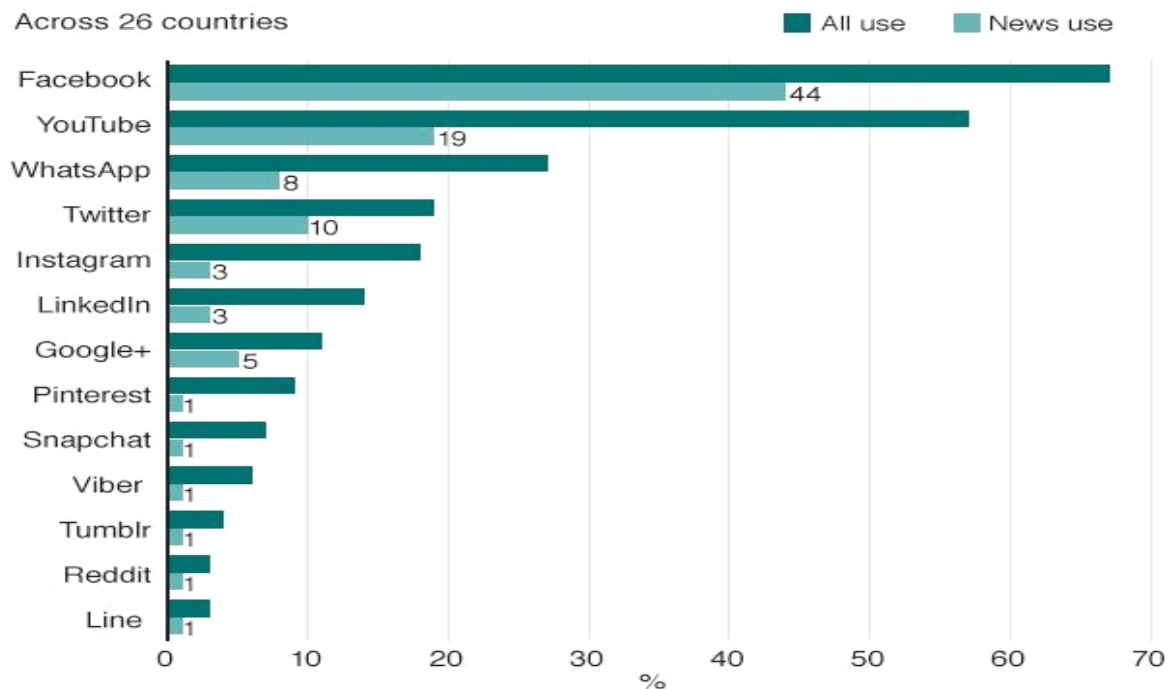
[Figure.8] Top social network for news.

Thanks to this fast and easy access to information, the quality of news on social media is lower than traditional news organisations. This low cost and fast dissemination of information through social media has led to an urgency to be informed every minute-. The insatiable appetite for fresh news has led to the multiplying and sharing of fake across all social platforms.

The Nature Communications Journal recently published the findings of a study conducted by Indiana University researchers. It contained an analysis of low-credibility stories posted on Twitter, and showed that 14 million messages spreading 400000 articles on Twitter during ten months in 2016 and 2017. 389,569 articles stemmed from low-credibility sources and- 15,053 articles from fact-checking sources were collected.

Furthermore was the level of public posts linking to these articles: 13,617,425 tweets linked to low-credibility sources and 1,133,674 linked to fact-checking. All of these were published on Nov 20th, 2018. Evidence showed that bots contribute significantly to the spread of low-credibility articles before they go viral. This is despite the fact that only 6% of accounts were identified as bots. However, it was enough to spread 31% of the low-credibility sources, and the *retweeting* The [Figure.9] shows the popularity and bot support for the top sources.
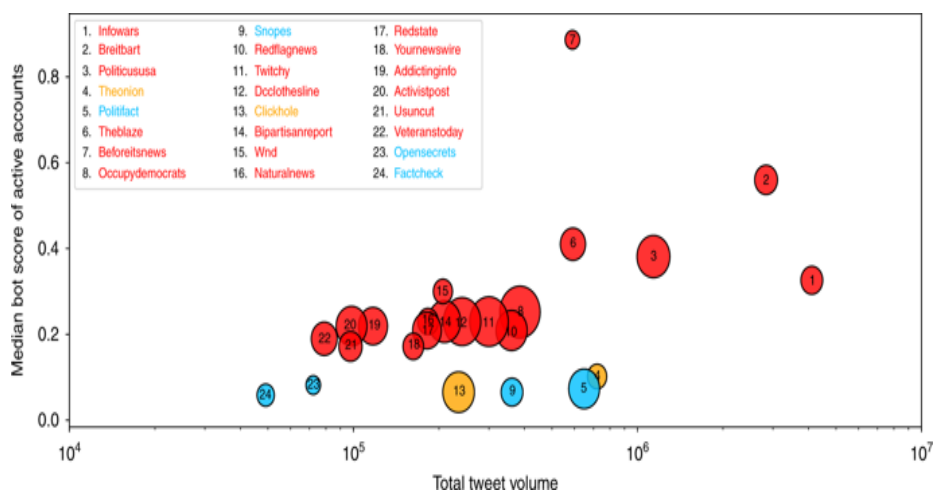
[Figure.9] Median bot score of active accounts.

Satire websites are shown in orange, fact-checking sites in blue, and low-credibility sources in red. Popularity is measured by total tweet volume (horizontal axis) and the median number of tweets per article (circle area). Bot support is gauged by the median bot score of the 100 most active accounts posting links to articles from each source (vertical axis). Low-credibility sources have greater support by bots, as well as greater median and/or total volume in many cases.

A *bot* is an automated application used to perform simple and repetitive tasks. Bots can also work in social network sites and simulate the internet users' behaviours in social networks, i.e., both are capable of different social interaction on Twitter that make may resemble the behaviours of people.

The functions of Bots are vast and include:

· Based on scripts they have the availability to reply to postings or questions from.

· They can contact users by sending them questions resulting in the exchange of communication and this way bots generate trust of this users.

· Generate debate by posting messages about trending topics.

Bot's algorithms allow them to respond to particular situation training from response patterns or input values and their resembling in people's behaviours helps to the propagation of fake news. They have the capability to search and retrieve information that has not been validated nor authenticated; they also post continuingly this non-authenticated information using strategies such as "trending topics" or "hashtags" to an audience.

As history shows the Fake News is a concept which has basically the intention of distort information by spreading it in different ways of communication to manipulate people. It has been around for long time and it is repeated constantly by government or powerful individuals.

At the present project Fake News will be defined as disinformation which is believed to be the most suitable term for the proposal. Fake news moulds people's perceptions and impacts people's reality by changing and confusing their thoughts and actions.

"Disinformation is defined as deliberately distorted information that secretly leaked into the communication process in order to deceive and manipulate."

Vladmir Bitman
KGB Director

Chapter 3. Twitter – Sentimental Analysis

Text Mining Techniques

Tidy Data Principles: In order to prepare the data a tidy data principles was used to handle data easier and more effective . At [Figure.10] a flowchart of a typical analysis using tidy data principles and [Figure.11] a small program code in R language using a small piece of text or token to and organizing it in a data frame (Slige, 2017a).
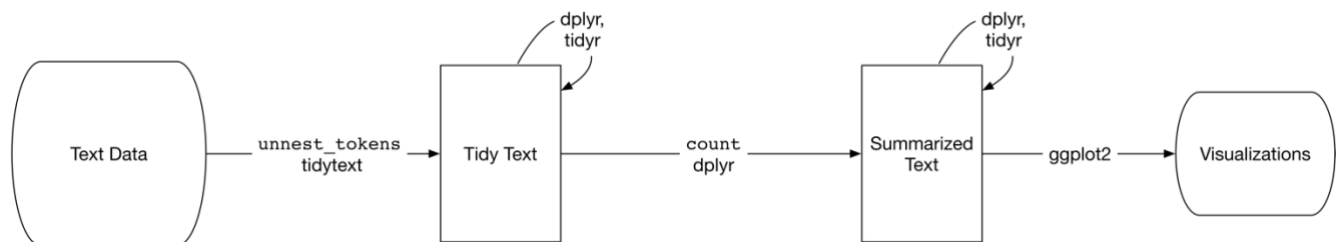
[Figure.10] Flowchart of a typical analysis using tidy data principles

```
> text1 <- c("The truth may be stretched thin, but it never breaks -",
+ "and it always surfaces above lies,as oil floats on water...  -",
+ "There is nothing either good or bad, -",
+ "but thinking makes it so...")
> text1
[1] "The truth may be stretched thin, but it never breaks -"
[2] "and it always surfaces above lies,as oil floats on water...  -"
[3] "There is nothing either good or bad, -"
[4] "but thinking makes it so..."
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> text_df <- data_frame(line = 1:4, text = text1)
>
> text_df
# A tibble: 4 x 2
  line text
  <int> <chr>
1     1 The truth may be stretched thin, but it never breaks -
2     2 and it always surfaces above lies,as oil floats on water...  -
3     3 There is nothing either good or bad, -
4     4 but thinking makes it so...
> library(tidytext)
```

[Figure.11] A Small R Program

As mentioned by Granskogen (2018), this kind of analysis can be done through two approaches:

The Linguistic approach: It is based only on the analysis of the content of the text itself. This approach involves using techniques that analyses frequency, usage, and patterns in the text (Granskogen (2018).

This approach is reasonable because news articles are usually intentionally created using inflammatory language and sensational headlines for specific purposes: i.e., to tempt readers to click on a link or to incite confusion. So, the linguistic analysis seeks to capture the writing styles in fake news articles. (Shu et al, 2016 and Chen et al, 2015).

The Contextual approach: This process incorporates most of the information that is not text. This includes data about users, such as comments, likes, re-tweets, shares and so on. It can also be information regarding the origin, both as who created it and where it was first published» (Granskogen, 2018). Some of the most common techniques that have been used for fake news detection are:

Linguistic approach: Sentiment analysis, Naive Bayes, Support Vector Machines

Contextual approach: Network analysis, Logistic regression, Trust Networks

To build the Machine Learning Model we need a dataset constituted by True and False news articles .This data or called the training data will be analysed using the techniques mentioned above such as sentiment analysis, Naive Bayes, Network Analysis, Logistic regression, etc. It is very important to notice that there is not a unique way for fake news detection. We can actually say that this is a recent problem that has been studied in the last years.

To generate an accurate Machine Learning Model, we have to perform a lot of tests using different techniques and approaches. That is why, at this point of the project, it is not possible to determine the exact methodology that will be used for building the Machine Learning Model.

In this project proposal we have decided to reduce the domain for fake news detection to Sport news. We have taken this decision because most of the research we have reviewed confirmed that Machine Learning Models for fake news detection have shown good results in closed domains (Conroy et al, 2015). However, more recent research indicates that a contextual approach must improve accuracy in open domains (Granskogen, 2018).

Sentiment Analysis

As stated before, our projects initial goal was to extract data from Twitter data that would later be used for the training and testing of our model. For this reason, the first practical interaction we carried out was sentimental analysis on twitter data.

When speaking of sentiments, we relate to feelings, attitudes, emotions, opinions, among others. Sentiment analysis refers to the practice of applying Natural Language Processing and Text Analysis techniques to identify and extract subjective information from a piece of text.

The models that were being created started from the very Basic Sentiment Analysis that follows a straightforward process:

1. Data extraction
2. Data processing
3. Sentiment analysis

Data Collection

The first time that we needed to start building our models was to collect data that we would be working with. Social media it has become a medium where people express their interests, share theirs view, displeasures, among others., therefore, analysing this data was the perfect fit for our project. We decided to work with Twitter data because Twitter developed an API that allows to extract tweets posted by users and their underlying metadata in structured format which can be easily analysed.

For this step it was necessary to create a Twitter application which will provide with customer and access key that will be needed to connect to R in order to extract the required tweets.[Figure.12]

```
# Authonitical keys
consumer_key <- '**********************'
consumer_secret <- '**********************'
access_token <- '**********************'
access_secret <-'**********************'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

tweets_h <- searchTwitter("#halamadrid", n=100,lang = "en")
tweets_f <- searchTwitter("#fcbarcelona", n=100,lang = "en")
tweets_m <- searchTwitter("#messi", n=100,lang = "en")
tweets_r <- searchTwitter("#realmadrid", n=100,lang = "en")
```

[Figure.12] Twitter Authentication setup in R code

Data Processing

Data processing requires a few steps in order to make it readable and easy to work with, the first thing we need to do is to convert the data to a data frame, that results in something like:

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID
1	Is #Moussa Wague destined to become another forgot...	FALSE	0	NA	2019–05–05 18:51:11	FALSE	NA	1125110710848184320	NA
2	RT @FootyGraphic: 🔴🔵 The (red–)blue print for that...	FALSE	0	NA	2019–05–05 18:49:06	FALSE	NA	1125110184374939650	NA
3	Hi, If you need any unique logo or graphics design pl...	FALSE	1	NA	2019–05–05 18:48:28	TRUE	NA	1125110027356758016	NA
4	@ryghtan #griezmann #fcbarcelona tic toc tic toc	FALSE	0	ryghtan	2019–05–05 18:48:22	FALSE	1125103892403036172	1125110001184518146	3635467035
5	Salah a doubt, Firminho out for Liverpool in Champio...	FALSE	0	NA	2019–05–05 18:45:07	TRUE	NA	1125109184012226561	NA
6	RT @FootyGraphic: 🔴🔵 The (red–)blue print for that...	FALSE	0	NA	2019–05–05 18:29:49	FALSE	NA	1125105334232678406	NA
7	Andy Robertson all the way..literally #FCBarcelona htt...	FALSE	0	NA	2019–05–05 18:18:04	FALSE	NA	1125102375981670401	NA

[Figure.13] Tweets converted into data frame

The part to be processed from the data frame is the text column, as shown in the pictures, it contains plenty of special character and unnecessary data that is not relevant, hence, it is vital to process this data before we move on to the analysis. This processing part consists in:

- Convert all character to lower case.

- Replace blan space "rt".

- Replace @username

- Remove punctuation.

- Remove links.

- Remove tabs

- Remove blank spaces at the beginning.

- Remove blank spaces at the end.

- Remove stop words.

Stop words are a set of words commonly used in any language. This last processing step is critical to applications like the one we are creating, by removing them the model can focus on the important words for our analysis.

Data Analysis

One way of finding frequents an important term being used in the analysed data is using word clouds, which is an image composed of words used in the text, the bigger and bolder it appears in the words cloud the more specific this word appears in the text.

[Figure.14] Word cloud plotted from tweets

To analyse the sentiments on the tweets, the inbuilt sentiment analyser in R, which uses the NRC sentiment dictionary to calculate the presence of eight different emotions and their corresponding valence in a text was used. The following plots now show the results on the sentiment analysis for each hashtag extracted.
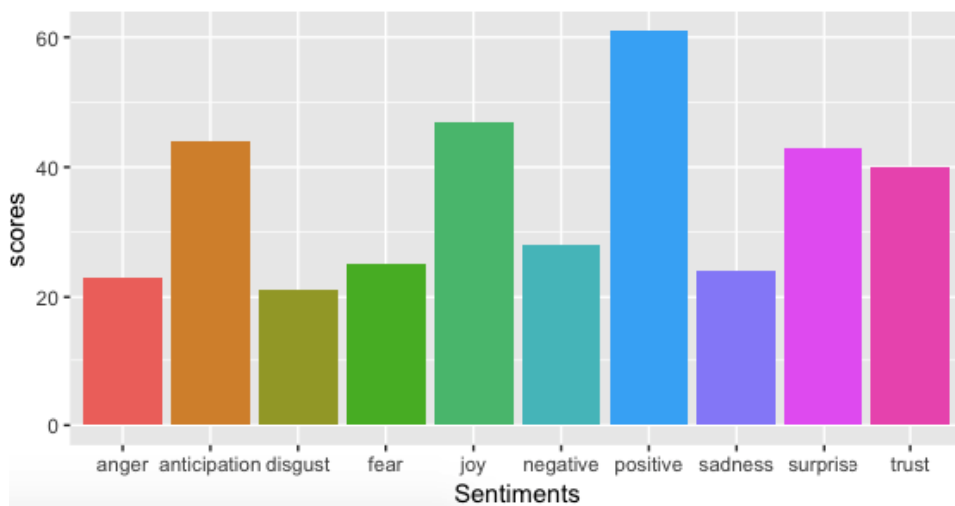
[Figure.15] Sentiments of people behind the tweets on #halamadrid
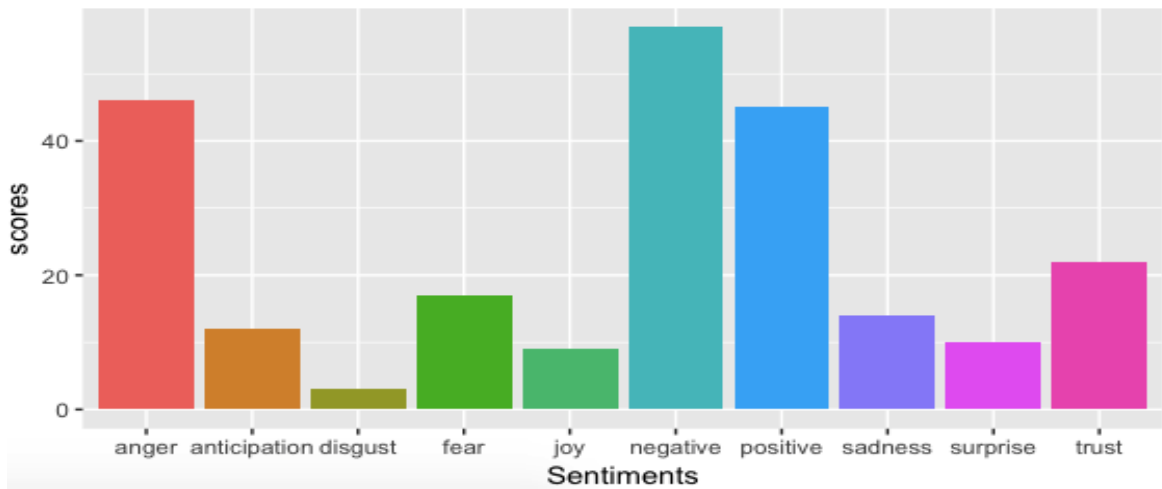
[Figure.16] Sentiments of people behind the tweets on #fcbarcelona
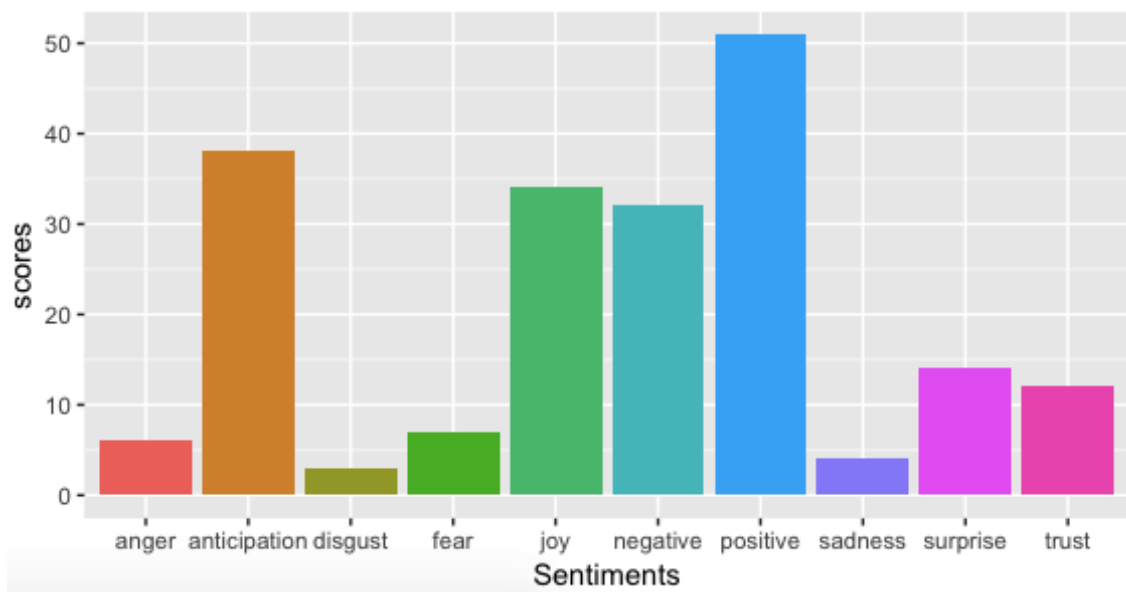
[Figure.17] Sentiments of people behind the tweets on #messi

Chapter 4. Training Model

Procedure

Problem Definition: Classify News articles as «fake» or «reliable»

As we have seen in previous sections, we are facing a Text classification problem. We want to classify News articles as «fake (1)» or «reliable (0)». The function that we want to build can be expressed this way:

Equation 1:

$$F(a) = \begin{cases} 1, & \text{if } a \text{ is a piece of fake news} \\ 0, & \text{if reliable} \end{cases}$$

Where α is the text of the article we want to verify its authenticity.

The solution: Supervised Machine Learning

We are implementing a Supervised Machine Learning Algorithm to build a Model (a function) that will take an Input Text (the news article) and will return a Output Tag: «fake (1)» or «reliable (0)».

Supervised Learning:

- Is the task of learning a function (the model) from labelled training data. The training data consists of a set of input-output pairs. We call it labelled data because we know the output.
- In our case, the input is the text of the news article and the output is the authenticity tag («fake (1)» or «reliable (0)»).
- The Function obtained will be used for mapping new inputs. In other words, the function will return the authenticity tag («fake (1)» or «reliable (0)») for new input data (not labelled news articles).
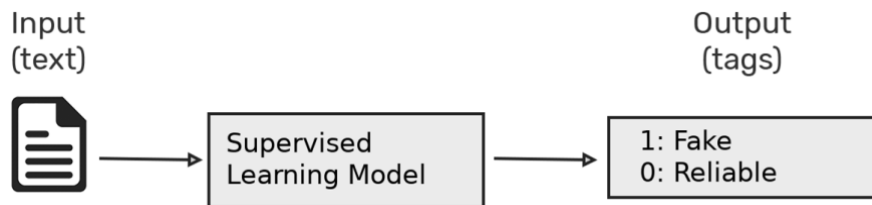
[Figure.18] Supervised Learning Model modified from MonkeyLearn (2019)

The training dataset

In this project we used four different labelled fake news datasets (Section Datasets), we called them:

- *Kaggle Fake News Dataset 1*

- *Kaggle Fake News Dataset 2*

- *Victory University Dataset*

- *Gofaas Dataset:* This is a small but very trustworthy dataset of 511 news articles. It has been built by our team in order to have reliable data that will be used as Test data to validate the models built with other datasets.

The first three datasets were built for similar fake news detector projects and are available online. In dataset you can find a detailed explanations of the datasets. To simplify the explanation of the process followed to train our model, consider the example data schema shown in [Figure.19]. This will be our training data from which our Machine Learning Algorithm will build the Classification Function (the model). Each input-output pair (News Article - Label) will be used for the algorithm to learn the writing styles of fake or reliable news articles.

Source	News article	Label	
https://web.archive.org/web/20170626190612/http://snoopack.com/2017/06/23/just-today-trey-gowdy-new-director-fbi/	A huge new development in Washington, D.C. may be putting American security back on track. This is great news for American patriots everywhere...	1	70%: Train Data
https://www.aljazeera.com/indepth/features/dying-sea-refugees-learn-love-water-190416201531713.html	After almost dying at sea, refugees learn to love the water.Survivors in Sicily, some of whom witnessed others perish, are integrating with locals and getting back in the sea...	0	
https://newspunch.com/irma-destroy-new-york-city/	All 5 Boroughs of New York City and most of New Jersey could be totally destroyed on September 10 by Hurricane IRMA, according to computer simulations...	1	
https://steemit.com/news/@gaucher/donald-trump-dead-from-a-fatal-heart-attack	Donald Trump was pronounced dead this morning following what some are describing as a violent heart attack. The world famous businessman and TV personality was...	1	
https://www.express.co.uk/celebrity-news/1110320/Emilia-Clarke-Game-of-Thrones-Instagram-season-8-premiere-pictures	EMILIA CLARKE warned Game of Thrones fans they would "freak out" when they saw the upcoming episodes in season 8 as she shared behind-the-scenes snaps...	0	
http://www.mysterious-times.com/2018/04/30/bill-gates-outlines-2018-plan-to-depopulate-the-planet/	Bill Gates has doubled down on his goal to depopulate the planet, using deceitful Orwellian doublespeak in a new video to bamboozle...	1	
https://www.theguardian.com/media/2019/apr/16/the-claims-against-assange-and-us-attitudes-to-international-law	Dr Kevin Bannon says the US demeans the international criminal court, while Dr Gill Gregory is troubled by...	0	
https://www.bloomberg.com/news/articles/2019-04-17/is-amazon-too-powerful-its-merchants-are-starting-to-wonder?srnd=premium-europe	Fifteen years ago, Jason Boyce made a bet on Amazon that changed his life for the better. He started selling...	0	30%: Test Data
https://eu.usatoday.com/story/news/politics/onpolitics/2017/07/15/fox-news-shepard-smith-lie-after-lie-after-lie-russia-meeting/481751001/	Fox News host Shepard Smith slammed what he called "lies" and "deception" pushed by Donald Trump Jr. in a fiery Friday...	1	
https://web.archive.org/web/20170731052757/http://politicspaper.com:80/breaking/breaking-hillary-clinton-third-heart-attack-docs-says-wont-survive/	Hillary Clinton had a third and most-likely fatal heart attack this afternoon after spending the morning being told...	1	

[Figure.19] Example of the dataset used to train the model.

The approach

There are different ways to face a text classification problem with supervised Learning. We are going to implement a simple approach known as «bag of words». In this method, the text (in our case the news article) is simply defined by the set of words that composed it. This way, the Machine Learning Algorithm will basically learn (based on the training data) which words are usually in a fake or reliable news article.

Procedure to build the Model

Because one of the goals of this project is to evaluate the performance of different algorithms, we built models for each dataset using four algorithms. Further description is provide at Algorithms Section.

- **Naive Bayes (NB):** Implemented through the *e1071* R Library
- **Support vector machine (SVM):** We used the *RTextTools* R Library, which depends on e1071.
- **Random forest (RF):** We used the *RTextTools* R Library, which depends on *RandomForest.*
- **Gradient Boosting (XGB):** Implemented using the XGBoost R Library.

So, we obtained 4 Models for each dataset:

- *Kaggle Fake News Dataset 1:*
  - *Model_NB*
  - *Model_ SVM*
  - *Model_ RF*
  - *Model_ XGB*
- The same for the other datasets

In this section, we explain the procedure followed to build the XGBoost Model using the *Victory University Dataset*.

The Models were built following these steps:

1. **Splitting the Dataset into Training and Test data:**

   - <u>70% of the dataset will be used as Train data</u>: The model will be built using this portion of the data.
   - <u>30% of the dataset will be used as Test data</u>: We will use the model's build using the 70% of the dataset to classify the 30% that hasn't been used to train the model. This will allow us to calculate the accuracy of the classification made with our model. We will explain later how the accuracy is calculated.

```
# =============================================================================== =
#  Splitting the data into Train and Test data ----
# =============================================================================== =
samp_id = sample(1:nrow(data),
                 round(nrow(data)*0.7),   # 70% will be used as Training data
                 replace = F)
train = data[samp_id,]
test  = data[-samp_id,]
```
[Figure.20] R Script used to split the data into Train and Test data.

2. Cleaning the data:

We need to remove everything that doesn't contribute to the analysis-from the data before we run the algorithm. The cleaning of the data used in this project includes:

- Convert to lower-case letters.
- Remove punctuation.
- Remove numbers.
- Remove blank space

Remove stopwords: Stop words are commonly used words (a, an, and, the, this, those). The reason why stop words should be removed is that we can focus on the words that really differentiate the articles and not the words that are in all the articles.

- Stemming Words: Trimming words such as 'calling', 'called' and 'calls' to call.

```r
# ================================================================================ =
#  Cleaning the data ----
# ================================================================================ =
library(tm)
library(SnowballC)

text <- data$text
text <- tolower(text)
text <- removePunctuation(text)
text <- removeNumbers(text)
text <- removeWords(text, stopwords("en"))    # stopwords
text <- stripWhitespace(text)                 # Remove blank space
text <- wordStem(text, language = "english")  # Stemming Words
```
[Figure.20] R Script used to clean the data.

3. Building the Document-Term Matrix:

To be able to run a Machine Learning algorithm, we first need to transform each news article into a numerical representation in the form of a vector. In [Figure.21], we show how build the DTM. This matrix will be the numerical representation that a Machine Learning algorithm is able to understand. As you can see, each column within this matrix represents a word in the training data. Thus, each document is defined by the frequency of the words that are in the dictionary composed for all the terms in our data.
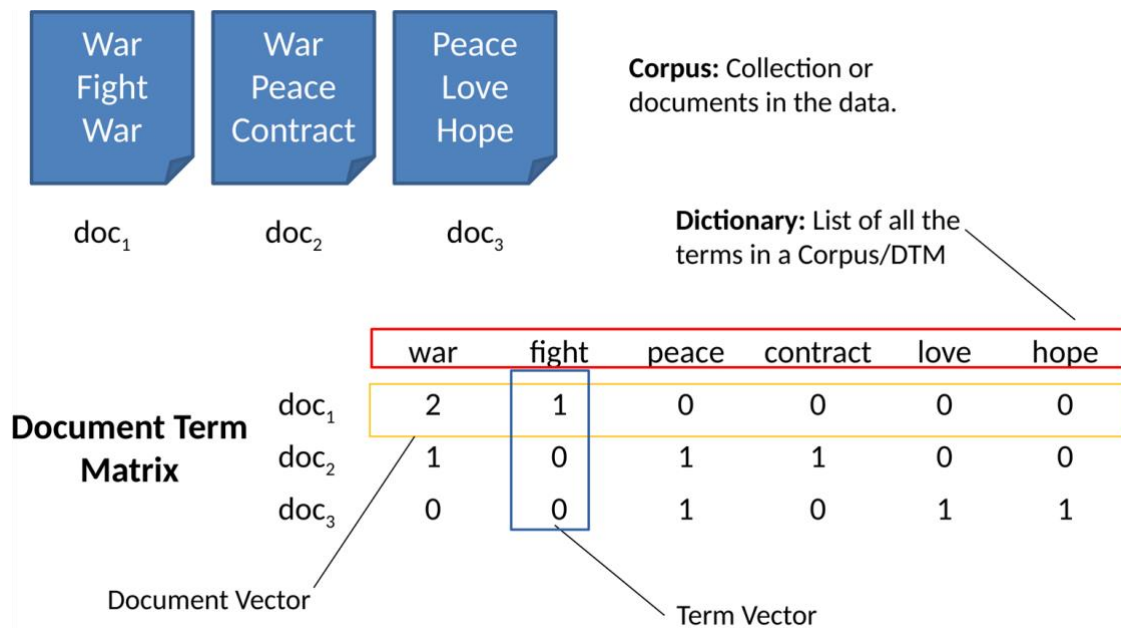
[Figure.20] The Document Term Matrix (DTM).

```
# ============================================================================= =
#  Building the Document-Term Matrix ----
# ============================================================================= =
library(text2vec)
library(xgboost)


# * Train data ----

# Vocabulary of tokens for the train data
vocab_train <- create_vocabulary(itoken(text_train,
                                        preprocessor = tolower,
                                        tokenizer    = word_tokenizer))
# Saving the vocab_train
saveRDS(vocab_train, file = vocab_train_filename)


# Document-term matrix for the train data
dtm_train_xgb <- create_dtm(itoken(text_train,
                                   preprocessor = tolower,
                                   tokenizer    = word_tokenizer),
                           vocab_vectorizer(vocab_train))
```

[Figure.21] R Script used to build the DTM.

4. Model Building:

This is the point where we create the function using a Machine Learning Algorithm. In [Figure.22], we show the code snippet we used to build the XGBoost Model. First we need to add the labels to our DTM to create the Matrix that we pass to the XGBoost function. We implement a decision trees based XGBoosting using the following parameters:

- *objective = "binary: logistic"* : To train a binary classification model (this is the case of fake news detection).

- *max.depth = 7* : Maximum depth of a tree.

eta = 0.01 : Step size shrinkage used in update to prevents overfitting.

- *nrounds = 10000 : The number of rounds for boosting.*

The model created with the XGBoost function has been saved into the *model_XGB* variable. This model will be used later to classify new data.

```
# ============================================================================= =
#  Xgboost model building ----
# ============================================================================= =

# Turn the DTM into an XGB matrix using the labels that are to be learned ----
xgbMatrix_train <- xgb.DMatrix(dtm_train_xgb, label = labels_train)


# Parameters for xgboost
xgb_params = list(
  objective = "binary:logistic",
  eta = 0.01,
  max.depth = 7,
  eval_metric = "auc")


# Creating the «xgb model» obj ----
model_XGB <- xgboost(data = xgbMatrix_train, params = xgb_params,
                     nrounds = 10000, print_every_n = 500)
# Saving the model
saveRDS(model_XGB, file = model_XGB_filename)
```

[Figure.22] R Script used to build the DTM.

5. Classifying news articles using the *model_XGB* and calculating the accuracy of the model:

Now that we have a model, we can use it to classify new data. Remember that we count with the labels of the data (fake (1) or reliable (0)). So, the goal is to verify the accuracy of the model by comparing the results returned for the Model with the true labels.

i.   **The accuracy of the Model over the Training data:**

A first measure of the accuracy of the model can be calculated by classifying the same data that has been used to build the model. Our Model must perform very well when classifying this data.

This accuracy doesn't really give a good measure of the efficiency of the model to classify new data, but it is used to validate that the algorithm has been correctly implemented.

It is expected to get an accuracy close to 100% when using the Train data. In [Figure.23], we show the accuracy and Confusion Matrix we got over the training data.

ii. The accuracy of the Model over the Test data:

A second measure of the accuracy of the model is calculated by classifying the news articles of our Test Data. This data was not used to build the Model. So, the results of the classification over the Test Data give us a good measure of the accuracy of the model.

When we used the Test data to calculate the accuracy, it must be noticed that, even if the Test data was not used to build our model, it comes from the same dataset of our Training data. It is therefore very likely that the news articles come from similar sources and have a similar writing style. That is why, when calculating accuracy using test data from the same dataset, it is also expected to get a good accuracy. The accuracy and Confusion Matrix is shown in [Figure.23].

iii. The accuracy of the Model over another dataset (the Gofaas Dataset):

The best measure of the accuracy of a model, is the one calculated by classifying data that is not related at all with the dataset used to build the model. In this project we use the Gofaas Dataset as our final Test data.

```
> XGBoost_VictoryDataset_GofaaasDataset
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 197   1
         1 107 206

               Accuracy : 0.7886
                 95% CI : (0.7507, 0.8233)
    No Information Rate : 0.5949
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5946
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6480
            Specificity : 0.9952
         Pos Pred Value : 0.9949
         Neg Pred Value : 0.6581
             Prevalence : 0.5949
         Detection Rate : 0.3855
   Detection Prevalence : 0.3875
      Balanced Accuracy : 0.8216

       'Positive' Class : 0
```

[Figure.23] *Accuracy and Confusion Matrix of the model built using XGBoost and the Victory University Dataset. These are the result obtained when classifying the articles of the Gofaas Dataset. Notice the good results we got when using the model created to classify data from a different dataset that the one used to build the model.*

Datasets

Kaggle Fake News Dataset 1

It is a full train dataset size of 20386 unique values provided from Kaggle competitions. Follows below its attributes:

- Id: unique id

- Title

- Author

- Text

- Label
    - 1 – Reliable
    - 0 – Unreliable

There are two types of articles fake and real News. It is basically constituted PolitiFact news articles and it is written by a different number of distinct authors about 89%. Only 1% of the articles are written by Pam Key and 9% by nan. Train.csv dataset is also designed to build a system to identify unreliable news article. It was the very first dataset used to build the model.

[Figure.24] train.csv dataset provide by Kaggle.com

Kaggle Fake News Dataset 2

It is a fake news detection dataset; size of 2831 unique values with the following attributes:

- URLs
- Headline
- Body
- Label
  - 1 – Reliable
  - 0 – Unreliable

There are two types of articles fake and real News. The nature of the dataset is PolitiFact and it contains: bbc, reuters, nytimes, cnn and others reliable URLs sources and for beforeitsnews for fake. There are more 0 or fake news articles then reliable ones.

[Figure.25] data.csv provide by Kaggle.com

Victory University Dataset

The dataset contains two types of articles fake and real News. This dataset was created by The University of Victoria, Canada.

Reliable articles were obtained by crawling articles from Reuters.com (News website), unreliable articles were collected from unreliable websites that were flagged by PolitiFact (a fact-checking organization in the USA) and Wikipedia. The dataset contains different types of articles on different topics.

Each article contains the following information:

- Type
- Text
- Label
  - 0 – Reliable
  - 1 – Unreliable

The initial dataset provided by the University was reviewed and cleaned, as there were some articles that contained only sources to videos, data was reviewed carefully reviewed.

 The [Figure.26] gives a breakdown of the categories and number of articles per category.

| News | Articles Size | Subjects | |
		Type	Articles Size
Reliable	20233	World	9697
		Politics	10536
Unreliable	15076	GovernmentNews	1089
		Middle-east	521
		US News	341
		left-news	2881
		politics	3780
		News	6464
TOTAL	35309		

[Figure.26] Dataset was created by The University of Victoria, Canada

Gofaas DataSet

The dataset contains two types of articles fake and real News, it contains only political types of articles. The original dataset was created from GitHub repository Fake Newsnet, which contains two files: politifact_fake.csv, which is the file from where unreliable articles were taken from. Each entry was reviewed manually to confirm if the URL was active and whether the article could be considered indeed unreliable. Reliable articles were obtained by crawling articles from trusted sources such as: Reuters.com (News website), Washington Post, Guardian, BBC, among others.

Each article contains the following information:

- Source / URL
- Text
- Label
  - 0 – Reliable
  - 1 – Unreliable

The [Figure.27] gives a breakdown of the categories and number of articles per category.

News	Articles Size
Reliable	304
Unreliable	207
TOTAL	511

[Figure.27] Gofaas an original Dataset made to check accuracy of the model

Algorithms

According to Hastie, (2014) it is an accurate state of dominance of the following algorithms:

XG-Boost > Boosting > Random Forest > Bagging > Naïve Bayes Single Tree

XG-Boost and Random Forest are based on Trees. Naïve Bayes and Trees comes from probability and as such they all rely on Bayes' Model in order to predict if an event is happening. The Bayes' Theorem or Bayes' Rule is a way to figure out conditional probability, in other words it to find out if there is a relationship between one or more events.
According to Glen (2014) the Theorem was named after English mathematician Thomas Bayes (1701-1761).

Naïve Bayes Algorithm

Naïve Bayes is based on the Bayesian theorem, there in order to understand Naïve Bayes it is important to first understand the Bayesian theorem. The name Naive is used because it assumes the features that go into the model is independent of each other. That is changing the value of one feature, does not directly influence or change the value of any of the other features used in the algorithm. In order to understand how Naïve Bayes works first is necessary to understand what 'Conditional Probability' is and what is the 'Bayes Rule'. Mathematically, Conditional probability of A given B can be computed as:

Equation 2:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Likelihood — Class Prior Probability — Posterior Probability — Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

[Figure.28] Bayes Rules

- P(c|x) is the posterior probability of *class (target)* given *predictor (attribute).*
- P(c) is the prior probability of *class.*
- P(x|c) is the likelihood which is the probability of *predictor* given *class.*
- P(x) is the prior probability of *predictor*.

Naïve Bayes simple example for Fake News Detection

To illustrate the way Naïve works, 4 segments from rows will be taken from the dataset "goofaasDataset.csv", which a dataset created by our team with proven authentic labelled as 0 and non-authentic labelled as 1 pieces of information.

	X	url	author	text	label
103	103	http://breaking13news.com/malia-obama-arrested-...	BREAKING13NEWS	Malia Obama may have done irreparable harm to her ...	1
104	104	http://thenewyorkevening.com:80/2017/04/10/brea...	NY Evening	Malia Obama, who has decided that as an adult she w...	1
105	105	https://www.bbc.com/sport/football/47939085	Matthew Henry	Manchester United's Champions League run ended in ...	0
106	106	https://www.bbc.com/news/av/world-us-canada-47...	bbc	Marine crawls to finish Boston Marathon for fallen co...	0

[Figure.29] Rows 104-105 from the will be taken for our training data

TRAIN DATA	LABEL
Malia Obama who has decided that as an adult she…	1
Manchester United's Champions League run ended…	0
Marine crawls to finish Boston Marathon from fallen…	0
TEST DATA	LABEL
Malia Obama may have done irreparable harm to…	?

[Figure.30] Row 103 will be used as our testing data.

Naïve Bayes is a probabilistic classifier, thus, the main goal here is to calculate the probability that the sentence from test data "Malia Obama may have done irreparable harm to…" is 1 (unreliable) and also the probability that is 0 (reliable).

Written mathematically it would translate to:

P(1| Malia Obama may have done irreparable harm to) -
the probability that the label of the sentence is non authentic.

The first step to be done when creating a machine learning model is to determine what it will be used as a feature. Features are pieces of information taken from the text and passed to the algorithm. In our case, news is composed of words; those words. Those words have to be converted into numbers in order for the model to do the calculations. This is accomplished by using word frequencies, which means that the order and sentence construction is not important, and each new is treated as a set of the words it contains, the feature in this case will be the count of each of these words.

Then it is necessary to transform the probability we want to calculate into something that can be calculated using word frequencies, to achieve this basic property of probabilities and Bayes' Theorem will be used.

Equation 3:

$$P(1|\text{ Malia Obama may have done irreparable harm to}) =$$
$$\frac{P(\text{Malia Obama may have done irreparable harm to}|1) \times P(1)}{P(\text{Malia Obama may have done irreparable harm to})}$$

Our main goal is to predict the label that has a bigger probability; therefore, the divisor can be discarded which is the same for both labels and just compare:

$$P(\text{Malia Obama may have done irreparable harm to}|1) \times (1)$$
With
$$P(\text{Malia Obama may have done irreparable harm to}|0) \times (0)$$

Probabilities could be calculated by counting how many times "Malia Obama may have done irreparable harm to" appears in the 1 label, divided by the total and obtain P(Malia Obama may have done irreparable harm to|0), although the probability in this case would be 0 because the sentence does not appear in the training data.

In this part is when the Naïve part comes into play, assuming that every word in the text is independent of the other ones, meaning that we stop looking at entire sentences but rather at individual words. And our formula would be written as:

Equation 4:

P(Malia Obama may have done irreparable harm to) = P(Malia) x P(Obama) x
P(may) x P(have) x P(done) x P(irreparable) x P(harm) x P(to)

The next step is to apply what we had before.

Equation 5:

P(Malia Obama may have done irreparable harm to|1) =
P(Malia|1) x P(Obama|1) x P(may|1) x P(have|1) x P(done|1)
x P(irreparable|1) x P(harm|1) x P(to|1)

If any of those individual words appear in the training data, we could proceed to calculate the probabilities by counting the training data.

First a priory probability is calculated of each label: for a given sentence in our training data, the probability that it is unreliable P(1) is 1/3. Then, P(Not 0) is 2/3.

Then, calculating P(Amalia|1) means counting how many times the word "Amalia" appears in unreliable texts(one) and divided by the total number of words in unreliable (1), therefore:

P(Amalia|1) = 1/10

When we come across with words that don't appear in any text, it will mean that P(may|1) = 0, multiplying 0 with other probabilities it will result in a zero-result nullifying the whole calculation. This issue is solved by implementing *Laplace smoothing*, which just add one to every count so it is never zero. To balance this, we add the number of possible words to the

divisor, so the division will never be greater than 1. In this case our possible words are:

Total possible words = 24

"Manchester" "United" "Champion" "League" "run" "ended" "Marine" "crawls" "to" "finish" "Boston" "Marathon" "from" "fallen" "Malia" "Obama" "who" "has" "decided" that" "as" "an" "adult" "she"

Words from labels 0 = 14
"Manchester" "United" "Champion" "League" "run" "ended" "Marine" "crawls" "to" "finish" "Boston" "Marathon" "from" "fallen"

Words from labels 1 = 10
"Malia" "Obama" "who" "has" "decided" that" "as" "an" "adult" "she"
Applying smoothing we get that:

Probability that "Malia Obama may have done irreparable harm to" is unreliable (0) is: **29.41%**

Word	P(word|1)	Total
Amalia	1 +1 / 10 +24	0,058823529
Obama	1 +1 / 10 +24	0,058823529
may	0 +1 / 10+24	0,029411765
have	0 +1 / 10+24	0,029411765
done	0 +1 / 10+24	0,029411765
irreparable	0 +1 / 10+24	0,029411765
harm	0 +1 / 10+24	0,029411765
to	0 +1 / 10+24	0,029411765
Total for P(word|1)		**0,294117647**

[Figure.31]. Probability results of 29.41% of labels 0

Probability that "Malia Obama may have done irreparable harm to" is reliable (1) is: **23.68%**

Word	P(word|0)	
Amalia	0+1/14+24	0,026315789
Obama	0+1/14+24	0,026315789
may	0+1/14+24	0,026315789
have	0+1/14+24	0,026315789
done	0+1/14+24	0,026315789
irreparable	0+1/14+24	0,026315789
harm	0+1/14+24	0,026315789
to	1+1/14+24	0,052631579
Total for P(word|1)		**0,236842105**

[Figure.32]. Probability results of 23,68% of labels 1

Although the results are very tight, our classifier would assign the piece of information to the label 1.

Single Tree (Hastie, 2014).

To understand the tree concept by applying the Bayes' Model. It is given a binary problem where, green is equal to 1 and red equals 0. The Bayes Error Rate of 0.25 can be visualized as a black ellipse. The model wants to classify red by given green data just using X1 and X2 coordinates.

It could also be interpreted as a given dataset or piece of text (green) in order to predict fake news(red) based on the Bayes' Decision Boundary. The Bayes' Model predicts as if we knew the whole population which the data came.
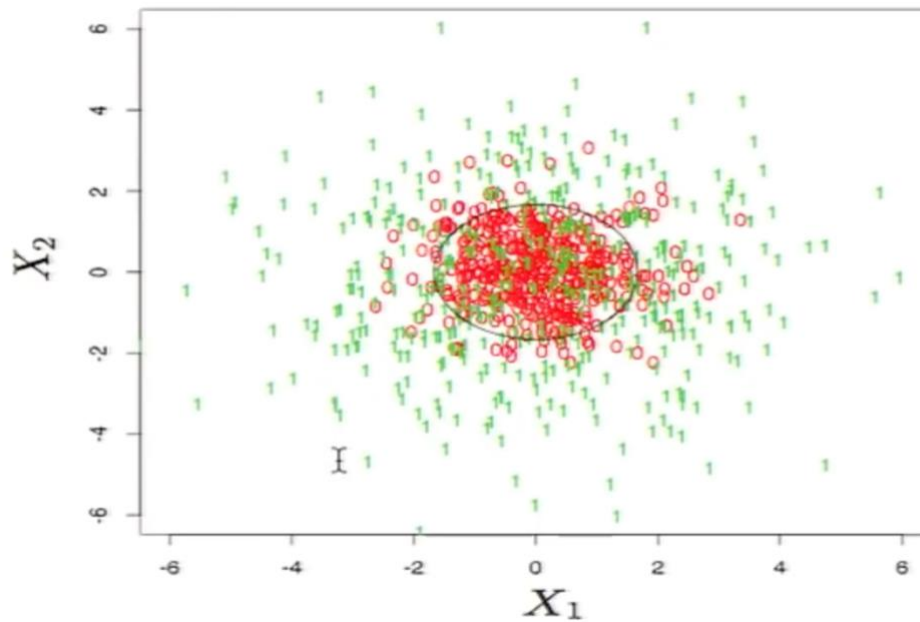
[Figure.32]. Bayes Error Rate of 0.25 illustrated by the black ellipse (Hastie, 2014)

Decision tree algorithm provide a possibility of narrow down the Boundary by decreasing the Error Rate. The mechanism asks an series of questions. It comes with an pair of code X1 and X2 and asks: *is X2 less then 1.6711?* if yes go left if no go right and keep looping over until it gets down to an terminal node and says you are inside (red) or outside (green)
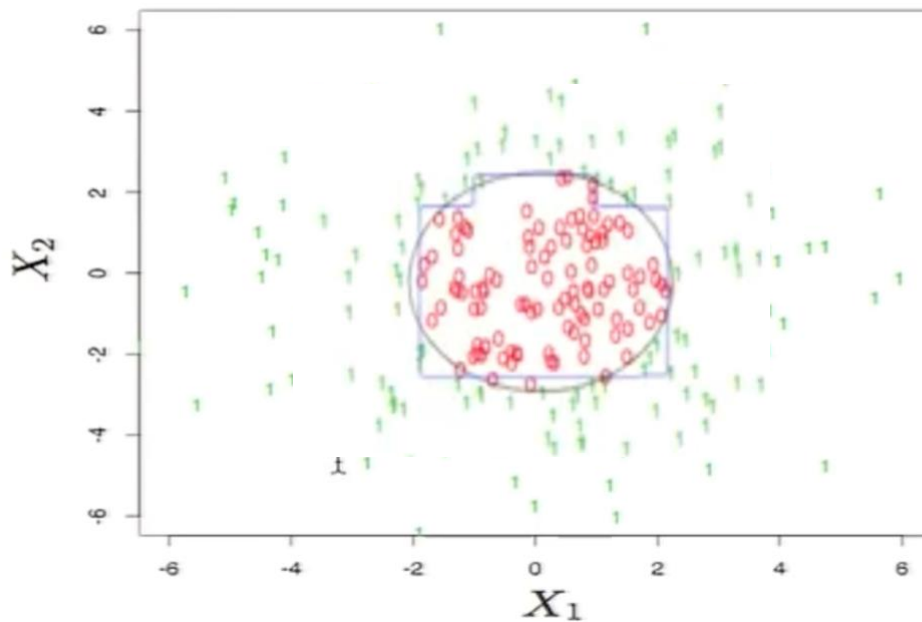
[Figure.32]. Representation of the classification tree and the error rate decreasing from 0.25 to 0.073. *(Hastie, 2014)*

The Trees advantages are: able to handle huge datasets, mixed predictors (quantitative and qualitative), easily ignore redundant variables, handle missing data through splits, small trees are easy to interpret. Whereas large trees are hard to interpret and the prediction performance is often poor because of high variance. Classification trees can be simple, but often produce noise(bushy) or weak (stunted) classifiers.

Improving performance Bagging < Random Forest < Boosting

Bagging and Random Forest are a way to bring down the high variance problem by averaging many different trees. Bagging does that by shaking the data up with different samples, and average then all to decrease the variance but the problem of variables correlation still present. Random Forest takes randomly different samples of data or uncorrelated variables, which decrease even more the variance problem.

Random Forest is a refinement of bagged trees at each tree split, a random sample of *m* feature are considered for splitting. Typically

Equation 6:

$$m = \sqrt{P} \text{ or } \log2\varphi,$$

where φ is the number of features. For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored. This is called the "out-of-bag" or OOB error rate. (Hastie, 2014).

Contributors (Louppe, 2014)

1. **Kwok and Carter, 1990**: an empirically observation of averaging multiple decision trees with different structure, consistently produces better result than any of the constituents of the ensemble.

2. **Breiman, 1994**: was one of the earliest to show, both theoretical and empirically, that aggregating multiples versions of an estimator into an ensemble can give substantial gains in accuracy.

3. **Dietterich and Kong, 1995**: building upon 1, 2 and Kong contributors propose to randomise the choice of the best split at a given node by selection uniformly at random one of the 20 best splits of node t.

4. **Breiman, 1996**: Bagging fits many large trees to bootstrap-resampled versions of the training data, and classify by majority vote. a randomised variant of the tree induction algorithm that consists in searching for the best split at each node over a random subsample of the variants.

5. **Amit et al., 1997**: propose a randomised variant of the tree induction algorithm that consists in searching for the best split at each node over a random subsample of the

trees in which both the(ordered) variable to split on and the variants.

6. **Ho, 1998**: inspired from 2 (bagging) and 4 (random subsets of variables) contributors proposes with the Random Subspace(RS) method to build a decision forest whose trees are grown on random subsets of the input variables - drawn once, prior to the construction of each tree- rather than on all p variables.

7. **Breiman, 2001**: combines bagging with random variable selection at each node.

8. **Cutler and Zhao, 2001**: with Perfect Random Tree Ensembles propose to grow a forest of perfectly fit decision discretisation threshold are chosen at random.

9. **Geurts et al. 2006**: empirically show that the variance of the optimal cut-point v (in the case of ordered input variables) may indeed be very high, even for large sample sizes.

Random Forest Algorithm

Methodology

Random forest or random decision forest are an *ensemble learning* method for classification and regression. It basically means they are methods that generate many classifiers and aggregate their results. (Liaw and Wiener, 2002).

According to Louppe, 2014, inspired by Breiman et al., 1984 figure a *tree-structure model (or decision tree)* can be defined as a **model** represented by a rooted tree (often binary, but not necessarily)

Equation 7:
$$\varphi = X \rightarrow y$$

Where any node t represents a subspace $X_t \subseteq X$ of the input space, with the root node t_0 corresponding to X itself. Internal nodes t are labelled with a split taken from a set s_t of questions Q. It divides the space X_t that node t represents into disjoint subspaces respectively corresponding to each of its children

If φ is classification tree, then
$$\hat{y}_t \in \{c_1, ..., c_J\}$$
Expression 1

As such, the predicted value $\varphi(x)$ is the label of the leaf reached by the instance x when it is propagated through the tree by following the splits s_t

In other words, **classification** tree is used for **categorical data** and it is based on the **mode** or what happens most often. While **regression** tree is used for **continuous data** and it based on the **mean average**.

The present work has a binary classification problem. The data is either reliable(authentic)/unreliable(fake) it is determined at the source level, this step is necessary to ensure the model does not just learn the mappings from known sources to labels. The goal is to find models that can not only produce accurate predictions, but also be used to extract knowledge in an intelligible way.

Result interpretations

GoFaasDataset.csv: it is a dataset created manually by the GoFaas group to ensure reliably. It is was done in order to set labels the data accurately. The data structure is composed by 510 entries which 304(60%) are 0/reliable and 206(40%) are 1/unreliable.

Comparison of top unreliable and reliable sources by article frequency

Top Five Unreliable News Sources		Top Five Reliable News Sources	
Before It's News	2066	Reuters	3898
Zero Hedge	149	BBC	830
Raw Story	90	USA Today	824
Washington Examiner	79	Washington Post	820
Infowars	67	CNN	595

[Figure.33] The reliable and unreliable source were based on Gilda (2017).
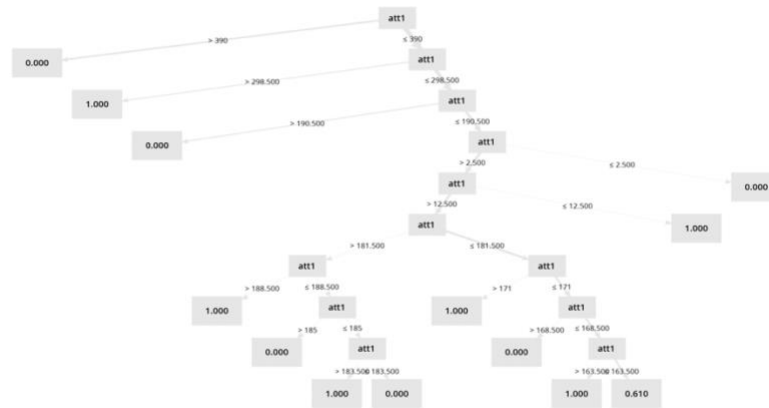
[Figure.34] Random Forest Model 1 create by RapidMinder Studio using Gofaas dataset.

Root node <- The attribute **att1** is the related to the row id or the row number is chosen as a root node.

Rule for splitting <- on **att1** value the algorithm checks if X1(root node) or att1 is greater or equal to 0.7 if so it goes left.

When to stop <- number of trees 100, criterion: least square, maximal depth 10.

```
RegressionTree

att1 > 390: 0.000 {count=134}
att1 ≤ 390
|   att1 > 298.500: 1.000 {count=79}
|   att1 ≤ 298.500
|   |   att1 > 190.500: 0.000 {count=84}
|   |   att1 ≤ 190.500
|   |   |   att1 > 2.500
|   |   |   |   att1 > 12.500
|   |   |   |   |   att1 > 181.500
|   |   |   |   |   |   att1 > 188.500: 1.000 {count=3}
|   |   |   |   |   |   att1 ≤ 188.500
|   |   |   |   |   |   |   att1 > 185: 0.000 {count=4}
|   |   |   |   |   |   |   att1 ≤ 185
|   |   |   |   |   |   |   |   att1 > 183.500: 1.000 {count=1}
|   |   |   |   |   |   |   |   att1 ≤ 183.500: 0.000 {count=3}
|   |   |   |   |   att1 ≤ 181.500
|   |   |   |   |   |   att1 > 171: 1.000 {count=12}
|   |   |   |   |   |   att1 ≤ 171
|   |   |   |   |   |   |   att1 > 168.500: 0.000 {count=3}
|   |   |   |   |   |   |   att1 ≤ 168.500
|   |   |   |   |   |   |   |   att1 > 163.500: 1.000 {count=7}
|   |   |   |   |   |   |   |   att1 ≤ 163.500: 0.610 {count=141}
|   |   |   |   att1 ≤ 12.500: 1.000 {count=10}
|   |   |   att1 ≤ 2.500: 0.000 {count=3}
```

[Figure.35] Random Forest Model 1 descriptions create by RapidMinder Studio using Gofaas dataset.

Boosting < AdaBoost < Gradient Booting < XG-Boost

According to Freund and Shapire (1996) *Boosting* refers to an general and provably effective method of producing an very accurate prediction rule by combining rough and moderately inaccurate rules of thumb. Boosting has its roots in an theoretical framework for studying machine learning called the "PAC"(Probably Approximately Correct) learning model by Valiant. Kearns and Valiant were the first to pose the question of whether an "weak" learning algorithm which performs just slightly better than random guessing in the PAC model can be "boosted" into an arbitrarily accurate "strong" learning algorithm.

Methodology

General Boosting works with a variety of different loss functions. Models include regression, resistant regression, K-class classification and risk modelling. (Hastie, 2014).

Boosting also works by averaging trees but does it in an way which learns from errors of previous trees. Boosting solve the performance problem by supplying directly the weak learning algorithms.

"*When this is possible, the booster's distribution D_t is supplied directly to the weak learning algorithm, a method call boosting by* **reweighting**." (Freund, Schapire, 1996)".

Boosting stumps is a two-nodes tree, after an single split. It works remarkably well on the nested-spheres problem (Hastie, 2014).

The analysing training error is the most basic theoretical property of AdaBoost concerns its ability to reduce the training error.

Gradient Boosting builds additive tree models, for example, for representing the logits in logistics regression. It inherits all the good features of trees (variable selection, missing data, mixed predictors), and improves on the weak features, such as prediction performance. (Hastie, 2014).

XG-Boost stands for *eXtreme Gradient Boosting* it is an large-scale machine learning to build scalable learning systems. It is fast and optimizes for out-of-core computations. (Chen & Guestrin, 2016). It solves the variable overlifting problem or random error, which is when the model or algorithm shows low bias but high variance.

Result interpretations

According to Hastie, (2014) Boosting is an *Stagewise Additive Modelling*. It builds an additive model by

Equation 8 $\qquad F(x) = \sum_{m=1}^{M} \beta_m \, b(x; ⅂_m)$

where $b(x; ⅂m)$ is a tree, and $⅂m$ parametrizes the splits.

They are based on statistics, traditionally the parameters are fit jointly (i.e. least squares, maximum likelihood).

Whereas XG-Boost regularises the weight which overfit the data in other words regularise the minor improvements in the regularized objective

$$I_1 = \{1\}$$
$$G_1 = g_1$$
$$H_1 = h_1$$

$$I_2 = \{4\}$$
$$G_2 = g_4$$
$$H_4 = h_4$$

$$I_3 = \{2,3,5\}$$
$$G_3 = g_2 + g_3 + g_5$$
$$H_3 = h_2 + h_3 + h_5$$

$$Obj = -\sum_j \frac{G_j^2}{H_j+\lambda} + 3\gamma$$

The smaller the score is, the better the structure is
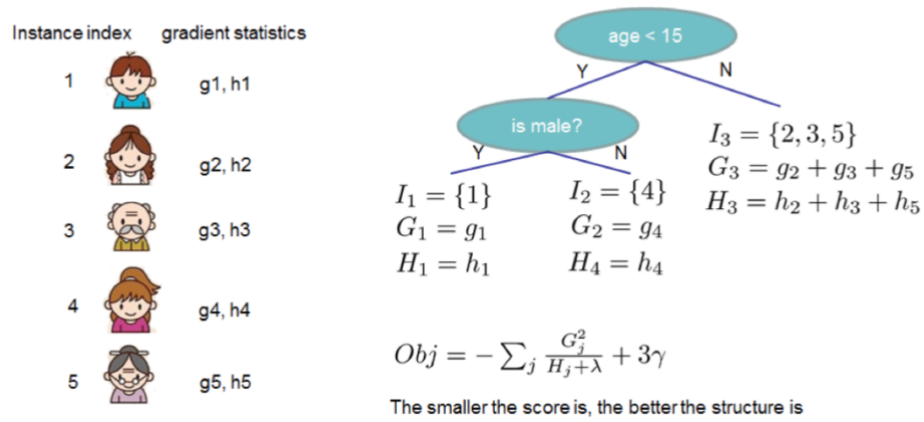
[Figure.36] XG-Boost Structure Score Calculation. (Chen and Guestrin, 2016).

To prevent overfitting which is a problem of all Boosting algorithms, two techniques were added, first one is shrinkage and the second is column (feature) subsampling.
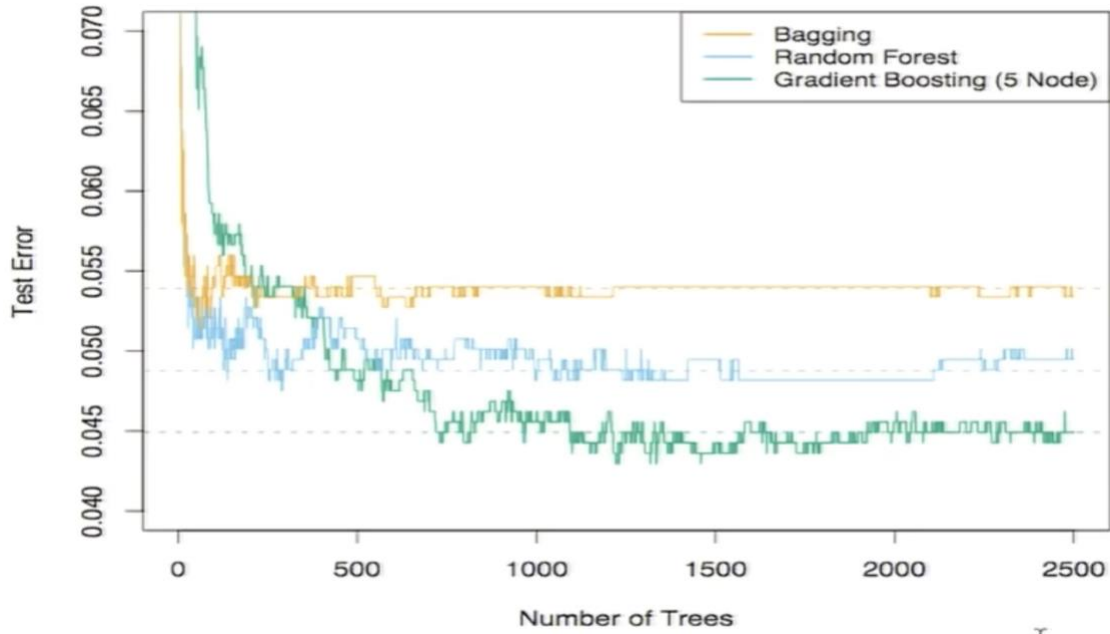
[Figure.37] States the dominance between algorithms: Gradient Booting > Random Forest > Bagging. The number of trees at x axis and test error at y axis.

Boosting Relation to Support Vector Machines

According to Freund, Schapire (1999) Support Vector Machine (SVM) and AdaBoost seem very similar when analysing the margin theory of points, or the distance from the data point to an decision boundary. Once the weak hypotheses which will be combined are found the only interest is in choosing the coefficients α_t . One reasonable approach suggested by AdaBoost's generalization error is to choose the coefficients so that the bound is minimized. The vector of weak-hypothesis prediction associated with the instance vector and weight vector.

Where, for boosting, the norms in the denominator are defined as,

In comparison, the explicit goal of SVM is to maximize a minimal margin of the hyperplanes, which will be explained in details at SVM Algorithm Section.

Several Important differences between Boosting and SVM

1. **Different norms can result in very different margins**: let the weak hypotheses have range {-1,+1} and the label *y* on all examples can be computed by an majority vote of *k* of weak hypotheses. *K* is a small fraction of the total number of weak hypotheses then the margin associated with AdaBoost will be much larger than one associated with SVM ;
2. **The computation requirement are different**: SVM corresponds to a *quadratic programming*, while AdaBoost corresponds to *linear programming*;

3. **A different approach is used to search efficiently in high dimensional space**: quadratic programming is more computational and demanding than linear programming. Also SVM deal with overfitting problem through the method of *kernels* which allow algorithms to perform low dimensional calculation that are mathematically equivalent to inner products in a high dimension "virtual" space. Whereas AdaBoost address overfitting by maximizing the margin, the computational problem with operating in high dimensional spaces remains.

Support Vector Machines Algorithm

Contributors (Cortes and Vapnik, 1995)

1. **Fisher, 1936**: suggest the first algorithm for pattern recognition.

2. **Rosenblatt, 1962**: explored a different kind of learning machines: perceptrons or neural networks.

3. **Rumelhart, Hinton & Williams, 1986, 1987; Parker, 1985, LeCun, 1985**: an algorithm that allows for all weights of the neural network to adapt in order locally to minimize the error on set of vectors belonging to a pattern recognition problem was found.

4. **Cortex and Vapnik, 1995**: construct a new type of machine learning, SVM.

Methodology

The support-vector machine also support-vector networks are supervised learning models with associated learning algorithms that analyse data used for classification problem. Given an set of training examples, each marked as belonging to one or the other of two categories. For instance fake and authentic news, an SVM training algorithms builds a model that assigns new examples to one category or the other, making it an non-probabilistic binary linear classifier. However when data is unlabelled, supervised learning is not possible. (Wikipedia, 2019)

According to Cortes and Vapnik, (1995) high generalization is the ability of support-vector networks utilizes polynomial input transformations. In other words to minimize the error on a set of vectors belonging to a pattern an optimal hyperplanes is bounded by the ratio between the expectation value of the number of support vectors and the number of training vectors:
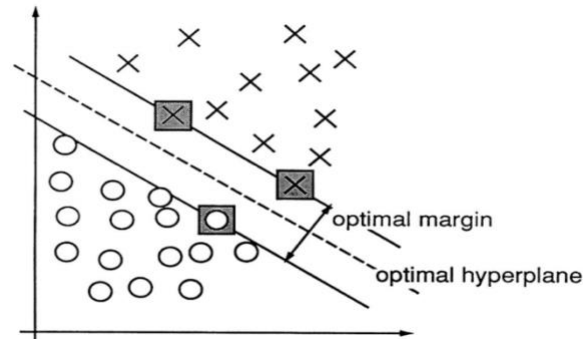
[Figure.38] An example of an separable problem in an two dimensional space. The SV, marked with grey squares, define the margin of largest separation between the two classes.(Cortes and Vapnik, 1995).

The reason support-vector machine utilizes polynomial input transformation is because the entire space is separates into half spaces. As a result the best choice will be the hyperplane that leaves the maximum margin from both classes.

Chapter 5. Gofaas R package

Chapter 6. Gofaas Web App

UI (User Interface):

In information technology, The UI (User Interface) is everything designed into an information device with which any person may interact. This can include display screen, keyboard, a mouse and the appearance on desktop.

For UI (User Interface) we used R Shiny (Library).

R Shiny:

R Shiny (is an R Package that help to build interactive web apps straight from R. it not only that it also can be deployed as a standalone app and useful to build dashboards.) is more suitable for our project. It not only using same Programming Language but also easy to embed, design and expend as needed.

The UI (User Interface) is designed on R Shiny. It will allow the user to interact with Application. It will provide user few option to select from like type, Copy and Paste news or upload a text file, give user ability to select algorithm of their choice or select more than one algorithm etc.

UI Design:

User interface (UI) design is the process of making interfaces in software or computerized devices with a focus on looks or style. Designers aim to create designs users will find easy to use and pleasurable. UI design typically refers to graphical user interfaces but also includes others, such as voice-controlled ones.

These are a few points we kept in mind while designing the UI (User Interface).

1. Attractive to everyone.
2. Easy and simple to use.
3. Easy to understand.
4. Easy to interact with.

UI explained:

The user will be provided few options to select form.

1. Text Input: User can type or Past the News.

2. File input: User can upload a text file which contain news (User have to select only one option from 1 or 2).

3. Checkbox allow user to select from different algorithm. One or more algorithm can be used for fake news detection. (User can use one or more algorithm at the same time).

4. Submit Button.

5. Results after submit news.
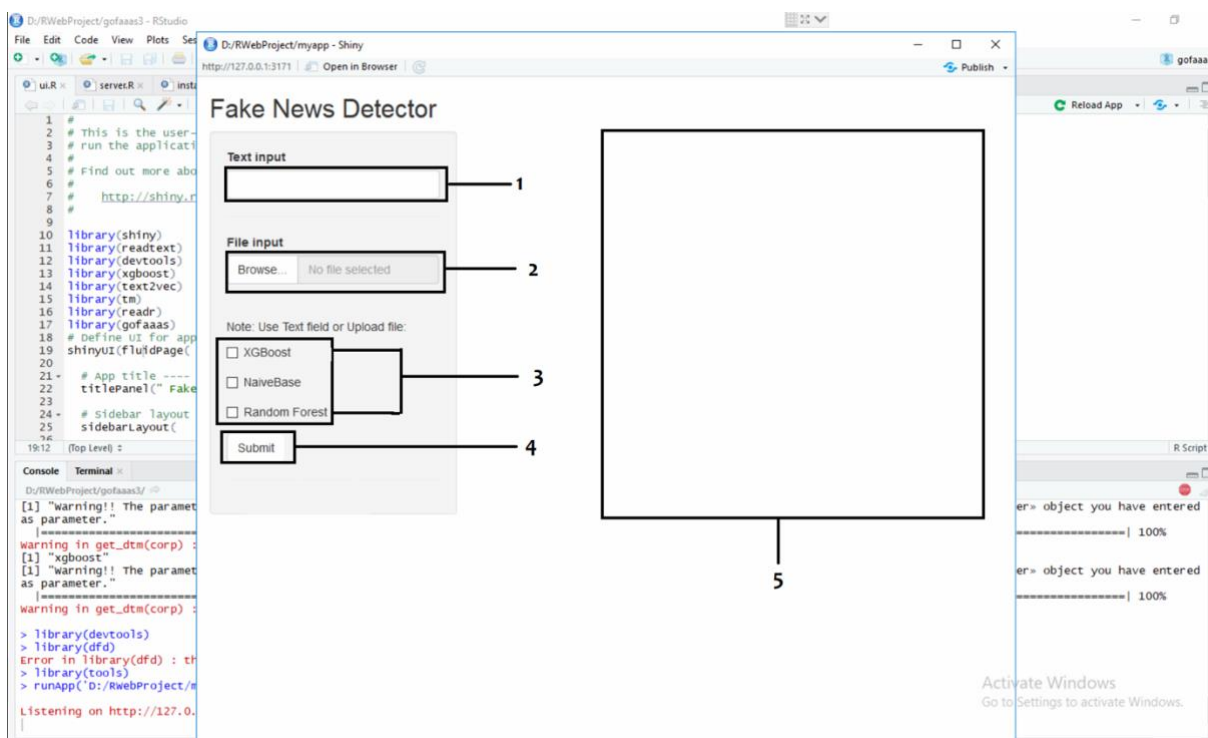
[Figure.39] UI Main Page

If user type a piece of text or upload file and select none of the algorithm will get an error to inform about options as shows. Also if user select one or more algorithm but neither upload file or enter text will get an error. It will ask user to enter text or upload file as shows **Error! Reference source not f ound.**.
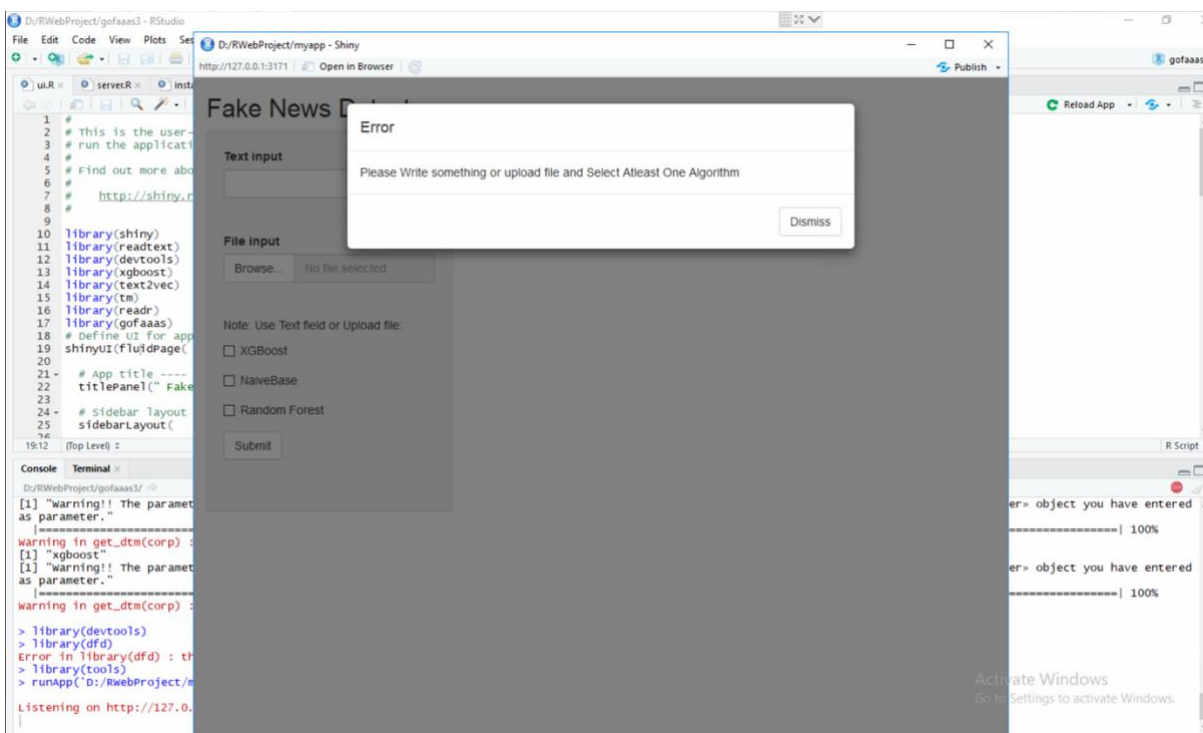
[Figure.40] Error

Finally results. Once we either type or upload news as text file select at least one algorithm will go to GitHub and run the algorithm and get the results and display as popup.
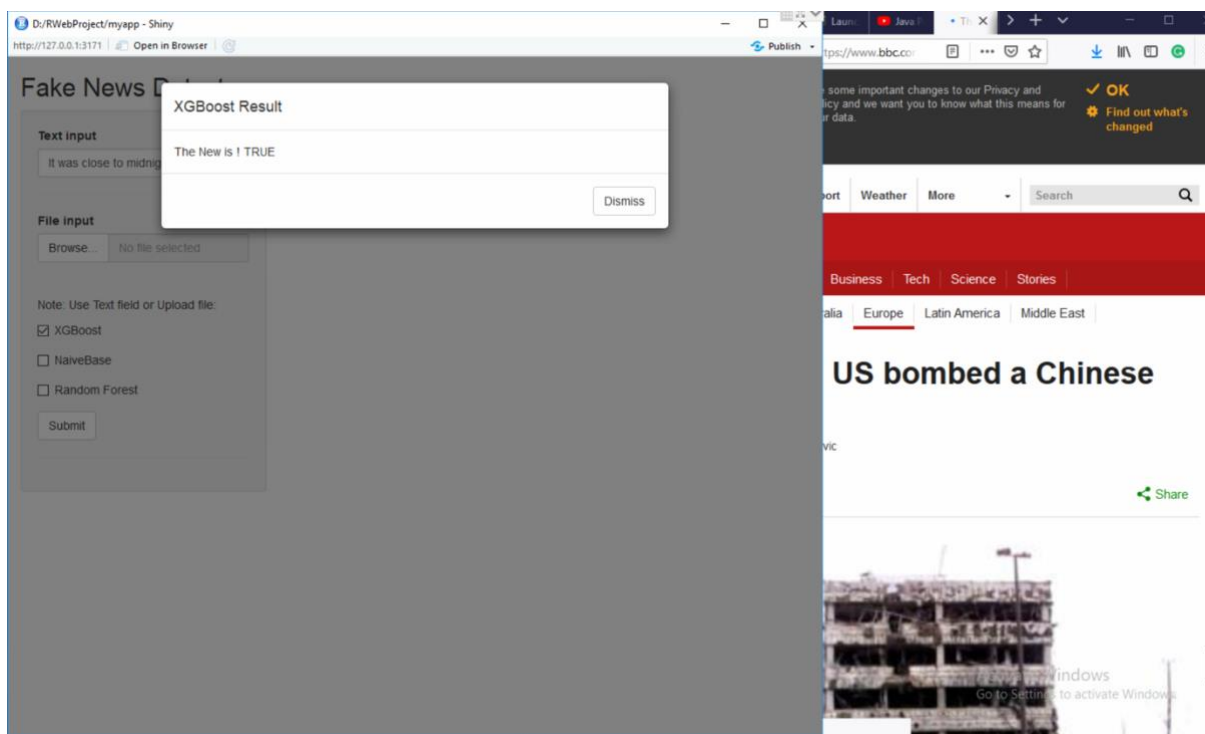
[Figure.41] Result

Chapter 7. Conclusion

Initial Conclusion

The Fake News concept has been around for ages and has been user for financial and political gain. Therefore it impacts extremely negative on individuals and society. Shu et al (2016) show that social media has been used to provide low quality news because it is cheap to provide and much faster and easier to disseminate. For this reason Fake News detection on social media is challenging and relevant. Machine learning promises to help us as an human to scale up the fake news.

Final Conclusion

The present project was a very slow and demanding process of discovery, processing and analyses. We had to change paths to be able to understand the limitations and the challenges of machine learning. At the very start for any beginner there are two main challenges to be faced. One of them is the subject definition itself, in our case fake news. The second challenge is how to detect the subject is study using machine learning techniques.

There are different approaches to use machine learning one of them is an supervised learning which required a labelled dataset. We did not know that until we tried on. We also did not know that if the data is not pre-processed it can impact in all results. Which makes sense when you think that data mining is extracting some important features from an bunch of mixed information. If the source is not clean enough you may get things you do not want. So a very clean, labelled dataset is the key to success in supervised machine learning. However is important to know how to label the dataset and create your own which you can trust. We still have a lot to learn and to be to honest is very hard to be brave and change paths, specially when the time is limited. Whereas as Erol Ozan said : "Some beautiful paths can't be discovered without getting lost".

Resource Requirements

Human Resources

Following are the roles & responsibilities of our team members.

Initial - Team & Roles

Initial commitments			
tasks	Student ID	Names	Roles
1, 3	2016288	ANDREA LOPEZ	Web Developer
2, 4	2017279	ADELO VIEIRA	Data Analyst
5, 6	2016245	ZAFAR AHSAN	Web Designer
8, 9	2015407	FAROOQ SAQIB	Database Administrator
7, 10	2016439	SHIRLEY MARINHO	Project Manager
MUHAMMAD IQBAL supervisor			

1. Problem Area/ Innovation Area,
2. Solution to the problem / Innovation Solution
3. Project Goals
4. Project Objectives
5. Resource Requirements
6. Project Scope
7. Summary Schedule
8. Risk Analyses
9. Conclusion
10.   Reference list

Software & Techniques requirements

For the completing of the project we will be relying on the software And techniques that will help us to complete different tasks.

Initial Intangible resources

Adobe Dreamweaver: HTML, CSS, JScript, MySQL etc.

Data Mining: for the purpose of turn raw data into useful information. By using software to look for patterns in large batches of data. That will help us to difference between authentic and fake news. According to Encyclopedia Britannica (2018). Data mining, also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. Data mining or DM, also known as knowledge discovery in databases (Fayyad and Uthurusamy, 1999), and information archaeology (Brachman et al, 1993).

Text Mining: According Marti (1999) Text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners. I suspect this has happened because people assume TDM is a natural extension of the slightly less nascent field of data mining (DM).

MYSQL: After mining Data we will require a database to save the results and that will help to train our software model to differentiate between authentic and fake news.

Bot: Automated software to complete simple & repetitive tasks that would be time consuming and impossible for an individual to complete within reasonable time frame. It would be time-consuming, mundane or impossible for a human to perform. (Technopedia, 2018).

RapidMiner: For purpose of Machine learning Model we will be using rapid miner software.

R Language: According GNU Project, 2018, R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

Machine Learning Model: we will design a Machine learning model by using R Language. This model will learn from different raw data from different resources and help it to learn. According to *Professor Tom Mitchell, Carnegie Mellon University:*

> "A computer program is said to learn from experience 'E', with respect to some class of tasks 'T' and performance measure 'P' if its performance at tasks in 'T' as measured by 'P' improves with experience 'E'.

Vast amounts of data are being generated in many fields, and the statisticians' job is to make sense of it all: to extract important patterns and trends, and to understand "what the data says". We call this *learning from data. (Hastie, 2016).*

Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field…(Bishop, 2016).

Machine learning (ML) can help you use historical data to make better business decisions. ML algorithms discover patterns in data, and construct mathematical models using these discoveries. Then you can use the models to make predictions on future data. (Amazon, 2018).

One of the most interesting features of machine learning is that it lies on the boundary of several different academic disciplines, principally computer science, statistics, mathematics, and engineering. ...machine learning is usually studied as part of artificial intelligence, which puts it firmly into computer science ...understanding why these algorithms work requires a certain amount of statistical and mathematical sophistication that is often missing from computer science undergraduates. (Marsland, 2009).

1. **Communications**
- Cell phones
- Email
- Internet
- Basecamp

2. **Physical resources**
- Equipment: Computers
- Data: Twitter data

Current Roles			
tasks	**Student ID**	**Names**	**Roles**
3, 6, 12, 14	2016288	ANDREA LOPEZ	Data Analyst
2, 4, 7, 12, 15	2017279	ADELO VIEIRA	Data Analyst
8, 12,14	2016245	ZAFAR AHSAN	Web Designer
8, 12,16	2015407	FAROOQ SAQIB	Web Developer
1, 2, 5, 9, 10, 12	2016439	SHIRLEY MARINHO	Project Manager
MUHAMMAD IQBAL supervisor			

1. Abstract
2. Initial and Final Proposal Summary
3. Chapter 3. Twitter – Sentimental Analysis
4. Chapter 4. Training Model(results)
5. Chapter 4. Training Model(Kaggle 1 and 2 datasets, SVM, Random Forest, XG-Boosting algorithm)
6. Chapter 4. Training Model – (Victory University, Gofaas dataset, Naïve Bayes algorithm)
7. Chapter 5. Gofaas R package
8. Chapter 6. Gofaas WebApp
9. Chapter 7. Final Conclusion
10.   Resources Requirement- Team Roles
11.   Resources Requirement - Technologies
12.   Summary Schedule
13.   Reference list
14.   Manual work of 411 entries of Gofaas dataset
15.   Manual work of 100 entries of Gofaas dataset
16.   Gofaas R package
17.   Gofaas WebApp

Current Intangible resources

The resources of the project has been changed dramatically over last six months because during the development phase we realised the some of the resources are not feasible or there are better option that will be more useful over the others.

Adobe Dreamweaver to R Shiny: R Shiny (is an R Package that help to build interactive web apps straight from R. It not only that it also can be deployed as a standalone app and useful to build dashboards.) is more suitable for our project. It not only using same Programming Language but also easy to embed, design and expend as needed.

The UI (User Interface) is designed on R Shiny: it will allow the user to interact with Application. It will provide user few option to select from like type, Copy and Paste news or upload a text file, give user ability to select algorithm of their choice or select more than one algorithm etc.

Data Mining or Text Mining: due to time and the resources restriction it was not feasible to clean and label the mined data so we had to scrape the whole idea to use the mining tools to collect the data for our model training. So we decided to go different authentic new websites and some fake websites to collect and prepare out data set for model training purpose.

R Language: since day first we were using R language to design our machine learning models and now UI (User interface) also based on R Shiny (a library of R language).

Summary Schedule

Milestone table First Semester

Milestone	Date	Time	Due Date	Who's Responsible	Details
The idea meeting	01/10/2018	2.30pm	Completed	group	Defining the idea – Fake news and Machine Learning
Fake News researching	09/10/2018	10.44pm	Complete	Andrea	Definition, origin, documentary
Twitter policies	09/10/2018	10.44pm	Postponed	Ahsan	Tweeter developer website
Database	09/10/2018	10.44pm	Postponed	Farooq	
Algorithms	09/10/2018	10.44pm	Still in process	Adelo, Shirley, Farooq	Decision Tree, Naïve and Bayne
Data Collection	09/10/2018	10.44pm	Still in process	Adelo, Shirley	tweepy
Graham review	17/10.2018	10.15am	complete	Andrea, Adelo, Shirley	Graham's review: fake news definition
First - Muhammed meeting	18/10/2018	5.30pm	Complete	group	Referencing books: Data Mining Chapter 3, RapidMiner videos
Presentation meeting	18/10/2018	8.15pm	Complete	group	Idea presentation
Group Logo	19/10/2018	10.54am	Complete	Shirley	
Basecamp setup	21/10/2018	7pm	Complete	Shirley	Adding groups member, supervisor and tasks to be completed
Presentation	24/10/2018	10.14am	Complete	group	Graham's review: be careful with biases
Week 1 – Self Assessment	From 28/10/2018	anytime	Until 04/11/2018	individual	Self-Evaluation Report
Group Meeting	30/10/2018	11.30am	Complete	group	GoFaaas the Web Application
Week 2 – Self Assessment	From 05/11/2018	anytime	Until 11/11/2018	individual	Self-Evaluation Report
Second - Muhammed meeting	8/11/2018	5.30pm	Complete	group	Referencing books: Text Mining Chapters 1,2,3, fake news

					articles and GitHub dataset
Week 3 – Self Assessment	From 12/11/2018	anytime	Until 18/11/2018	individual	Self-Evaluation Report
Group Meeting	21/11/2018	11.15am	Complete	group	Proposal Project tasks and self-assessment
Week 4 – Self Assessment	From 19/11/2018	anytime	Until 25/11/2018	individual	Self-Evaluation Report
Group Meeting	26/11/2018	11.30am	Complete	group	Problem Area [Andrea and Farooq] Problem Solution [Adelo and Ahsan] Keep posted to Basecamp Portal
Week 5 – Self Assessment	From 26/10/2018	anytime	Until 02/11/2018	individual	Self-Evaluation Report
Third - Muhammed meeting	29/11/2018	5.30pm	Complete	group	Small fake news program in R, Read the article about fake news provided
Week 6 – Self Assessment	From 03/12/2018	anytime	Until 09/12/2018	individual	Self-Evaluation Report
Week 7 – Self Assessment	From 10/12/2018	anytime	Until 16/12/2018	individual	Self-Evaluation Report
Fourth - Muhammed meeting	13/12/2018	5.30pm	In progress	group	
Week 8 – Self Assessment	From 17/12/2018	anytime	Until 23/12/2018	individual	Self-Evaluation Report
Group Meeting	10/12/2018	11.30am	In progress	group	The final discussion to deliverables of the project proposal
Week 9 – Self Assessment	From 24/12/2018	anytime	Until 31/12/2018	individual	Self-Evaluation Report
Due date to Project Proposal	11/12/2018	11.27pm	13/12/2018	group	Delivery of Project Proposal

Milestone table Second Semester

Milestone	Date	Time	Due Date	Who's Responsible	Details
Self Assessment	From 01/01/2019	anytime	Until 06/01/2019	individual	Self-Evaluation Report
Self Assessment	From 07/01/2019	anytime	Until 13/01/2019	individual	Self-Evaluation Report
Self Assessment	From 14/01/2019	anytime	Until 20/01/2019	individual	Self-Evaluation Report
Self Assessment	From 21/01/2019	anytime	Until 27/01/2019	individual	Self-Evaluation Report
Self Assessment	From 28/01/2019	anytime	Until 03/02/2019	individual	Self-Evaluation Report
Self Assessment	From 04/02/2019	anytime	Until 10/02/2019	individual	Self-Evaluation Report
First Group Meeting	05/02/2019	11.30am	Complete	group	Andrea, Farooq, Shirley. Planning to meet Muhammad and delivery the small R program
Self Assessment	From 11/02/2019	anytime	Until 17/02/2019	individual	Self-Evaluation Report
First Muhammed meeting	18/02/2019	11.30am	Complete	group	Small R program for Twitter sentimental analysis presentation. Work for next meeting: Algorithms(Decision Tree-Ahsan, Naïve Bayes - Andrea, Random Forest - Shirley, XG-Boosting Farooq and Adelo)
Self Assessment	From 18/02/2019	anytime	Until 24/02/2019	individual	Self-Evaluation Report
Self Assessment	From 25/02/2019	anytime	Until 03/03/2019	individual	Self-Evaluation Report

Second Muhammed meeting	04/03/2019	11.30am	Complete	group	Keep same model using same fake news dataset train() and test() from Kaggle
Self Assessment	From 04/03/2019	anytime	Until 10/03/2019	individual	Self-Evaluation Report
Group Meeting	06/03/2019	9.30am	Complete	group	Presentation preparation: Fake News definition used in the project; Sentimental Analysis for Twitter, Algorithms Decision Tree, Random Forest, XG Boosting; The fake news dataset, the data store problem; next milestones
GitHub code version control	In progress			group	https://github.com/ShiMarinho/MachineLearning-FakeNews
Third - Muhammed meeting	11/03/2019	11.30pm	Complete	group	Read the articles: 1- Helmstetter and Paulheim, 2018; 2- Gilda, 2017;
Self Assessment	From 11/03/2019	anytime	Until 17/03/2019	individual	Self-Evaluation Report
Presentation	13/03/2019	10.30am	Complete	group	We had to change the route from extracting data from Twitter to a fake news. The problem of large amount of data to store is the current discussion.
Self Assessment	From 18/03/2019	anytime	Until 24/03/2019	individual	Self-Evaluation Report
Fourth - Muhammed meeting	25/03/2019	11.30am	Complete	group	Algorithms performance analysis to check the accuracy between them. XG-Boosting, Naïve Bayes, Random Forest.
Self Assessment	From 25/03/2019	anytime	Until 31/03/2019	individual	Self-Evaluation Report
Group Meeting	01/04/2019	2.30pm	3.30pm	group	**Final Prototype:** Back-End algorithms: XG-Boosting (Adelo), Naïve Bayes(Andrea), Random

					Forest(Shirley) to create the model . Front-End Web App, input piece of text or file process the data and return, if is fake or not.(Ahsan, Farooq). **Technologies**: R language, R Studio, ShinyRStudio.
Online documentation for Algorithms result table	In progress			Adelo	http://perso.sinfronteras.ws/index.php/Establishing_an_authenticity_of_sport_news_by_Machine_Learning_Models
Self Assessment	From 01/04/2019	anytime	Until 07/04/2019	individual	Self-Evaluation Report
Applied Technology Group Project class	03/04/2019	9am	11.30am	Shirley	Schedule an meeting to check the documentation (Graham) and the technical overview (Mark)
Self Assessment	From 08/04/2019	anytime	Until 14/04/2019	individual	Self-Evaluation Report
Creation of R package for gofaaas library	15/04/2019		In progress	Adelo	Dependencies: install.packages('devtools') install.packages('xgboost') install.packages('text2vec') install.packages('tm') install.packages('readr') install_github('https://github.com/adeloaleman/gofaaas')
Self Assessment	From 15/04/2019	anytime	Until 21/04/2019	individual	Self-Evaluation Report
Graham and Mark meeting	16/04/2019	11.30am	1pm	group	Mark's comments: the dataset label, do a small one by hand to ensure accuracy and control. Graham's comments: all the changes, challenges and fails should be documented
Gofaas dataset made by hand	18/04/2019		In progress	Andrea, Ahsan	https://github.com/ShiMarinho/MachineLearning-FakeNews/blob/master/fakeNewsBackEnd/gofaasFakeNewsDataset/goofaasDataset.csv
The Final Documentatio	19/04/2019	For discussion - Adelo Introduction Chapter 1 - Project proposal Chapter 2 - Training a Supervised Machine Learning Model for fake news detection			

n Structure proposal		2.1 Procedure
		2.2 Results
		   2.2.1 Summary of Results
		   2.2.2 Results for the Kaggle fake news dataset
		   2.2.3 Results for the fake news detector dataset
		   2.2.4 Results for the Gofaaas fake news dataset
		2.3 Datasets used
		   2.3.1 Description of the Kaggle fake news dataset
		   2.3.2 Description of the Fake news Detector dataset
		   2.3.3 Description of the Gofaaas fake news dataset
		2.4 Algorithms
		   2.4.1 Support vector machine
		   2.4.1 Random forest
		   2.4.1 Extreme Gradient Boosting
		    2.4.1.1 The XGBoost R package
		   2.4.1 Naive Bayes
		Chapter 3 - Gofaaas Fake News detector Web App
		Conclusion

The Final Documentation Structure proposal	19/04/2019	For discussion - Shirley
		
		Front Page
		Declaration
		Acknowledgement (Thanks for Muhammad, Graham and Mark)
		Abstract
		Table of Contents
		List of Figures
		List of Tables
		Chapter 1 Introduction
		1.1 Motivation
		1.2 Team & Roles
		     1.2.a initially
		     1.2.b the changes
		1.3 Project Proposal
		     1.2.a initially
		     1.2.b the changes
		1.4 Problem area/ Innovation area
		     1.2.a initially
		     1.2.b the changes
		1.5 Project Goal
		     1.2.a initially
		     1.2.b the changes
		1.6 Project Objectives
		     1.2.a initially
		     1.2.b the changes
		1.7 Resources Requirement
		     1.2.a initially
		     1.2.b the changes
		1.8 Project Scope
		     1.2.a initially
		     1.2.b the changes
		1.9 Summary Schedule
		1.10 Risk Analysis
		     1.2.a initially
		     1.2.b the changes
		Chapter 2 Literature review
		2.1 Fake News Historic Foundation
		Chapter 3 - The Machine Learning
		Phase - Discovery
		3.1 Data Extraction and Cleaning techniques
		3.1.a Twitter API for data using #arsenal as a searching key word
		3.1.b Text Mining techniques
		3.1.c The Linguistic approach
		3.1.d The Contextual approach
		3.2 Data Processing
		3.2.a Confusion matrix to separate positive and negative tweets
		3.3.b Sentimental analysis application
		3.3 Data Visualisation
		3.3.a Positive and Negative Tweets
		3.3.b Word Cloud
		Phase - Transition
		3.4.a Train and test to achieve 70 to 80% accuracy
		3.4.b Crossfold validation
		3.4.c Tweeter been substituted to fake and authentic news extraction
		Chapter 4 - The outcomes of the Model
		Phase - The algorithms
		3.5.a - Naive Bayes
		3.5.b - XGBoost
		3.5.c - Tree
		3.5.d Random Forest
		Phase - The dataset

		3.5 Sample dataset as a solution for running the algorithms 3.5.a Fake news detector (5000 rows) 3.5.b Kaggle fake news (20000 rows) 3.5.c Fake news challenge **Phase - The accuracy problem** 3.6 Labelling process 3.6.a The algorithms table comparator results 3.6.b Gofaas dataset as a label solution 3.6.c Gofaas library Chapter 5 - Gofaas Web App a way to interact, test and display the model results 4. The Web App layout 4.1 - Input data 4.1.a - text 4.1.b - file 4.2 - Processing 4.2.a - text 4.3.b - file 4.3 - Displaying results 4.3.a algorithms accuracy 4.3.b histogram of accuracy's result 4.4 - Public Target Chapter 6 - Conclusion Referencing			
Fifth - Muhammed meeting	29/04/2019	11.30am	Complete	group	Algorithms reports. XG-Boosting, Naïve Bayes, Random Forest (done). WebApp Simulation
Group meeting to define the final documentation structure	29/04/2019	Complete - group Chapter 1. Introduction Initial Proposal Final Proposal Chapter 2. Literature review Chapter 3. Twitter – Sentimental Analysis Chapter 4. Training model Chapter 5. Gofaas R package Chapter 6. Gofaas Web App Chapter 7. Conclusion Resource Requirements Summary Schedule Referencing			
Group meeting to define the final presentation	29/04/2019	Complete - group Introduction – Shirley Fake news quiz – Andrea The classification approach for fake news detection - Andrea Procedures and results - Adelo Algorithms brief - Shirley Gofaas R package - Farooq WebApp Simulation - Ahsan			
Self Assessment	From 29/04/2019	anytime	Until 05/05/2019	individual	Self-Evaluation Report
last - Muhammed meeting	07/05/2019	11.30am	Complete	Adelo, Farooq, Shirley	Documentation and WebApp Simulation review last comments
last -Group meeting	07/05/2019	12.30am	Complete	Adelo, Farooq, Shirley	Making the Video

Referencing

Andrew, O. 2018.
> *The History and Evolution of the Smartphone: 1992-2018* Available at:
> https://www.textrequest.com/blog/history-evolution-smartphone/
> [Accessed 3rd December 2018].

An Quick Guide to Fake News Detection on Social Media. 2018.
> Fake News Detection on Social Media. [ONLINE] Available at:
> https://www.kdnuggets.com/2017/10/guide-fake-news-detection-
> social-media.html.
> [Accessed 1 November 2018].

Architecture of the Web Inspector. 2018.
> *Unweaving the web*. [ONLINE] Available
> at: https://blogs.igalia.com/dpino/2015/11/22/Architecture-of-the-
> Web-Inspector/
>   [Accessed 3 November 2018].

Best Machine Learning Software . 2018.
> *Reviews of the Most Popular Systems.* [ONLINE] Available at:
> https://www.capterra.com/machine-learning-software/
> [Accessed 27 November 2018].

Bishop, C. M., 2006.
> *Pattern Recognition and Machine* Learning. Springer-Verlag New York,
> LLC

Brachman et al.1993.
> R. J. Brachman, P. G. Selfridge, L. G. Terveen, B. Altman, A Borgida, F.
> Halper, T. Kirk, A. Lazar, D. L. McGuinness, and L. A. Resnick.
> 1993. Integrated support for data archaeology. *International Journal of
> Intelligent and Cooperative Information Systems*, 2(2):159-185.

Breiman, L., 2002.
> *Manual on setting up , using, and understanding random forest v3.1.*
> [ONLINE] Available at http://oz.berkeley.edu/users/breiman/
> Using_random_forests_V3.1.pdf. 18,19 [Accessed 3 May 2019].

Cambridge English Dictionary, 2018.
THE INTERNET | meaning in the Cambridge English Dictionary. [ONLINE] Available at: https://dictionary.cambridge.org/dictionary/english/internet. [Accessed 30 November 2018].

Conroy, N. J. , Rubin V. L., Chen Y. 2015.
Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology.

Chen, T., Guestrin, C., 2016.
XG-Boost: A scalable Tree Boosting System[ONLINE] Available at: http://delivery.acm.org/10.1145/2940000/2939785/p785-chen.pdf?ip=109.78.66.19&id=2939785&acc=CHORUS&key=4D47 02B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35 %2E6D218144511F3437&__acm__=1557086126_29c7bfabf71a6 08b5957e450c08ba827 [Accessed 3 May 2019].

Chen Y., Conroy N. J. , Rubin V. L, 2015.
Misleading online content: Recognizing clickbait as false news. In Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection.

Dataquest Data Science Blog. 2018.
Using Linear Regression for Predictive Modeling in R. [ONLINE] Available at: https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/
 [Accessed 3 November 2018].

Darnton, R. 2017.
The True History of Fake News. Available at: https://www.nybooks.com/daily/2017/02/13/the-true-history-of-fake-news/ [Accessed 3rd December 2018].

Digital Single Market. 2018.
Fake News and online disinformation. [ONLINE] Available at: https://ec.europa.eu/digital-single-market/en/fake-news-disinformation[Accessed 13 November 2018].

Encyclopedia Britannica. 2018.
Neural network | computing | Britannica.com. [ONLINE] Available at: https://www.britannica.com/technology/neural-network. [Accessed 29 November 2018].

Entrepreneurship, Business Incubation, Business Model % Strategy Blog,. 2018.
Review of 20 Business Incubation Models – InfoDev Process Model_2009 (Part 15 of 20) | Entrepreneurship, Business Incubation, Business Models & Strategy Blog. [ONLINE] Available at: https://worldbusinessincubation.wordpress.com/2013/10/09/review-of-20-business-incubation-models-infodev-process-model_2009-part-15-of-20/ [Accessed 3 November 2018].

Fayyad and Uthurusamy, 1999.
Usama Fayyad and Ramasamy Uthurusamy. Data mining and knowledge discovery in databases: Introduction to the special issue. *Communications of the ACM*, 39(11), November.

Freund, Y. Schapire R. E., 1996.
Experiments with a News Boosting. Machine Learning: Proceeding of the Thirteenth International Conference, AT&T Laboratories, NJ 07974-0636. pp. 1-9.

Gilda S., 2017.
Evaluating Machine Learning Algorithms for Fake News Detection. 15th Student Conference on Research and Development (SCOReD), vol. 978-1-5386-2126-4, pp. 110-115.

GitHub. (2018a).
FakeNewsNet/Data at master-KaiDMML/FakeNewsNet - GitHub [ONLINE] Available at: https://github.com/KaiDMML/FakeNewsNet/tree/master/Data. [Accessed 10 November 2018].

GitHub. (2018b).
ShiMarinho/MachineLearning-FakeNews: Final-Project. [ONLINE] Available at https://github.com/ShiMarinho/MachineLearning-FakeNews [Accessed 5 October 2018].

Glen. (2014).
>   *Statistics How To.* [ONLINE] Available at: https://
>   statisticshowto.datasciencecentral.com/bayes-theorem-problems/
>   [Accessed 3 May 2019].

Granskogen T., 2018.
>   Automatic Detection of Fake News in Social Media using Contextual
>   Information. Norwegian University of Science and Technology.

Hastie, T., 2016.
>   The Elements of Statistical Learning: Data Mining, Inference, and
>   Prediction, Second Edition (Springer Series in Statistics).Springer.

Helmstetter S., Paulheim H., 2018.
>   *Weakly Supervised Learning for Fake News Detection on Twitter.*
>   International Conference on Advances in Social Networks Analysis and
>   Mining (ASONAM), August 28-31, pp. 274-277.

Hendricks, D. 2013.
>   *Complete History of Social Media: Then and Now.* Available at:
>   https://smallbiztrends.com/2013/05/the-complete-history-of-social-
>   media-infographic.html [Accessed 3rd December 2018].

Kai H., Lars A., Arvidsson A., Arup N., Finn, C. E., Etter, M. 2011.
>   *Good Friends, Bad News – Affect and Virality in Twitter.* The 2011
>   International Workshop on Social Computing, Networking, and
>   Services, SocialComNet.

Korting, T., S., 2014.
>   *How SVM(Support Vector Machine) algorithm works.* Available at:
>   https://www.youtube.com/watch?v=1NxnPkZM9bc. [Accessed 5 May
>   2019].

Lexalytics.
>   Sentiment Analysis Explained [ONLINE] Available at:
>   https://www.lexalytics.com/technology/sentiment-analysis [Accessed
>   05 May 2019].

Liaw, A., Wiener, M., 2002.
Classification and Regression by RandomForest. RapidMiner, Tree (CART) – Machine Learning Fun and Easy [ONLINE] Available at: https://www.youtube.com/watch?v=DCZ3tsQIoGU . [Accessed 27 April 2019].

Life Science. 2018.
Who Invented the Printing Press? [ONLINE] Available at: https://www.livescience.com/43639-who-invented-the-printing-press.htm[Accessed 6 December 2018].

Louppe, G. 2014.
I understanding random forests from theory to practice. University of Liege. Faculty of Applied Sciences. Department of Electrical Engineering & Computer Science PhD dissertation. Advisor: Professor: Pierre Geurts.

Marti A. H., 1999.
Untangling Text Data Mining.[ONLINE] Available at: http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html.
[Accessed 29 November 2018].

Marsland, S., 2009.
Machine Learning: An Algorithmic Perspective. Taylor & Francis, Inc.

Machine Learning – Computer Science CMUQ. 2018.
Machine Learning - Computer ScienceCMUQ.[ONLINE] Available at: http://www.contrib.andrew.cmu.edu/~hmoazam/machine-learning.html.
[Accessed 1 November 2018].

Machine Learning Concepts – Amazon Machine Learning. 2018.
Machine Learning Concepts - Amazon Machine Learning.[ONLINE] Available at: https://docs.aws.amazon.com/machine-learning/latest/dg/machine-learning-concepts.html.
[Accessed 20 November 2018].

Machine Learning Finds "Fake News" with 88% Accuracy. 2018.
Fake News Detection on Social Media. [ONLINE] Available at: https://www.kdnuggets.com/2017/10/guide-fake-news-detection-social-media.html. [Accessed 1 November 2018].

Marco Bonzanini . 2018.
Mastering Social Media Mining with Python. [ONLINE] Available at:
https://developer.twitter.com/en/developer-terms/agreement-and-
policy[Accessed 17 November 2018].

Murphy J., Roser M. (2018).
"*Internet*".   Available   at: 'https://ourworldindata.org/internet'
[Accessed 3rd December 2018].

MLDB. 2018.
MLDB: the open-source Machine Learning Database. [ONLINE]
Available at:
https://mldb.ai[Accessed 27 October 2018].

Machine Box, Inc . 2018.
Docs · Machine Box · Machine learning in a box. [ONLINE] Available
at:https://machinebox.io/docs [Accessed 27 November 2018].

MindTools.com. 2018.
How to Spot Real and Fake News - From MindTools.com. Available at:
https://www.bbc.com/news/uk-36528256. [Accessed 3rd December
2018].

Mishra, D (2018).
Text Processing and Sentiment Analysis of Twitter Data [ONLNE]
Available at: https://hackernoon.com/text-processing-and-sentiment-
analysis-of-twitter-data-22ff5e51e14c [Accessed 05 May 2019].

MIT News. 2018.
Detecting fake news at its source | MIT News. [ONLINE] Available at:
http://news.mit.edu/2018/mit-csail-machine-learning-system-detects-fake-
news-from-source-1004[Accessed 10 December 2018].

MonkeyLearn. 2019.
Text Classification. An Comprehensive Guide to Classifying Text with
Machine   Learning   [ONLINE]   Available   at:
https://monkeylearn.com/text-classification/ [Accessed 5 May 2019].

PMTIPS.Net. 2018.
PMTIPS 2018 source available at
https://pmtips.net/blog-new/3-types-of-essential-resources-for-your-project[Accessed 7 December 2018].

Python Tutorial. 2018.
Extract links from webpage(BeautifulSoup). [ONLINE] Available
at: https://pythonspot.com/extract-links-from-webpage-beautifulsoup/
[Accessed 29 October 2018].

R: What is R? 2018.
R: What is R? [ONLINE] Available at: https://www.r-project.org/about.html [Accessed 3 November 2018].

Root Report. 2018.
6 FREE Fake News Generator To Prank Your Friends. [ONLINE]
Available at:
https://www.rootreport.com/fake-news-generator/
[Accessed 27 October 2018].

Sagar, C., (2018).
Twitter Sentiment Analysis Using R [ONLINE] Available at:
http://dataaspirant.com/2018/03/22/twitter-sentiment-analysis-using-r/ [Accessed 05 May 2019].

Saxena, R., 2017
How the naïve bayes classifier works in machine learning [ONLINE]
Available at: http://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/[Accessed 27 April 2019].

SERWoo. 2018.
35 Sites Which Increase Your Domain's Trust. [ONLINE] Available
at: https://www.serpwoo.com/blog/experts/increasing-your-domains-trust/. [Accessed 25 October 2018].

Slige, S., T. (2017a).
Text Mining with R: An Tidy Approach. Chapter 1 *The Tidy Text Format*
1-12. O'Reilly Media.

Slige, S., T. (2017b).
Text Mining with R: An Tidy Approach. Chapter 2 *Sentiment Analysis with Tidy* Data 13-29. O'Reilly Media.

Slige, S., T. (2017c).
Text Mining with R: An Tidy Approach. Chapter 3 *Analyzing Word and Document Frequency: tf-idf* 31-44. O'Reilly Media.

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. 2016.
Fake News Detection on Social Media: A Data Mining Perspective. [ONLINE] Available N00014-16-1-2257, 1-15.

Shu K., Liu H. 2018.
A Quick Guide to Fake News Detection on Social Media. [ONLINE] Available at: https://www.kdnuggets.com/2017/10/guide-fake-news-detection-social-media.html. [Accessed 1 November 2018].

Social Media Sentiment Analysis- Sinfrotereas. 2018.
Social Media Sentiment Analysis - Sinfronteras. [ONLINE] Available at: http://perso.sinfronteras.ws/index.php/Social_Media_Sentiment_Analysis[Accessed 3 November 2018].

Sistilli A., 2015.
Twitter Data Mining: A Guide to Big Data Analytics Using Python. https://www.toptal.com/python/twitter-data-mining-using-python

Soll, J. 2016.
The Long and Brutal History of Fake News. Available at: https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535 [Accessed 3rd December 2018].

Sports Journalism – Wikipedia. 2018.
Sports journalism. Wikipedia.[ONLINE] Available at: https://docs.aws.amazon.com/machine-learning/latest/dg/machine-learning-concepts.html. [Accessed 19 November 2018].

Stecanella, B., 2017.
A practical explanation of an Naïve Bayes classifier.[ONLINE] Available at: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/ [Accessed 27 April 2019].

Sunil, R 2017.
 6 Easy Steps to Learn Naïve Bayes algorithm(with codes in Python and R).[ONLINE] Available at:
 https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [Accessed 27 April 2019].

TensorFlow. 2018.
 TensorFlow. [ONLINE] Available at: https://www.tensorflow.org.
 [Accessed 25 October 2018].

Technopedia, 2018.
 Bot. Available at:
 https://www.techopedia.com/definition/24063/internet-bot
 [Accessed 4 December 2018].

The Fake News Generator. 2018.
 The Fake News Generator. [ONLINE] Available at:
 https://www.thefakenewsgenerator.com [Accessed 17 November 2018].

The New York Times. 2018.
 Opinion | Operation Infektion: A three-part video series on Russian disinformation. [ONLINE] Available
 at: https://www.nytimes.com/2018/11/12/opinion/russia-meddling-disinformation-fake-news-elections.html
 [Accessed 7 November 2018].

The Telegraph. 2018.
 Fake News: What exactly is it – and how can you spot it? . [ONLINE]
 Available at: https://www.telegraph.co.uk/technology/0/fake-news-exactly-has-really-had-influence/
 [Accessed 6 December 2018].

Torgo, T. 2010.
 Data Mining with R: Learning with Case Studies.
 Norwell, Massachusetts, U.S.A.: Chapman & Hall.
 Chapter 3 Predicting Stock Market Returns 95 – 163

Twitter Developers . (2018a).
 Developer Agreement and Policy. [ONLINE] Available at:
 https://developer.twitter.com/en/developer-terms/agreement-and-policy. [Accessed 17 November 2018].

Twitter Developers . (2018b).
 Rate Limiting. [ONLINE] Available at:
 https://developer.twitter.com/en/docs/basics/rate-limiting
 [Accessed 17 November 2018].

Toptal | Engineering Blog . 2018.
 Twitter Data Mining: Analysing Big Data Using Python. [ONLINE]
 Available at:https://www.toptal.com/python/twitter-data-mining-
 using-python
 [Accessed 27 November 2018].

United States Department of State. 1987.
 *Soviet Influence Activities: An Report on Active Measures and
 Propaganda, 1986-87.* Department of State Publication 9627.
 USA.

YouTube. (2018a).
 01 RapidMiner Studio - GUI Intro - YouTube. [ONLINE] Available at:
 https://www.youtube.com/watch?v=ophGqpUexKI&list=PLssWC2d
 9JhOZLbQNZ8OuOxLypglgWqbJA
 [Accessed 19 November 2018].

YouTube. (2018b).
 #5: Front End - Shiny Web Application in R - YouTube. [ONLINE]
 Available
 at: https://www.youtube.com/watch?v=sIrmioxztYQ&index=5&list=
 PLYOWzFZIEZiXGKcvmC-8tZu_drpgPk2Eh.
 [Accessed 20 November 2018].

YouTube. (2018c).
 *How To Extract URLs From A Website In Chrome? (No Downloads
 Required) – YouTube.* [ONLINE] Available
 at: https://www.youtube.com/watch?v=85GqVYeyn18&feature=yo
 utu.be
 [Accessed 19 November 2018].

YouTube. (2018d).
 Predicting the Stock Value using RapidMiner – YouTube [ONLINE]
 Available at:
 https://www.youtube.com/watch?v=LbtZU1_i9Qk&feature=youtu.
 be
 [Accessed 19 November 2018].

YouTube. (2018e).

Karl Marx & Conflict Theory: Crash Course Sociology #6 - YouTube
[ONLINE] Available at:

: https://www.youtube.com/watch?v=gR3igiwaeyc

[Accessed 03 December 2018].

YouTube. (2018f).

POLITICAL THEORY - Karl Marx – YouTube [ONLINE] Available at:
https://www.youtube.com/watch?v=fSQgCy_iIcc [Accessed 03
December 2018].

YouTube. (2014).

Hastie, T., 2014 – *Gradient Boosting Machine Learning – YouTube*
[ONLINE] Available at:
https://www.youtube.com/watch?time_continue=61&v=wPqtzj5VZ
us [Accessed 02 May 2019].

Wakefield, J. 2016.

Social Media Outstrips TV as news source for young people. Available
at: https://www.bbc.com/news/uk-36528256.   [Accessed   3rd
December 2018].

Wikipedia, 2019.

Support-vector Machine. Available at:
https://en.wikipedia.org/wiki/Support-vector_machine. [Accessed 5
May 2019].