TABLE OF CONTENTS

Abstract

This Master's thesis presents the process and results of the analysis of articles on artificial intelligence published from 2014 to 2023 by two digital media. It is explained how the process of extracting the content by means of and the date of each article in the two media has been carried out, how the data has been cleaned, visualizations have been made and text analysis has been carried out. All this was done using the RStudio tool. First of all, web scraping techniques were used to obtain the information. For this purpose, it was necessary to use the RSelenium R package, which allows remote interaction with the websites. For the visualizations the package ggplot2 has been used and, finally, several text analysis tools have been used in R. That is why this work is composed of several documents: this PDF and several files in rmd format in which all the code necessary to replicate the work is found. In these files the commented code is presented in such a way that anyone who needs it can make the modifications they wish.

All the files can be found in the GitHub repository:
https://github.com/isabelml/tfm_r_project.git

Key words: Artificial intelligence, press, newspapers, data scraping, text analysis, data visualization, RStudio, RSelenium, ggplot2

1 INTRODUCTION

In the last decade in Spain, as in the whole of the Western world, there has been a growing interest in the development of Artificial Intelligence (AI). This interest has been preceded by the development of different tools that incorporate this technology. Examples of this are intelligent virtual assistants such as Siri from Apple, Google Assistant, Alexa from Amazon, or Microsoft Cortana. Also, companies such as Tesla have made great strides in the development of autonomous vehicles that incorporate AI to recognize the environment and even make autonomous decisions. Other fields in which this technology has brought about a revolution are the recommendation systems in different entertainment platforms (e.g., Netflix or Spotify); in medicine, where algorithms have been developed for medical diagnosis; and in natural language processing, which has enabled the development of chatbots such as Chat-GPT or Midjourney.

The arrival of artificial intelligence in society has raised several ethical and social challenges associated with its use. More and more people are showing interest in this technology and asking questions about it. The media have shaped this debate and at the same time have been opinion leaders by spreading the word about some issues that had not yet reached society. Moreover, in recent years we have seen how this issue has gone further and further, making it necessary for political leaders to sit down and make decisions on the regulation of AI. In the context of the European Union, the proposal of the Artificial Intelligence Act (European Parliament, 2023) is being discussed, which has among other purposes to establish a system to assess the risk that a technology incorporating AI may pose to the health and safety of people (Newtral, 2023).

This Final Master's Thesis will explore the debate through the news about Artificial Intelligence published in two Spanish digital media. Through web scraping techniques, the texts of the news will be obtained and later analyzed by using text analysis techniques. This work aims to contribute empirically to the analysis of news, therefore all the code used to perform the process of data extraction, cleaning, analysis, and visualization will be properly compiled in several rmd format files that will be uploaded in the following GitHub link. The work is structured in six different sections: first the objectives and relevance of this work are explained, making a small analysis of the existing literature on the subject. Subsequently, the methodology and data sources used will be briefly presented. The main section of the paper explains how data extraction, data cleaning, text analysis and visualization were carried out. Finally, the results are presented and conclusions are outlined.

2 OBJECTIVE AND RELEVANCE

The aim of this work is to make an empirical contribution in the field of news data extraction from digital newspapers and its text analysis through the R programming environment. The code created to perform the extraction of news content will be replicable, so anyone who wants to perform a similar analysis will be able to use this code and adapt it to extract the selection of news they need. This is important because at the moment these newspapers do not have their articles in any API or other accessible format, so data extraction using the web scraping technique is the only possible way to analyse the content. On the other hand, text analysis through R also makes it much easier to draw conclusions from large volumes of text without having to spend a lot of time reading it in detail. R has a large number of specialized natural language processing (NLP) packages that allow for more rigorous and efficient text processing and analysis, as well as textual data visualization tools that help to communicate these results in an understandable way. This code will also be available for anyone who wants to perform the analysis of a text database other than the one to be used in this paper.

The analysis of news published in the media is useful especially when there is no other source of data on the opinion of citizens on a particular topic. As the irruption of Artificial Intelligence is relatively recent, there are hardly any surveys or qualitative material (interviews, focus groups...) that allow us to know what citizens think about this technology. In this context, analyzing the press allows us to obtain valuable information on what are the main topics that are being discussed, what kind of data are being disseminated and what are the conflictive issues that are being put on the table. In addition, this work is going to include articles from newspapers of different ideologies, which also allows us to compare the positioning of each of them with respect to artificial intelligence. In short, the analysis of the text of the news on artificial intelligence that will be carried out in this work will make it possible to identify trends and main topics, evaluate the tone and the predominant opinion, detect biases and analyze the media coverage of this topic in the period of time analyzed.

There are examples of other research papers that have also applied text analysis techniques to study the content of media news. Fast and Horvitz (2017) conducted a text analysis of news about artificial intelligence published in the American newspaper "New York Times" for 30 years. In this study, they focused on analysing the text of the news and classifying it as "optimistic" or "pessimistic". In this way, they were able to show that not only was there more and more talk about this topic, but also more optimism. This study differs from the present work in that the authors were able to access the data through the newspaper's API. Garvey and Maskal (2019) also conducted a study in which they analyzed news published in this same newspaper (New York Times) about artificial intelligence between 1956 and 2018. These authors were able to show that media coverage of artificial intelligence was not negative. To do so, they performed text analysis using the Google Cloud Natural Language API Sentiment Analysis tool and, as in the previous case, they extracted the news using the newspaper's API.

The work is structured in several sections. First, the methodology used and the sources from which the information has been extracted will be explained. This section is

especially important since, as this is an empirical study, the data extraction work carried out is of special relevance. For this reason, a separate section will be dedicated to explaining how the data extraction process was carried out using the R programming language and the RStudio tool. Subsequently, it will be explained how the data have been cleaned. The fourth section will consist of a text analysis where the visualizations carried out will also be presented to allow further analysis. The work ends with a brief section with the conclusions and limitations of the work.

3 METHODOLOGY

3. 1 Data sources

This research work will use data extracted from two primary sources, the web sites of two Spanish digital newspapers: elDiario.es (https://www.eldiario.es/) and El Mundo (https://elmundo.es). These two newspapers have been chosen because they are two national generalist newspapers, therefore, they are comparable to each other. In addition, they have an ideologically different editorial line, which makes the comparison more interesting. While Diario.es is on the left of the ideological spectrum, El Mundo is on the right.

Data extraction was carried out using data harvesting techniques with the RStudio program. To this end, a code was developed to extract three different variables from each article published in these two newspapers: the title, the content of the article and the date of publication. Section 3 will explain the data extraction process in more detail.

3. 2 Sample

The sample used in this work is made up of all the articles published on the websites of the newspapers elDiario.es and El Mundo that contain the Spanish words "artificial intelligence" in the text of the article. In addition, it has been filtered by all articles published from July 2014 onwards, in order to have a comparable sample in both newspapers. Although the oldest article of these characteristics published by elDiario.es dates back to 2009, in the case of El Mundo the oldest date is July 2014. That is why it has been decided to remove all the rows corresponding to articles published in previous dates. On the other hand, the most recent article collected in the database created for this work dates from May 18, 2023. Both newspapers have at least one article published on that date.

The sample of articles does not include those with restricted access. Both digital newspapers have some articles for which it is necessary to pay a subscription to access them. These are the most recent articles, since they are made available to the general public after a certain period of time. Therefore, since the number of articles is small and limited, this is not a problem for the research.

A total of 1251 articles from El Mundo and 1056 from elDiario.es were analyzed. The following table shows the number of articles analyzed by newspaper and year of publication:

Table 1: Number of articles published by newspaper and year of publication.

Year	*elDiario.es*	*El Mundo*
2014	4	1
2015	21	10
2016	50	111
2017	69	229
2018	103	174
2019	99	175
2020	212	157
2021	122	137
2022	224	188
2023	152	69

3. 3 Data extraction

As explained in previous sections, the data analyzed in this research work were obtained by using "web-scraping" techniques performed with the RStudio tool. This is a not so conventional way of obtaining data but very useful when the data are not directly accessible. That is, in this case, there was no way to download the texts of the articles from the newspaper's website or from any other website. Therefore, the only possible alternative would be to save all the information manually (by copying and pasting the text into a dataframe). However, given the large amount of news published by newspapers this would be very time consuming. On the other hand, one of the main advantages of creating a reproducible R code is that it can be used to extract other types of news published on different dates. To do so, only minor modifications would have to be made to the original code.

All the code used in this work is accessible through GitHub, in rmd format. The file containing the data extraction part is called "web_scraping_ai.rmd". In this file is all the code used to generate a csv file with the extraction of the text of the articles, the title, and the date of publication. The file is structured in 14 different sections in which the data extraction process is explained step by step. It is, therefore, an easily replicable code in case someone needs to perform a similar extraction in some of the two newspapers used in this work. Next, we will proceed to explain how this process has been carried out and some of the problems that have arisen during the process.

3.3.1 Google News

At the beginning, the idea was to extract the text of the articles through Google News. Google News is a news search engine that allows you to enter a term or several

terms and a range of dates (e.g., *eldiario.es: "inteligencia artificial" site:eldiario.es after: 2023-04-03 before: 2023-05-03*) and it returns a list of links with the articles filtered by the newspaper or website you want. The advantage of using this search engine is that the code for both newspapers would be very similar. In addition, the article search criteria would also be similar which makes it more comparable. Although the rmd file shows all the code needed to extract the articles in this way, it was finally decided not to use that data. This is because Google News offers a limited number of links in its searches. That is, for each search made in this search engine, a maximum of 30 pages appear with 10 links per page. This limits the final number of articles to 300. However, it has been considered useful to keep the code that allows the extraction of data in this way, in case someone needs to use it.

3.3.2 RSelenium: extract the links

Once the option of extracting the data through Google News was ruled out, the only possible option was to use each newspaper's own search engines. To do so, the words "artificial intelligence" in quotation marks were entered into the search engines. In both cases a page appears showing the title (with a hyperlink to the page where the full article is located) and the date of publication of a list of articles. To see more, you need to click on the "next" button at the bottom of the page. The first thing to do, therefore, is to create a dataframe with three variables: the title, the date of publication and the link to the full article. To be able to do this it is necessary to use a specific R package called "RSelenium" that allows to interact from RStudio with the web page. It allows you to navigate and click on the links and web buttons remotely.

Before starting to use RSelenium it is convenient to install a Docker (https://www.docker.com/). This is a container where you can run applications and that makes the process of interacting with a website easier and more organized when using RSelenium. Once RSelenium is installed and initialized, it is necessary to create a code that allows reading the html of each page, extracting the necessary information through the XPaths and clicking on the "next" button to turn the page and repeat the process. This process is done through a loop that at the end saves all the information in a previously created dataframe. In addition, beforehand, it is necessary to use the "developer tool" on the web page of each newspaper's search to locate in the html where the information to be extracted is located and to create the appropriate XPath in R. This is done with the help of the "scrapex" library that incorporates functions that allow reading and extracting

As each newspaper has a different web design, the code is not interchangeable, but a separate XPath must be created for each newspaper's variable. Therefore, it is also necessary to create two loops that generate two dataframes, one for each newspaper. In addition, both websites have a different way of displaying the news once the "next" button is clicked. While on the website of the newspaper El Mundo each time this button is clicked a different page appears with the new links, on the website of elDiario.es new links appear after the previous ones. Thus, in the first case you have to extract the information and then click the button and extract again. In the second case, click on the "next" button as many times as possible and then extract all the information at once. In order to interact with the web, in the case of elDiario.es it is also necessary to accept cookies. To do this you must find the XPath that identifies the accept button and press it remotely.

3.3.3 Extract the text of the articles

Once the dataframe with the link to each article has been created, it is necessary to find again in the html the part where the text to be extracted is located and create the appropriate XPath. Once it has been verified that the XPath finds the necessary information, it is necessary to create again a loop that iterates through each link in the vector created with the links and extracts the text and saves it in a dataframe. Once this process has been carried out for the two newspapers, a dataframe with the three necessary variables is available. However, when reviewing the extracted data a complication arises: there are a large number of empty rows or "NAs" in the "text" column of the El Mundo dataset. This is because probably due to a change in the web, not all articles extract the text with the same XPath. This is why it is necessary to create a new loop that iterates through the links that do not have text and extracts it with the new XPath.

3. 4 Data cleaning

Once you have the two dataframes with the text of the articles, the date of publication and the title, you have to clean the data so that you can perform visualizations and text analysis. The code to perform this data cleaning is available in the rmd file "data_cleaning_ai". Before performing the data cleaning, duplicate rows are removed and the dataframes are merged: the one containing the text and the one containing the publication date and article title. The rows that initially contained empty values in the text column are also added. Subsequently, the dates are transformed to a format suitable for R.

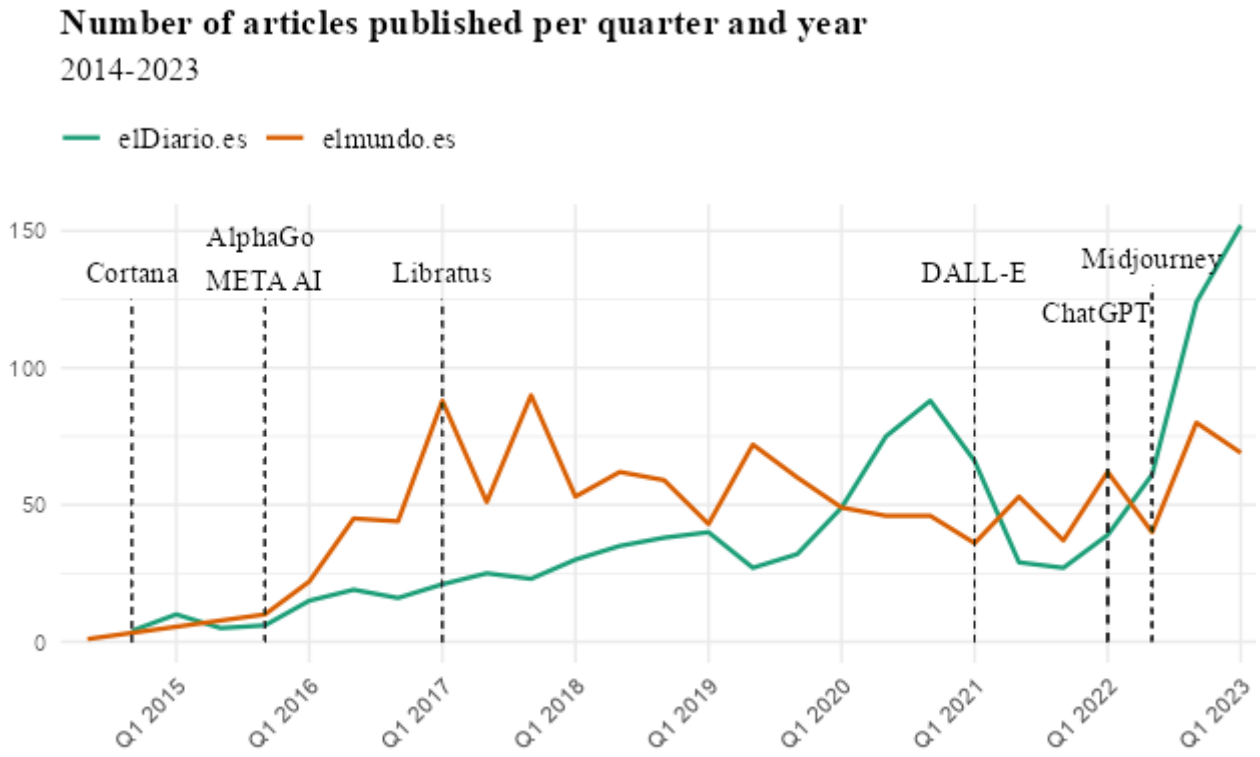
4 ANALYSIS AND RESULTS

4.4 Data visualization: trends in the number of articles published between 2015 and 2023

In order to view the number of articles published over time by both newspapers, a line graph has been made. First, the data have been filtered by the column "text" in order to keep only those containing the words "artificial intelligence" in their content, so that they can be compared. In addition, as explained in the section on the sample, the articles were limited to those published between 28-07-2014 and 30-04-2023. As the time period represented in the graph is quite broad, a column has also been created to indicate the four-month period of the year in which each date and year is included. This variable is the one represented on the X-axis. On the other hand, the Y-axis represents the absolute number of articles published during each semester. The green line represents the articles published by elDiario.es and the red line represents those published by El Mundo.

Key release dates of some artificial intelligence tools have also been included in the graph. This makes it possible to see if the pattern of article releases varies in relation to these events. All the code needed to perform this visualization is available in the rmd file "visualizations_ai.rmd".

Figure 1: Number of articles published per quarter and year (2014-2023)



Note: the y-axis shows the absolute frequency of news published on the website of each newspaper containing the words 'inteligencia artificial'

Source: own elaboration

The first thing that can be observed when looking at the graph is that both newspapers have followed a different trend in the publication of articles on artificial intelligence. While it is true that the number of articles published between 2015 and 2016 is very low in both cases, from 2016 onwards in the case of El Mundo the absolute number of publications increases significantly. It goes from having about 25 articles published at the beginning of 2016 to having more than three times as many in the first four months of 2017 and 2018. The trend during those years for elDiario.es is also upward, but much more slowly. It is not until 2020 that the trend experiences a very significant growth reaching almost 90 articles published during the third quarter of that year.

While between 2015 and 2020 El Mundo leads the ranking of number of articles published on IA, from 2020 onwards elDiario.es takes the lead. It is worth noting the peak of articles published between 2020 and 2021 by elDiario.es. Although no notable IA tool was launched at this time, it is true that as a consequence of the Covid-19 pandemic there was much talk at that time of the health advances that could be achieved with the introduction of IA in the field of medical science. Finally, eldiario.es reaches its maximum in the first quarter of 2023 with more than 150 articles published. At the same point El Mundo does not reach 75 articles, less than half.

4.3 Text analysis

In this section we will proceed to explain how the text analysis has been performed in R and to present the main results. Some data visualizations related to the text analysis have been made to allow a better understanding of the content of the news being analysed.

4.3.1 Tokenize the text

In order to perform the text analysis in R, the first step has been to clean the text column of some news rows. In the case of the news of the newspaper elDiario.es, the text referring to an international news agency ("Agencia EFE") has been removed, as well as the location of the publication, since it did not provide useful information. Subsequently, the text of the article was separated into words ("tokens"), so that a new dataframe was generated with a row for each word ("Word" column). The next step has been to eliminate those words that are not useful for text analysis. This is done by using a library ("stopwords") that allows to store in a vector the useless words in Spanish. This list of words includes different verb forms (estar, ser, haber, tener and hacer), about 20% of prepositions, conjunctions (8%), and articles (7%).

4.3.2 Word frequencies

Once the text has been split into words, it is worthwhile to see the frequency of words in absolute numbers per newspaper. For this purpose, two bar charts have been created in which we filter by the 15 most repeated words in each newspaper.

Figure 3: word frequency plot (eldiario.es)

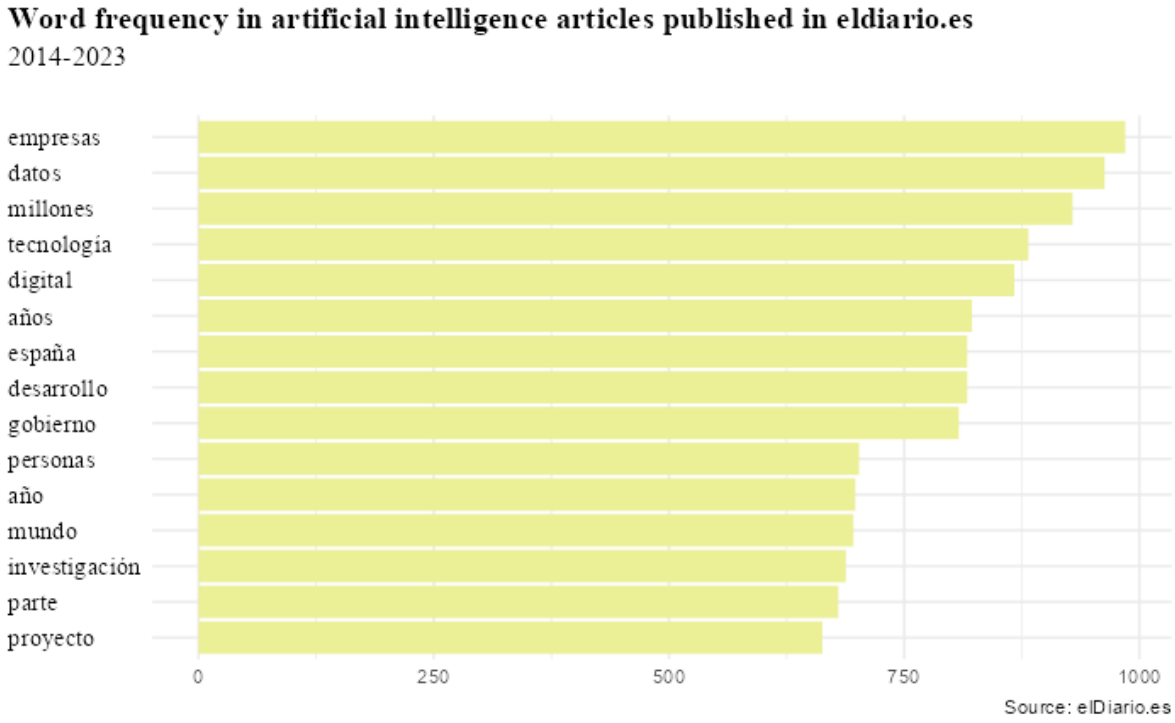
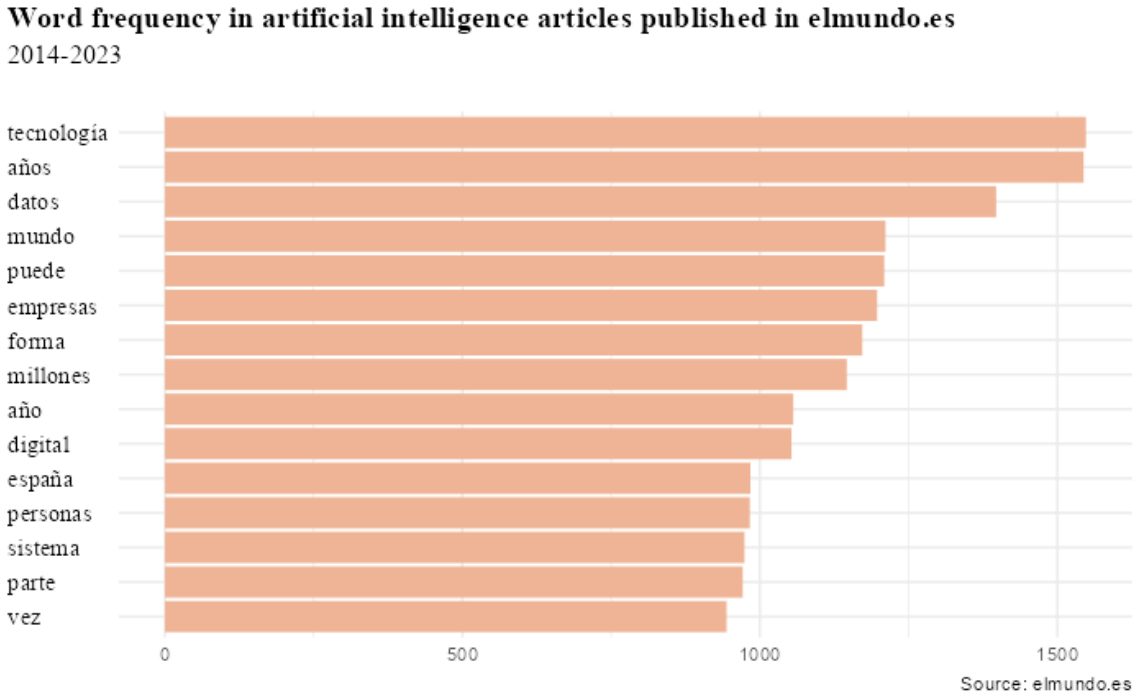


Figure 4: word frequency plot (elmundo.es)



In both cases the most repeated words are similar: companies, technology, data, digital, year or years... There are nine of the 15 words that are common to both newspapers. It is worth noting in the news of eldiario.es that words related to the field of research (research, project, government, development) are repeated very frequently, which suggests that this newspaper relates the use of artificial intelligence more to the scientific field. If we calculate the correlation coefficient (Figure 5) in both data sets we obtain a value of 0.97, which indicates that the texts of both newspapers in the period analysed are very similar.

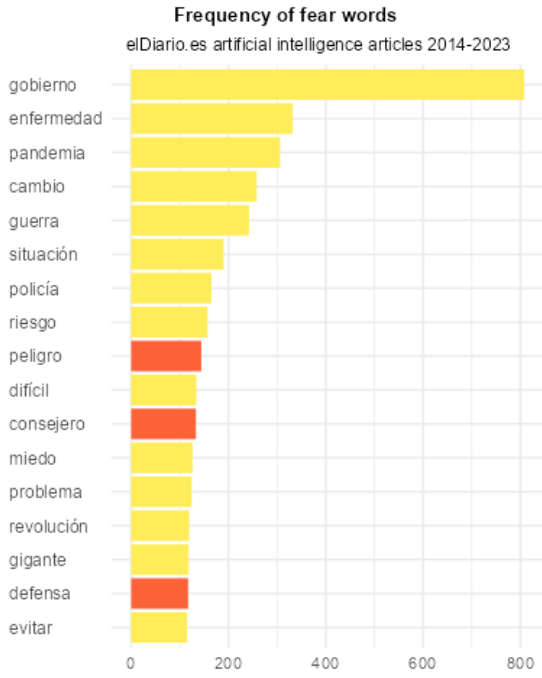
Figure 5: correlation coefficient

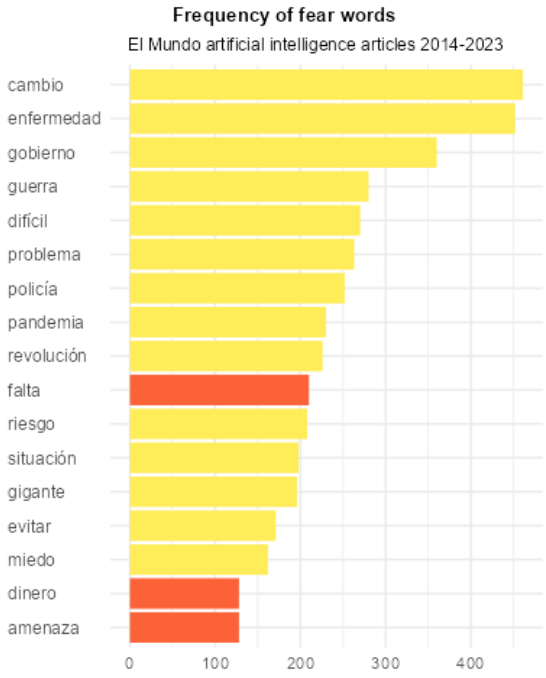
```
        Pearson's product-moment correlation

data:  proportion and elmundo.es
t = 660.34, df = 22096, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9749431 0.9762151
sample estimates:
      cor
0.9755872
```

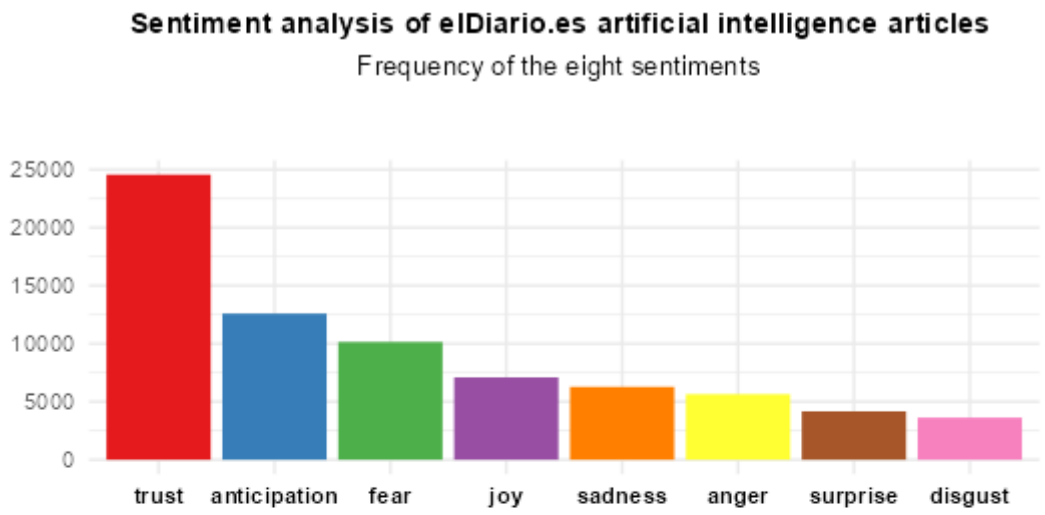
Below, you can also see another way of visualizing word frequency by using word clouds. In this case the most repeated words are shown in larger size. Words with a similar absolute frequency are also grouped by color.

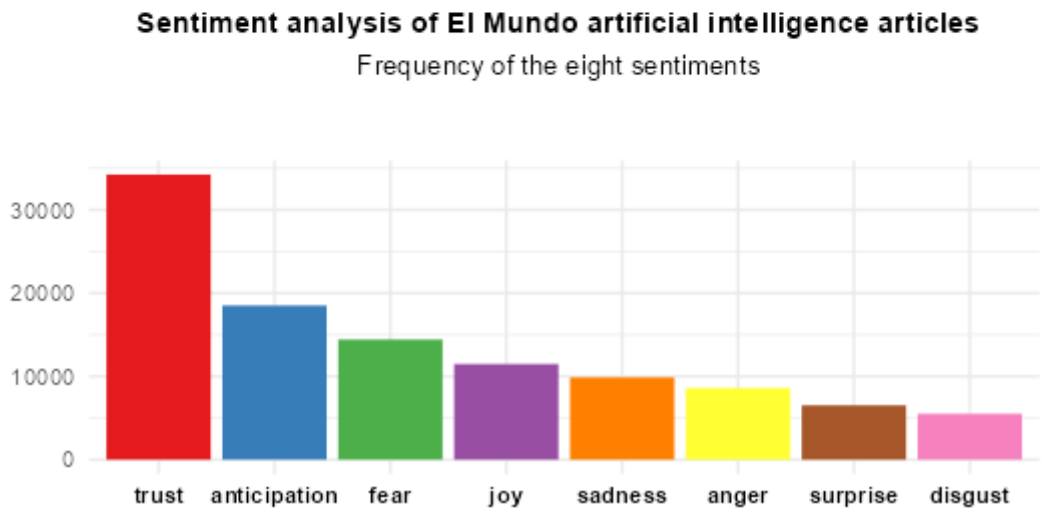
Figure 6: Word clouds

4.3.4 Sentiment analysis









5 CONCLUSIONS AND LIMITATIONS

6 REFERENCES

- Newtral. “Bruselas ultima una propuesta de ley para regular el desarrollo de la Inteligencia Artificial en la UE”. Retrieved from: https://www.newtral.es/ley-inteligencia-artificial-ia-chatgpt-union-europea-ue/20230427/
- European Parliament. EU AI Act: first regulation on artificial intelligence. Retrieved from: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

- elDiario.es. Búsqueda: “inteligencia artificial”. Retrieved from: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence https://www.eldiario.es/busqueda/%22inteligencia%20artificial%22
- elmundo.es. Buscador: “inteligencia artificial”. Retrieved from: https://ariadna.elmundo.es/buscador/archivo.html?q=%22inteligencia+artificial%22&b_avanzada=
- Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 31, No. 1).
- Garvey, C., & Maskal, C. (2020). Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence. *Omics: a journal of integrative biology*, 24(5), 286-299.