Comparative Analysis of Word Frequency in Science News: Colombia vs. USA and Colombia vs. Spain

Inti Leandro Ruiz Castro, Leidy Paola Alfonso Acosta
Universidad Nacional de Colombia

July 5, 2023

Abstract

This article presents an analysis of the content of scientific news in selected media outlets from Colombia, Spain, and the United States. The study aims to identify patterns, differences, and similarities in the topics addressed by the press in each country, using web scraping techniques and natural language processing. The data was collected from the Google News feed during a 15-day period in the first semester of 2023. The media outlets selected for each country were chosen based on popularity and the presence of a dedicated science section. The collected news articles were processed using natural language analysis techniques. With the data, the frequency of words was calculated to identify the most common topics in each country, the number of publications, the probability density of word frequency, and entropy. The results showed that Colombia has the highest number of news articles, while Spain has the longest articles on average. It was evident that all three countries consistently publish a constant number of articles day by day. Finally, differences in entropy were found, revealing variations in the diversity of the language used.

Palabras clave: web scraping, noticias de ciencia, NLP

1 Introducción

The media can be considered a reflection of the interests of its audience, and the coverage of scientific news, in particular, can provide insight into the public's interest in science. The media plays an important role in bridging the gap between the public and scientific advancements by presenting science in an accessible language and a more engaging narrative. On the other hand, science coverage in different countries can help identify trends and patterns in how science is presented and perceived, and understand how cultural, social, and economic factors can influence the way science is reported in different parts of the world.

Web scraping is a data extraction technique used to obtain information from web pages in an automated manner, making it a powerful and useful technique for gathering information. The use of this technique has progressively increased in recent times due to technological advancements that facilitate its access. It has been employed in various fields, ranging from particle physics to create a web portal dedicated to Higgs bosons for experts and the general public [1], to meteorology for collecting data in regions where traditional methods may be limited or unavailable [2], and even in anthropology, where it has been used to study racial disparities in policing practices [3].

Natural language extraction and processing techniques have experienced exponential development in recent years. From the early published works (applied to news) for filtering redundant or useless information [4] or extracting themes and quotes from news [5], to current methods that use neural networks to predict future news feed content [6]. Currently, these techniques represent a very important and versatile tool for analysis, applicable in a wide variety of areas. Additionally, the analysis of scientific news has been used as a pedagogical tool to improve cognitive development and performance in students [7], as well as for social and demographic analysis of entire countries [8]. Motivated by these considerations, works combining both concepts have been published. By using natural language analysis on scientific news, exaggerated claims [9] or fake news [10] have been identified, and truthfulness indices of news articles have been generated [11], all with the aim of reducing misinformation on scientific topics. Finally, the idea of a comparative study between scientific news in newspapers from different countries has already been explored, specifically between the Danish and British press, driven by the differences in how press is conducted in each country [12]. In this article, it was found that scientific news represents approximately 4

Conducting such an analysis through data science tools on the content of science news is not a new endeavor. However, as of the time of writing this document, there is no such study considering news from Colombian and American media outlets, as well as Colombian and Spanish media outlets. Analyzing news articles by country using web scraping and data science provides an objective way to identify patterns and trends, gain international perspectives, and study thematic coverage to pinpoint the topics of interest to different populations regarding science. Comparing Colombia vs. the United States and Colombia vs. Spain in terms of science coverage in the media is useful due to the cultural differences and distinctive characteristics that these countries present. This can help understand how cultural, social, linguistic, and economic factors can influence the way science is reported in different parts of the world, and potentially bridge the development gap.

The objective of this study is to analyze the content of science news in selected Colombian and American media outlets, as well as Colombian and Spanish media outlets, available on the Google News feed during a 15-day period in the first semester of 2023. Through web scraping and data science techniques for natural language analysis, the frequency of words will be obtained from the extracted information to identify patterns, differences, and similarities in the topics covered by the press in each country. The process of data preparation, the conditions regarding the considered news articles, and the procedure are described in Section 2. The results obtained and their respective analysis are presented in Section **??**. Finally, the conclusions are outlined in Section **??**.

2 Methodology

2.1 Data Source

The data consists of the content of news articles, including both the headlines and the body text, as well as suggested readings within the article. The news articles must appear in the Google News feed and belong to one of the selected media outlets for each country. Five media outlets were considered for each country, which must meet three conditions: having a dedicated section for science-related topics, being among the most consulted media outlets in their respective countries, and being locally based in the respective nation. The popularity of the media outlets was determined based on the German online statistics portal *Statista*, which collects statistical data on over 80,000 topics from more than 22,500 sources and makes them available to users.

For Colombia[1], the five most popular media outlets that meet the aforementioned conditions were: *El Tiempo*, *Semana*, *Pulzo*, *El Espectador*, and *Caracol.* On the other hand, the most widely read Spanish media outlets[2] that meet the criteria are: *El país*, *El Mundo*, *La Vanguardia*, *ABC*, and *La voz de Galicia.* Finally, for the United States[3], the selected media outlets for the database are: *CNN*, *MSN*, *Fox News*, *The Washington Post*, and *Yahoo.* It is worth noting that *The New York Times* fulfilled all the conditions mentioned, but its digital platform does not allow the web scraping process using the method considered in this work.

2.2 Procedure

A program was created from scratch using the *Python* programming language. The news articles are obtained using the *gnewsclient* library, filtering by the category "Science" and the specific countries. This allows extract-

[1] https://www.statista.com/statistics/1012047/colombia-news-websites/

[2] https://es.statista.com/estadisticas/476795/periodicos-diarios-mas-leidos-en-espana/

[3] https://www.statista.com/statistics/381569/leading-news-and-media-sites-usa-by-share-of-visits/

ing the news articles of that classification for each country present in the Google News feed. The first 50 news articles present in the feed per day were considered, and the selection criteria mentioned in subsection 2.1 were applied to this group. For the news articles that meet the criteria, their headline, URL, and content are saved in an external file for each country. Finally, a manual review was conducted to ensure that the saved news articles correspond to a science topic and to eliminate any potential mistakenly filtered news articles. The news articles were collected daily for a period of 15 days, from March 29th to April 12th, at 22:00 (UTC-5).

Once the database was completed, the content of each news article goes through a cleaning process, which involves removing links, numbers, special characters, punctuation marks, accents, emojis, and specific headers and "hook phrases" used by each media outlet. Figure 1 shows a fragment of the content of a news article before and after the cleaning process.

Next, the *nltk* library is used to remove stop words (also known as empty words), which are words without content for the considered topic, from the news article content in each language. Additionally, a set of custom-defined empty words by the authors is also eliminated. Once the text is free of stop words, the *Stanza* library is employed to perform lemmatization, which involves transforming the inflected forms of words into their corresponding lemma. For example, "jugadoras" is changed to "jugador" and "corríamos" is replaced by "correr."

After performing the aforementioned processes on each news article from the three countries, the frequency of occurrence of each word across the entire dataset is obtained for each country. The relative frequency is used for making comparisons between them.

The frequency distributions of words in natural languages approximately follow a power law [13] of the form:

$$p(x) = Cx^{-\alpha} \tag{1}$$

where C and α are fitting parameters. This behavior can be quickly verified by plotting the frequency on a logarithmic scale. Following this, the distributions were plotted, and a fitting process was carried out using equation 1.

On the other hand, information entropy is a measure that reflects the diversity or variety of information in a dataset. High entropy indicates greater diversity, while low entropy suggests lower diversity, implying that the data is more predictable or uniform. For the calculation of entropy, Shannon entropy is used in the following form:

$$\mathrm{H}(X) := -\sum x \in \mathcal{X} p(x) \log_2 p(x) \tag{2}$$

where $p(x)$ corresponds to the probability of each word, and the summation is performed over all words. The probability of each word can be estimated using the maximum likelihood estimation method [14]:

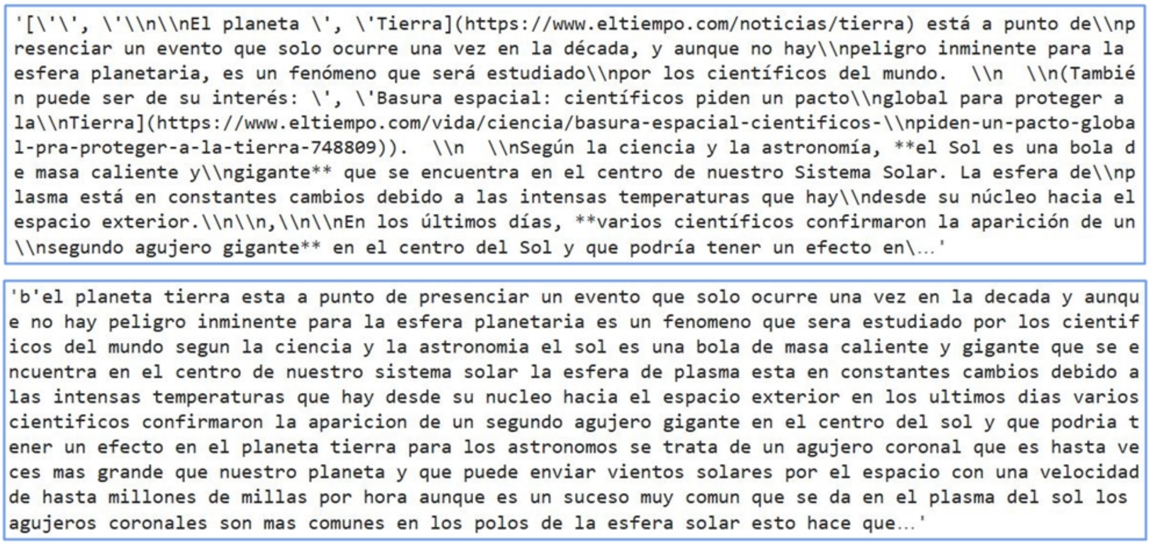
$$p(x_i) = \frac{f_i}{\sum_{i=1}^{N} f_i} \tag{3}$$

where f_i is the frequency of occurrence of the word x_i, and the denominator is the sum of frequencies over a finite

The dataset of science news for Colombia, collected as mentioned above, consists of a total of 224 news articles, with an average length of 507 words. On the other hand, for Spain, a total of 142 news articles were collected, with an average size of 821 words. Finally, for the United States, 53 news articles were obtained, with an average length of 709 words. The length of the news articles was obtained before the process of removing stop words. It should be noted that a longer news article does not necessarily mean it has richer or more diverse content. Therefore, a diversity analysis of words, similar to the one conducted in [15], and a brief study of entropy according to [14], are performed.

In general, based on the above information, it can be affirmed that Colombia published the highest number of news articles during the 15-day period, followed by Spain, and lastly, with a significant difference, the United States. However, in terms of article length, Spain had the longest news articles on average, followed by the United States, while Colombia had the shortest ones. More specifically, Colombia publishes 1.5 times more news articles than Spain. However, Spanish news articles are 1.6 times longer than Colombian ones. On the other hand, Colombia publishes 4.2 times more news articles than the United States, but the length of the American news articles is 1.5 times larger than the Colombian ones.

An important point to highlight in this comparison is that when filtering by the *Science* category and the country *United States* in the Google News feed, a large portion of the top news articles comes from specialized



```
'[\'\', \'\\n\\nEl planeta \', \'Tierra](https://www.eltiempo.com/noticias/tierra) está a punto de\\np
resenciar un evento que solo ocurre una vez en la década, y aunque no hay\\npeligro inminente para la
esfera planetaria, es un fenómeno que será estudiado\\npor los científicos del mundo.  \\n  \\n(Tambié
n puede ser de su interés: \', \'Basura espacial: científicos piden un pacto\\nglobal para proteger a
la\\nTierra](https://www.eltiempo.com/vida/ciencia/basura-espacial-cientificos-\\npiden-un-pacto-globa
l-pra-proteger-a-la-tierra-748809)).  \\n  \\nSegún la ciencia y la astronomía, **el Sol es una bola d
e masa caliente y\\ngigante** que se encuentra en el centro de nuestro Sistema Solar. La esfera de\\np
lasma está en constantes cambios debido a las intensas temperaturas que hay\\ndesde su núcleo hacia el
espacio exterior.\\n\\n,\\n\\nEn los últimos días, **varios científicos confirmaron la aparición de un
\\nsegundo agujero gigante** en el centro del Sol y que podría tener un efecto en\...'
```

```
'b'el planeta tierra esta a punto de presenciar un evento que solo ocurre una vez en la decada y aunqu
e no hay peligro inminente para la esfera planetaria es un fenomeno que sera estudiado por los cientif
icos del mundo segun la ciencia y la astronomia el sol es una bola de masa caliente y gigante que se e
ncuentra en el centro de nuestro sistema solar la esfera de plasma esta en constantes cambios debido a
las intensas temperaturas que hay desde su nucleo hacia el espacio exterior en los ultimos dias varios
cientificos confirmaron la aparicion de un segundo agujero gigante en el centro del sol y que podria t
ener un efecto en el planeta tierra para los astronomos se trata de un agujero coronal que es hasta ve
ces mas grande que nuestro planeta y que puede enviar vientos solares por el espacio con una velocidad
de hasta millones de millas por hora aunque es un suceso muy comun que se da en el plasma del sol los
agujeros coronales son mas comunes en los polos de la esfera solar esto hace que...'
```

Figure 1: Example of the content of the news article *Nasa halla agujero gigante en el Sol: sus efectos se podrían ver desde la Tierra* published by the newspaper *El Tiempo* before and after text cleaning.

journals such as *Nature* or science government institution portals like *NASA*. However, these types of media were not considered given the objective and focus of this work. This behavior was not observed in the case of Colombia.

In Table ??, the first 10 most used words per country in the preprocessed science news are shown. From this information, it is evident that both Colombia and the United States frequently use language related to the field of astronomy (moon, planet, galaxy, telescope). On the other hand, although Colombia and Spain do not reflect the same topic of interest, they coincide in several words such as *año* (year) and *científico* (scientific). The content of the news articles in Colombia and Spain uses words to refer to individuals such as *investigador* (researcher) and *científico* (scientist), while these words do not appear in the United States, indicating a more impersonal character. In all three countries, the word *nuevo*/*new* appears in the ninth and eighth positions, respectively.

Colombia		Spain		USA	
Word	Frecuency	Word	Frecuency	Word	Frecuency
año	323	año	355	space	208
tierra	275	estudio	252	nasa	156
científico	247	cientifico	240	galaxy	119
luna	246	solo	178	study	113
planeta	239	vida	173	time	111
estudio	239	persona	173	telescope	110
investigador	196	humano	169	egg	110
persona	190	universidad	167	new	108
nuevo	178	nuevo	159	webb	104
eclipse	176	ciencia	142	planet	100

Table 1: Top 10 de palabras más usadas en las noticias de ciencia junto al número de veces que estas fueron empleadas en todos los artículos considerados por país.

The content of the news, in terms of vocabulary, cannot be analyzed by limiting it solely to the top 10 most used words. Therefore, the analysis was extended up to the 50th position. Figure ?? shows the number of occurrences of each word in the group of science news per country, normalized by the highest frequency, for the top 50 most used words in the news.

Firstly, it is evident that all three graphs exhibit a power-law behavior, which was confirmed by plotting the

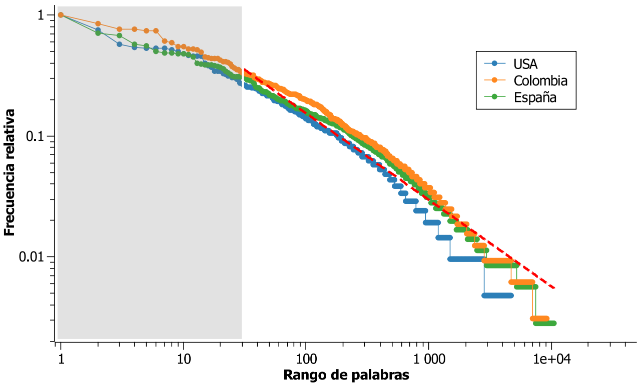


Figure 2: Logarithmic scale word frequency

data on a logarithmic scale (shown in Figure 2) and fitting it to the equation **??**. The results of the fitting for each country are presented in Table 2.

	Colombia	Spain	USA
C	5.189	4.16	4.66
α	0.73	0.71	0.77

Table 2: Fitting parameters for word frequency

The coefficient of determination (R^2) for all three countries is 0.97. It is important to note that the fitting was performed over a specific interval of the data, which is common in the analysis of systems exhibiting power-law behavior [13].

Regarding the most frequent words in positions 10 to 20, both Colombian and American media maintain a vocabulary related to astronomy, although for the United States, more diverse words begin to appear. On the other hand, in the case of Spain, the words in these positions can be associated with global terms that do not indicate a specific theme. However, it is worth noting that Spain is the only country to mention the word ”woman” among the top 50 most frequent words in the news.

In the range of positions 20 to 30, all countries exhibit greater diversity in vocabulary, with words that can be related to different fields of science (other than astronomy). For the United States, words like ”dinosaur,” ”ant,” and ”water” are found. In the case of Colombia, words like ”water,” ”brain,” or ”plant” appear, while for Spain, words like ”mathematician” or ”cell” are present. It is evident that biology can be associated with at least one of the mentioned words in all three countries, but there is also content from other sciences such as paleontology or mathematics.

In positions 30 to 40, Colombia shows a vocabulary oriented towards result analysis, with words like ”result,” ”data,” or ”greater.” Meanwhile, the United States returns to the astronomical theme, and Spain maintains a different focus, without focusing on a specific topic, using words like ”change,” ”society,” and ”system.” Finally, in the interval from 40 to 50, astronomical terminology reappears in the news from the United States, with words like ”solar” or ”James” (referring to the recent James Webb telescope, with its surname ranking ninth). In the case of Colombia, there is a vocabulary that cannot be exclusively related to science, with words like ”country” or ”hand.” Finally, for the Spanish news, there is a presence of nouns that are easily related to scientific work, such as ”physicist,” ”data,” or ”effect.”

The verb ”explicar” (to explain) was the most commonly used verb in both Spain and Colombia, while the second most frequently used verb was ”encontrar” (to find) for Colombia and ”saber” (to know) for Spain. As for the United States, the top two verbs were ”find” and ”show,” although it should be noted that ”like” has been omitted due to its ambiguity in the English language.

In general, these are words that are commonly used in science, so it is not surprising that they are frequently employed in the news.

In Figure **??**, the total number of news articles published per day, along with the cumulative number of news articles, is shown for each country. It is evident that Colombia generally publishes a higher number of news articles compared to Spain and the United States. On day 12, corresponding to April 9th, all three countries had

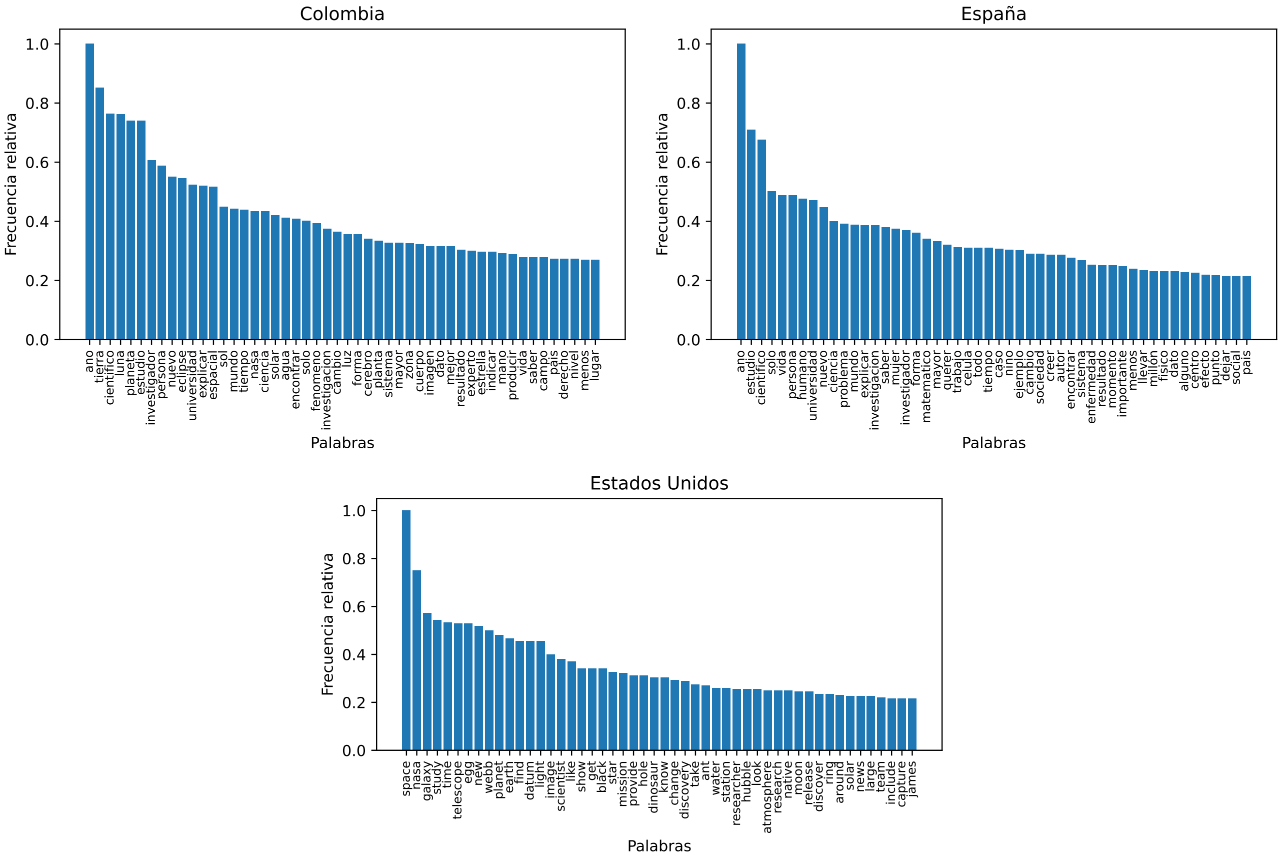


Figure 3: Relative frequency for the first 50 words per country. These words do not include accents or ”ñ” due to the cleaning process.

a similar daily number of news articles. The daily number of news articles exhibits an approximately constant behavior, and correspondingly, the growth of the cumulative news articles is linear, albeit with different slopes for each country. This indicates that the Google News feed, in relation to the considered media outlets, does not distinguish between the quantity of news articles published on weekdays or weekends, at least not in a combined manner.

The graph 5 shows the probability density distribution of word occurrence frequencies in science news, normalized for the United States, Colombia, and Spain. For all three countries, two regimes can be observed, divided by a relative frequency value of approximately 10^{-1}. For lower relative frequencies, there is a linear decay trend in the probability density, which is consistent with language behavior. It is also observed that the data for Spain and Colombia exhibit a very similar distribution, which is expected since these countries share the same language. However, for the regime corresponding to relative frequencies greater than 10^{-1}, there is a highly scattered behavior compared to the previous one, where a linear trend cannot be identified. This behavior has been observed in studies such as [16], where they investigated scaling properties in Twitter corpora at the city level in the United States.

The graph indicates that there are mostly words with low relative frequencies (and therefore, low frequencies). As the relative frequency increases, the number of words with that value of relative frequency decreases. In the scattered regime, words with high relative frequencies are located, but their density reaches the lowest values, indicating that there are few words that share the same relative frequency value. This behavior has been studied in systems that follow a power-law distribution and is generally categorized as noise, with its intensity varying depending on the system [13].

Finally, the entropy was calculated for the news collected per day and per country, and the values are shown in Figure 6. It can be observed that both Spain and Colombia exhibit approximately constant behavior with a

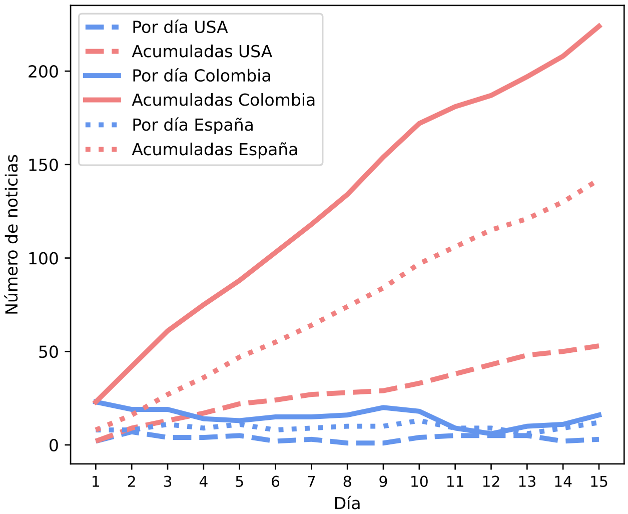


Figure 4: Distribución de la densidad de probabilidad de la frecuencia relativa de aparición de palabras en noticias de Ciencia para cada Estados Unidos, Colombia y España.

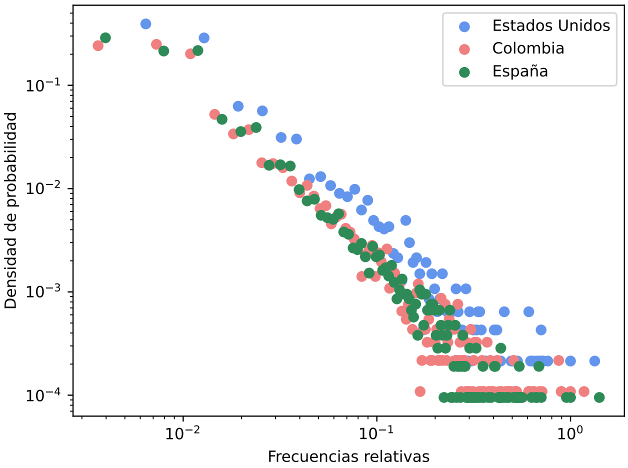


Figure 5: Probability density distribution of the relative frequency of occurrence of words in Science news for the United States, Colombia, and Spain.

mean value close to 10.5 bits. Colombia shows a global minimum on day 11, coinciding with the day when the lowest number of science news was recorded for this country. Overall, it is noteworthy that both Colombian and Spanish media maintain a notable stability in the diversity of their news throughout the days. This pattern is intriguing, as variations in the quantity and diversity of news would be expected, especially between weekdays and weekends.

On the other hand, the United States shows a variable behavior over time with a mean value close to 9 bits. On day 7, a global minimum is observed, during which only astronomy news was present. Thus, science news from Colombian media is more diverse than that of US media.

3 Observations

It is important to note that the results obtained in this study do not necessarily provide an exhaustive description of the approach to scientific news in the media of each country. This is because the news used in the dataset were obtained through the Google News feed, which has its own filtering algorithm. This implies that

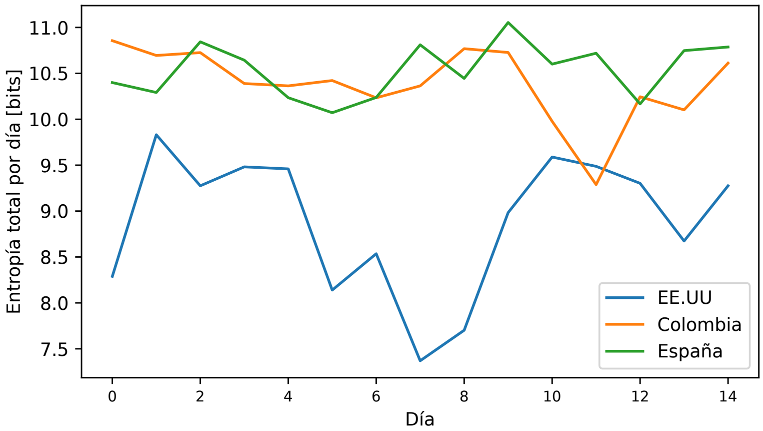


Figure 6: Entropy measured in bits for each day and country.

the news was not directly selected from the portals of each media outlet and their daily publications. Therefore, it is possible that some specific aspects of the media's focus on science coverage are not fully reflected in this work. Instead, the main objective of this study is to provide a more general overview of trends in scientific news coverage, avoiding a narrow focus solely on the media.

Additionally, through algorithms and preference analysis, Google News seeks to offer users a personalized and relevant experience by showing news that aligns with their interests and preferences. Thus, although the news used in this study do not directly come from the portals of each media outlet, they likely reflect to some extent the preferences and general trends of audiences regarding science coverage.

During the preparation phase, while selecting the media outlets to be used, it was evident that the science section of several news portals most read by the Colombian population was either not updated or had low publication frequency (once a month). For Spain, a large number of the most read outlets are portals dedicated exclusively to sports, without a dedicated science section. On the other hand, in the United States, as mentioned earlier, the majority of publications come from specialized media outlets.

Upon reviewing the collected news from the three countries, articles that remain registered in the Google News feed for more than one consecutive day have been found, suggesting that the system does not completely refresh its content on a daily basis.

Lastly, it is important to mention that, due to the low volume of data, it is not possible to assert with absolute certainty that some variations in the results correspond to system behavior, as they may be attributed to inherent statistical noise.

4 Conclusiones

- The word frequency in the three countries follows a power-law distribution, which is a common characteristic in natural languages. Additionally, it was found that the three countries maintain an approximately constant number of daily publications.
- Colombia had the highest number of news articles published during the 15-day period; however, it also had the lowest average article length. This may suggest that the goal of the media is to achieve mass dissemination. In contrast, Spain had a lower daily number of science news compared to Colombia. However, it is important to note that the limited quantity of news is compensated by a greater extent and diversity in the topics covered. Although the absolute number may appear small, it is evident that the landscape of scientific news in Spain is characterized by a more detailed and comprehensive approach in each article.

- The analysis of word frequency indicates that Colombian and American news primarily focus on topics related to astronomy, suggesting that other branches of science are excluded from the publications. Despite the lower frequency of publication, news articles from the United States are characterized by considerably greater length compared to Colombian news.

  On the other hand, although it was observed that the entropy of news in Colombia is higher than that of news in the United States, it is important to consider that this difference may be attributed, at least partially, to the language disparity between the two countries. Therefore, it is not possible to assert that the difference originates from the diversity of the topics covered.

- The use of web scraping techniques and data science in the collection and analysis of science-related news has provided the opportunity to systematically examine a significant volume of information sources. This approach has allowed for the identification of both similarities and differences among news articles, revealing patterns and trends that would be difficult to detect using conventional methods.

For future work, it is recommended to use a larger database by extending the period of news collection. This would allow for a more representative sample of the scientific theme. Additionally, it is essential to carry out a more thorough and refined data cleaning process, ensuring the quality and reliability of the analyzed information. To further enrich the analysis, it is suggested to include a Spanish-speaking country geographically close to Colombia, and likewise for the United States. This would enable the examination of similarities and differences in the coverage and content of scientific news between countries with similar cultural and geographical contexts.

References

[1] Kaushik De et al. “A Portal Dedicated to Higgs Bosons for Experts and the General Public”. In: *Computing in Science & Engineering* 15.5 (2013), pp. 46–53.

[2] Riko Hendrawan Marpaung, Jumail Sitorus, and Andika Yudha Sembiring. “Web Scraping Techniques to Collect Weather Data in South Sumatera”. In: *Journal of Physics: Conference Series* 1155.1 (2019), p. 012060.

[3] R. Tibshirani, S. Wager, and S. Athey. “A large-scale analysis of racial disparities in police stops across the United States”. In: *Nature Human Behaviour* 3.2 (2019), pp. 126–132. DOI: 10.1038/s41562-018-0528-8.

[4] Ian Garcia and Yiu-Kai Ng. “Eliminating redundant and less-informative RSS news articles based on word similarity and a fuzzy equivalence relation”. In: *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’06)*. IEEE. 2006, pp. 465–473.

[5] Luis Sarmento and Sérgio Nunes. “Automatic extraction of quotes and topics from news feeds”. In: *DSIE’09-4th Doctoral Symposium on Informatics Engineering*. 2009.

[6] Vasiliy Osipov et al. “Neural network forecasting of news feeds”. In: *Expert systems with applications* 169 (2021), p. 114521.

[7] Pei-Ying Tsai et al. “Effects of Prompting Critical Reading of Science News on Seventh Graders’ Cognitive Achievement.” In: *International Journal of Environmental and Science Education* 8.1 (2013), pp. 85–107.

[8] Gunver Lystbæk Vestergård and Kristian H Nielsen. “From the preserves of the educated elite to virtually everywhere: A content analysis of Danish science news in 1999 and 2012”. In: *Public Understanding of Science* 26.2 (2017), pp. 220–234.

[9] Yingya Li, Jieke Zhang, and Bei Yu. “An NLP analysis of exaggerated claims in science news”. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. 2017, pp. 106–111.

[10] Ray Oshikawa, Jing Qian, and William Yang Wang. “A survey on natural language processing for fake news detection”. In: *arXiv preprint arXiv:1811.00770* (2018).

[11] Mauricio Gruppi et al. “SciLander: Mapping the Scientific News Landscape”. In: *arXiv preprint arXiv:2205.07970* (2022).

[12] Gunver Lystbaek Vestergaard and Kristian Hvidtfelt Nielsen. “Science news in a closed and an open media market: A comparative content analysis of print and online science news in Denmark and the United Kingdom”. In: *European Journal of Communication* 31.6 (2016), pp. 661–677.

[13] Mark EJ Newman. “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary physics* 46.5 (2005), pp. 323–351.

[14] Christian Bentz et al. “The Entropy of Words—Learnability and Expressivity across More than 1000 Languages”. In: *Entropy* 19.6 (2017). ISSN: 1099-4300. DOI: `10.3390/e19060275`. URL: `https://www.mdpi.com/1099-4300/19/6/275`.

[15] José Alberto Ruiz Gayosso. “Complejidad y diversidad de palabras en publicaciones científicas”. Tesis de Maestría en Ciencias. Universidad Autónoma de México, 2023.

[16] Eszter Bokányi, Dániel Kondor, and Gábor Vattay. *Scaling in Words on Twitter*. 2019. arXiv: `1903.04329 [physics.soc-ph]`.

[17] G. R. Hayes et al. “Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 15.1 (2021), pp. 228–232.