Comparative analysis of word frequency in science news from Colombia vs. the United States and Colombia vs. Spain.

Leidy Paola Alfonso Acosta, Inti Leandro Ruiz Castro
Facultad de Ciencias, Departamento de Física
Universidad Nacional de Colombia

May 31, 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 covered by the press in each country, utilizing web scraping techniques and natural language processing. The data was collected from the Google News feed over 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. From the data, word frequency was calculated to identify the most common topics in each country, the number of publications, the probability density of word frequencies, and entropy. The results showed that Colombia has the highest number of news articles, while Spain has the longest articles on average. It was observed that all three countries consistently publish a steady number of articles day by day. Finally, differences in entropy were found, indicating variations in the diversity of the employed language.

Keywords: web scraping, science news, NLP.

1 Introduction

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

Web scraping is a data extraction technique used to gather information from web pages in an automated manner, making it a powerful and useful method for collecting information. Its usage has progressively increased due to technological advancements that have facilitated 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 gathering 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 processing (NLP) extraction and processing techniques have experienced exponential growth in recent years. From early publications focusing on filtering redundant or irrelevant information in

news articles [4] or extracting themes and quotations from news [5], to current applications using neural networks to predict future content in news feeds [6], these techniques have become essential and versatile tools for analysis in various fields. Additionally, the analysis of scientific news has been employed as a pedagogical tool to improve cognitive development and performance in students [7], as well as for social and demographic analyses on a national scale [8]. Motivated by these factors, studies have been conducted that combine both concepts. By applying natural language processing analysis to scientific news, exaggerated claims [9] or fake news [10] have been identified, and truthfulness indices for news articles have been generated [11], all aimed at reducing misinformation in scientific topics.

Furthermore, the idea of a comparative study between science news in newspapers from different countries has already been explored, specifically between Danish and British press, driven by the differences in journalistic practices for each country [12]. In this study, it was found that in both countries, scientific news represented around 4% of the total news articles. However, it was observed that in Denmark, scientific news was often triggered by political events, prioritized national stories, and contained more coverage of humanities and social sciences. In contrast, in the United Kingdom, scientific news was more traditional and focused on health and natural sciences, usually based on journal articles.

While conducting data-driven analyses of the content of science news articles is not a new endeavor, at the time of writing this document, no such study has been conducted specifically considering news from Colombian and U.S. media outlets, as well as Colombian and Spanish media outlets. Analyzing news by country using web scraping and data science provides an objective way to identify patterns, trends, and international perspectives, as well as study thematic coverage and highlight the topics of interest for different populations regarding science. Comparing Colombia vs. the United States and Colombia vs. Spain in terms of science coverage in the media is valuable due to the cultural differences and distinctive characteristics that exist, which can help understand how cultural, social, linguistic, and economic factors may influence

The objective of this study is to analyze the content of science news articles from selected Colombian, American, and Spanish media outlets available in the Google News feed during a 15-day period in the first semester of 2023. Through the combined use of web scraping and data science techniques for natural language analysis, the extracted information is utilized to determine word frequency, aiming to identify patterns, differences, and similarities in the topics covered by the press in each country. The process of data preparation, the criteria for selecting news articles, and the methodology employed are described in Section 2. The results obtained and their respective analysis are presented in Section 3, and finally, the conclusions are presented in Section **??**.

2 Methodology

2.1 Data Source

The data consists of the content of news articles, including both the headline 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, and they 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 using the 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 top five most popular and eligible media outlets are: *El Tiempo*, *Semana*, *Pulzo*, *El Espectador*, and *Caracol*. For Spain[2], the selected and most widely read media outlets meeting the criteria are: *El País*, *El Mundo*, *La Vanguardia*, *ABC*, and *La Voz de Galicia*. Finally, for the United States[3], the chosen media outlets for the database are: *CNN*, *MSN*, *Fox News*, *The Washington Post*, and *Yahoo*. It is worth noting that while *The New York Times* meets all the mentioned conditions, its digital platform does not allow web scraping using the method considered in this study.

[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/

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 respective countries. This allows extracting the news articles classified under science for each country from 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, thus eliminating any potential misfiltered articles. The news articles were collected on a daily basis 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 headlines and "hook phrases" specific to each media outlet. Figure **??** shows a snippet of the content of a news article before and after the cleaning process.

Next, the *nltk* library is used to eliminate stop words (also known as empty words) specific to each language from the news article content. Additionally, a set of custom stop words created by the authors was also removed. Once the stop words are removed, the *Stanza* library is used to perform lemmatization, which involves transforming the inflected form of a word into its corresponding lemma. For example, "players" is changed to "player" and "ran" is replaced by "run".

After applying the aforementioned processes to each news article from the three countries, the frequency of occurrence of each word across the entire dataset is obtained for each country.

The 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.

To calculate the entropy, Shannon's entropy is employed in the following form:

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{1}$$

where $p(x)$ corresponds to the probability of each word, and the sum is taken over all words. The probability of each word can be estimated using the maximum likelihood method [13]:

$$p(x_i) = \frac{f_i}{\sum_{i=1}^{N} f_i} \tag{2}$$

where f_i is the frequency of occurrence of the word x_i, and the denominator is the sum of frequencies over a finite group of size N, i.e., the total number of elements.

3 Results and Analysis

The dataset of science news for Colombia, collected as mentioned above, consists of a total of 224 articles, with an average length of 507 words. On the other hand, for Spain, a total of 142 articles were collected, with an average size of 821 words. Finally, for the United States, 53 articles were obtained, with an average length of 709 words. The length of the articles was obtained before the process in which stop words are removed.

In general, based on the above information, it can be stated that Colombia is the country with the highest number of published 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 has the longest articles on average, followed by the United States, while Colombia has the shortest articles. More specifically, Colombia publishes 1.5 times more articles than Spain, however, Spanish articles are 1.6 times longer than Colombian ones.
On the other hand, Colombia publishes 4.2 times more news than the United States, but the length of the US news is 1.5 times larger than the Colombian ones. One noteworthy point in this comparison is that when filtering the Google News feed by the category "Science" and the country "United States," a large portion of the top news comes from specialized journals like "Nature" or science government institution portals like "NASA." However, these types of sources were not considered given the objective and focus of this work. This behavior was not observed in the case of Colombia.

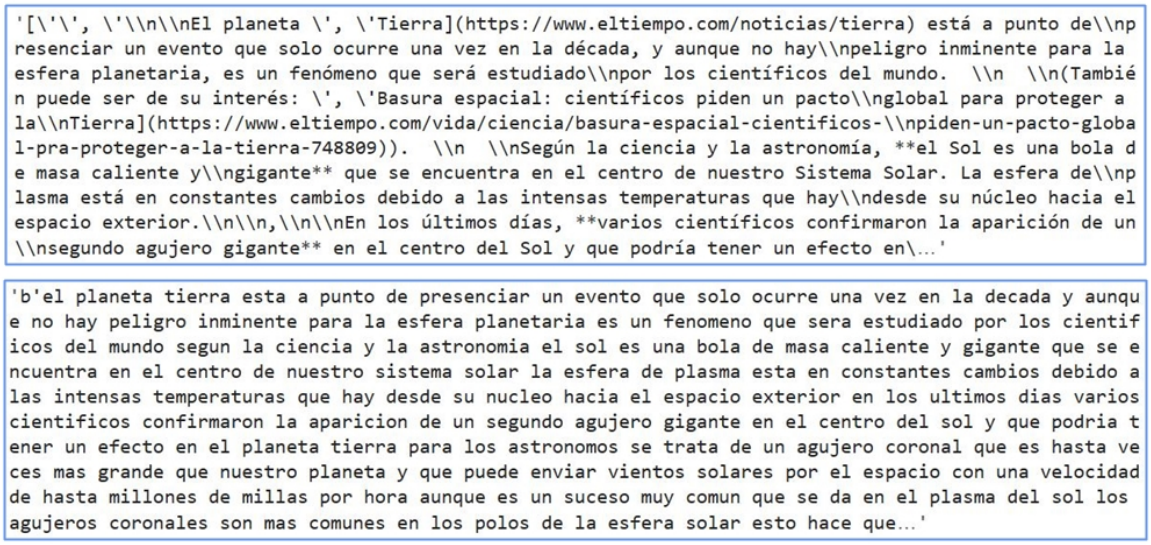

Figure 1: Example of the content of the news article *Nasa halla agujero gigante en el Sol: sus efectos se podrían ver desde la Tierra* published by the newspaper *El Tiempo* before and after text cleaning.

In Table 1, the first 10 most used words by country in the preprocessed science news are shown. Based on 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 "year" and "scientist." The content of the news in Colombia and Spain uses words to refer to individuals such as "researcher" and "scientist," while in the United States, these words do not appear, indicating a more impersonal nature. In all three countries, the word "new" appears in the ninth and eighth positions, respectively.

Colombia		Spain		EEUU	
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: TTop 10 most frequently used words in the science news articles along with the number of times they were used in all the articles considered, sorted by country.

The content of the news articles, in terms of vocabulary, cannot be analyzed by limiting it solely to the top 10 most used words, so the analysis was extended up to the 50th position. Figure 2 shows the number of times each word appeared in the group of science news articles per country, normalized by the highest frequency, for the top 50 most used words in the news articles about science per country. It is evident that all three graphs exhibit an exponential decay behavior.

In positions 10 to 20, both Colombian and American media maintain the trend of vocabulary related to astronomy, although for the United States, more diverse words start to appear. On the other hand, in the case of Spain, the words in these positions can be related to global terms that do not indicate a specific topic. However, it should be noted that Spain is the only country that mentions the word "mujer" (woman) among

the top 50 most frequent words in the news. In the range of positions 20 to 30, all countries show a greater diversity in vocabulary, with words that can be related to different fields of science (other than astronomy). For the United States, words such as "dinosaur," "ant," and "water" are found. In the case of Colombia, there are words like "agua" (water), "cerebro" (brain), or "planta" (plant), and for Spain, "matemático" (mathematician) or "célula" (cell). It is evident that biology can be related to 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 vocabulary oriented towards result analysis, with words like "resultado" (result), "dato" (data), or "mayor" (greater), while the United States returns to the astronomical theme, and Spain maintains a different focus, not centered on a specific topic, with words like "cambio" (change), "sociedad" (society), and "sistema" (system). Finally, in the interval from 40 to 50, astronomy-related terminology reappears in the news from the United States, with words like "solar" or "james" (referring to the recent "James Webb" telescope, with the surname "James" appearing in the ninth position). In the case of Colombia, the vocabulary includes words that cannot be solely related to science, such as "país" (country) or "mano" (hand). Finally, for the Spanish news, there is a presence of nouns that are easily related to scientific work, such as "físico" (physicist), "dato" (data), or "efecto" (effect).

The verb "explicar" (to explain) was the most used in both Spain and Colombia, while the second most used verb was "encontrar" (to find) for Colombia and "saber" (to know) for Spain. As for the United States, the two most used 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 used with high frequency in the news.

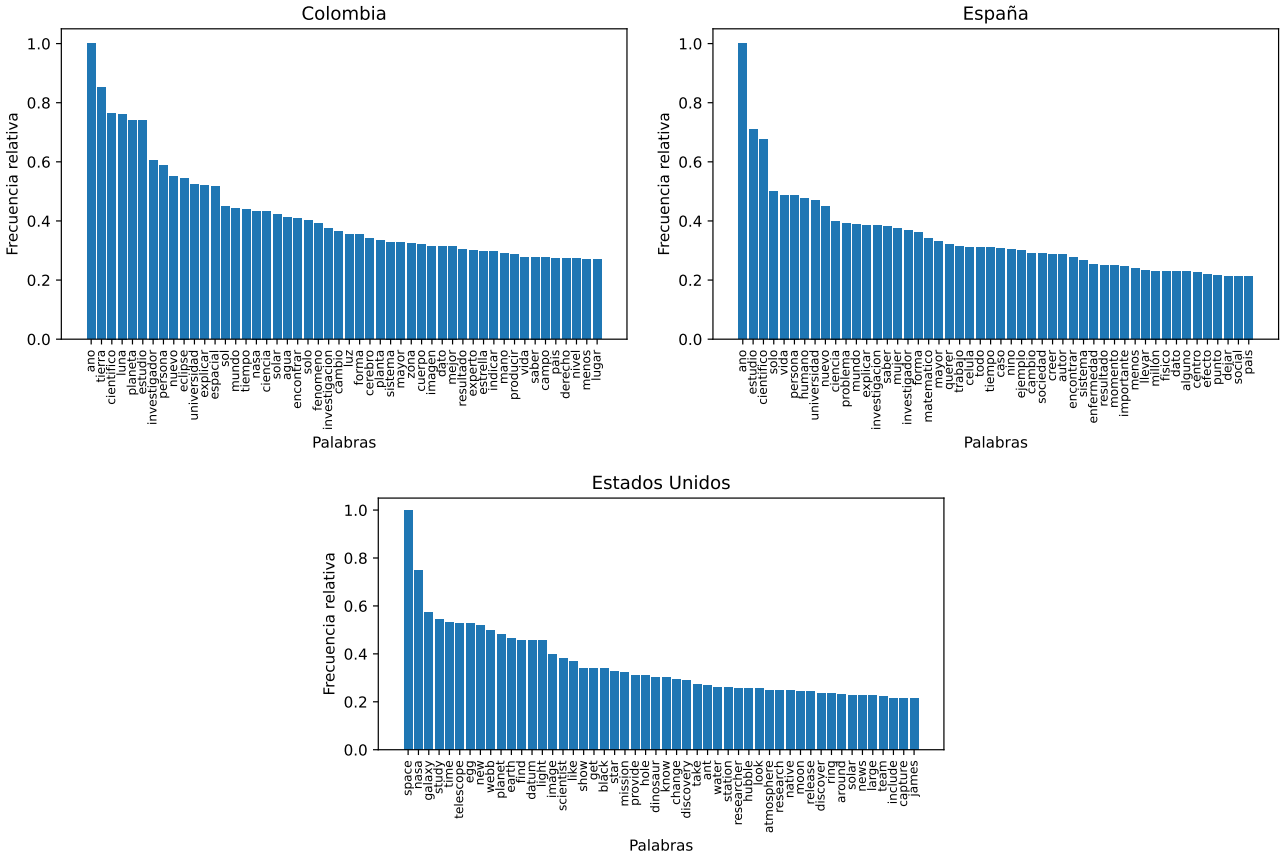

Figure 2: Relative frequency for the first 50 words per country. These words do not contain accents or "ñ" due to the cleaning process.

In Figure **??**, the total number of news articles published per day, along with the accumulated number of news articles, is shown for each country. It can be observed that Colombia generally publishes a higher number of news articles compared to Spain and the United States. On the 12th day, corresponding to April 9th, all three countries obtained a similar daily number of news articles. The behavior of the number of news articles per day is approximately constant. Similarly, the growth of the accumulated news articles is linear, although with different slopes depending on the country. This indicates that the considered media outlets do not distinguish between the quantity of news articles published on weekdays or weekends, at least not collectively.

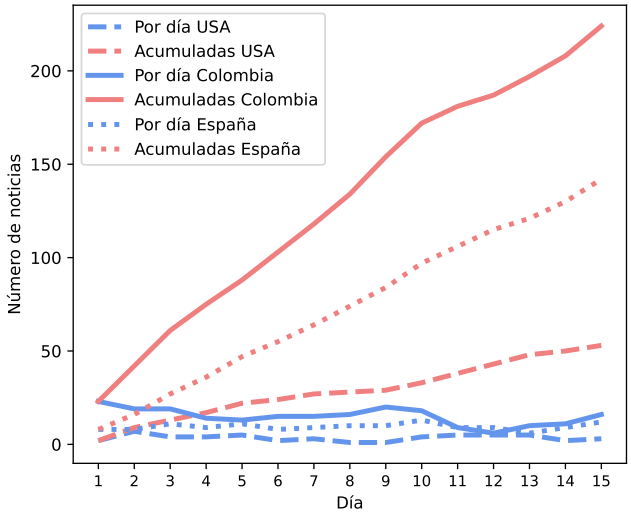

Figure 3: The number of news articles per day (blue) and the accumulated number of news articles (pink), for Colombia (solid line), Spain (dots), and the United States (dashed line).

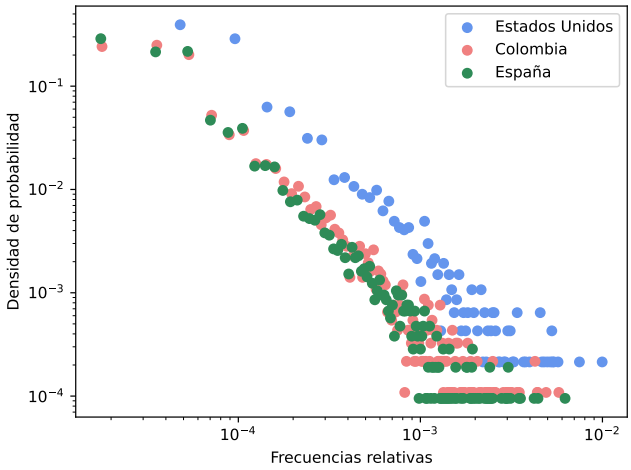

Figure 4: Distribution of the probability density of word occurrence frequency in Science news for the United States, Colombia, and Spain.

The graph 4 shows the distribution of the probability density of word occurrence frequency in science news for the United States, Colombia, and Spain. For all three countries, two regimes can be observed, divided by a frequency value of approximately 10^{-3}. For lower frequencies, a linear decay trend can be observed 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 an expected result since these countries share the same language. However, for the regime corresponding to frequencies greater than 10^{-3}, there is a highly dispersed

behavior compared to the previous one, where no linear trend can be identified.

The abnormal distribution of data in the second regime may be due to two fundamental factors. Firstly, during the cleaning process, the most frequently used words in both languages (articles, pronouns, and prepositions) were removed, which could have influenced the usual statistical behavior of language. Secondly, the cleaning process may contain imperfections due to the infinite possible variants that require a more thorough and detailed treatment. Thus, when reviewing words with very low frequencies, the presence of typos is evident, such as multiple words joined together (e.g., "cancerdiagnosis") or collections of meaningless letters, which originated from the content of the news or the source code of the webpage.

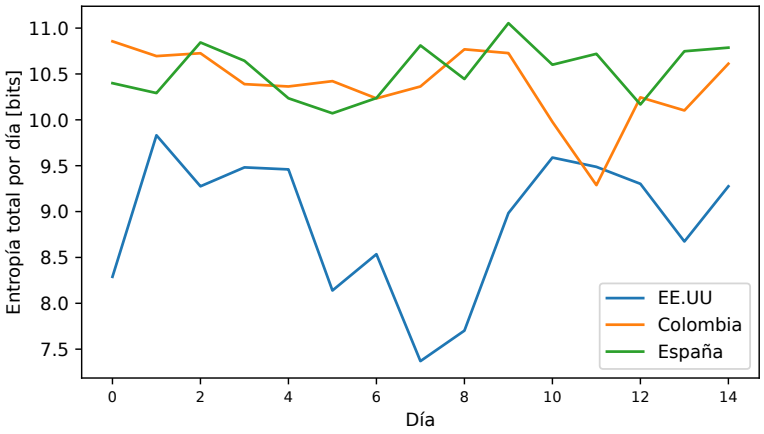

Figure 5: The entropy measured in bits for each day and country.

Finally, the entropy was calculated for the news collected per day and per country, and the values are shown in Figure 5. It can be observed that both Spain and Colombia exhibit a roughly constant behavior with an average 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 articles was recorded for this country. In general, it is noteworthy that both Colombian and Spanish media maintain a remarkable stability in the diversity of their news throughout the days. This pattern is intriguing, as one would expect variations in the quantity and diversity of news, especially between working days and non-working days.

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

4 Observations

During the preparation phase, while selecting the media to use, it was evident that the science section of several news portals most read by the Colombian population was either not updated or had a low publication frequency (once a month). For Spain, a large number of the most read media outlets are exclusively dedicated to sports and do not have a dedicated science section. On the other hand, as mentioned earlier, in the United States, the majority of publications come from specialized media.

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

Finally, it is important to mention that, due to the low volume of data, it cannot be stated with complete certainty that some variations in the results correspond to the system's behavior, as they may be due to inherent statistical noise.

5 Conclusions

- The frequency of words in the three countries follows an exponential decay pattern, which is a common characteristic in natural languages. Additionally, it was found that the three countries maintain a roughly 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 outlets is to achieve mass dissemination. In contrast, Spain recorded a lower daily number of science news compared to Colombia. However, it is important to highlight that the limited quantity of news is compensated by a greater depth and diversity in the topics covered. Although the absolute number may seem reduced, 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 articles 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 longer length compared to Colombian news articles.

  On the other hand, although it was observed that the entropy of news articles in Colombia is higher than that of news articles 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 and data science techniques in the collection and analysis of science-related news has provided the opportunity to systematically examine a considerable volume of information sources. This approach has allowed for the identification of both similarities and differences among the news articles, revealing patterns and trends that would be difficult to detect using conventional methods.

For future work, it is recommended to employ a larger database by extending the news collection period. This would allow for a more representative sample of scientific topics. Additionally, conducting a more thorough and refined data cleaning process is crucial to ensure 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 similarly 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] Maurıcio 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] 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.

[14] 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.