Characterizing Documents about Colombian Indigenous Peoples using Text Analytics

Jonathan Piñeros-Enciso and Ixent Galpin

Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano
Bogotá, Colombia
{jonathan.pinerose,ixent}@utadeo.edu.co

Abstract. The indigenous peoples of Colombia have a considerable social, political and cultural wealth. However, issues such as the decades-long armed conflict and drug trafficking have posed a significant threat to their survival. In this work, publically available documents on the Internet with information about two indigenous communities, the Awá and Inga people from the Cauca region in southern Colombia, are analyzed using automated text analytics approaches. A corpus is constructed comprising general characterization documents, media articles and sentences from the Constitutional Court. Topic analysis is carried out to identify the relevant themes in the corpus to characterize each community. Sentiment analysis carried out on the media articles indicates that the articles about the Inga tend to be more positive and objective than the Awá. This may be attributed to the significant impact that the armed conflict has had on the Awá in recent years, and the productive projects of the Inga. Furthermore, an approach for summarizing long, complex documents by means of timelines is illustrated with a sentence issued by the Constitutional Court. It is concluded that such an approach has significant potential to facilitate understanding of documents of this nature.

Keywords: Text Analytics · Sentiment Analysis · Indigenous Communities

1 Introduction

The indigenous populations of southern Colombia have historically suffered from drug trafficking and wars by illegal armed groups. These peoples live in a constant fight for their territory and their rights [14,20]. Although violence is a protagonist in their history, they are also characterized by their high resilience. Further, they stand out for their strong cultural component, their productive projects, and the great variety of fauna and flora they have in their areas of influence. Given the demographic and social particular characteristics of these populations, it is assumed that the information publicly available on the Internet is complete enough to carry out a robust text analytics process that allows the extraction of relevant and succinct information in an automated manner.

The first problem to face is the determination of the corpus. Documents on the Internet can be so diverse that the number of topics could be unmanageable. Therefore, the search was reduced to three categories, which are (1) characterization of the population, (2) journalistic documents, and (3) legal files. Each category was simplified because it continued to contain information with a high degree of heterogeneity. In this way, the characterization of the population was limited to official documents and/or issued by territorial entities, journalistic documents to the news in a range of dates and from specific media, and legal files to Constitutional Court sentences because these are of vital importance in the protection of human rights. The scope of the corpus is defined in Section 3 of this article.

Once the corpus has been determined, it is necessary to establish how the information will be approached. It was decided to go from the general to the specific. First, it was worked with the complete corpus, making a general analysis of the themes and establishing the diversity of the information collected. In Section 4 an analysis of the topics present in the entire corpus is carried out. This allows confirming if the oscillation between violent, cultural, and productive themes will be reflected in the extracted documents.

In Section 5, sentiment analysis is carried out on the texts of some emblematic Colombian media. The documents are evaluated and classified in positive or negative connotations as the case may be. The purpose of this process was to check if the characteristics of violence were evident and the force in which they were reflected in the collected material. Furthermore, an evaluation of the objectivity and subjectivity of these documents was carried out. This type of analysis can be useful when examining the media. This is important since indigenous peoples in Colombia and other countries around the world have repeatedly denounced false accusations by the media [7], lack of access and participation in the media [12], and even attacks against fundamental rights by the media[10].

Finally, in Section 6, a proprietary algorithm is presented that identifies date patterns (both exact dates and time intervals) and uses entity recognition techniques to determine events, actors, organizations, and actions. The result of this algorithm is a visualization that contains a timeline with a small summary of each event. This allows for understanding the sequence of events in a summarized and automated way. Since legal texts are usually long and difficult to understand, the operation of this algorithm is demonstrated using a Constitutional Court sentence as an example.

The main limitation of the present study is language. This is because the documents are in Spanish and the programming components used are more efficient for the English language. So the quality of the information can be lost in translation. On the other hand, this research does not address predictive analytics problems, as they are beyond the proposed scope. The research will only develop a descriptive stage, of classification, transformation, and presentation of the data.

2 State of the Art

For the present investigation, the existing state of the art was scanned from three different approaches.

2.1 Research work that performs text analytics on documents involving indigenous communities

Colquhoun *et al.* [11] present a study about the *Aboriginal and Torres Strait Islander* communities in Australia. This study aims to verify the correlation between maintaining cultural traditions and the happiness of the population. For this, they analyze qualitative data with the Leximancer[1]. text analytics software. This allows grouping topics and generate graph visualizations to generate the results.

On the other hand, Shah *et al.* [24] explore the challenges of sentiment analysis in the indigenous languages of India. This is due to the differences in the structures concerning languages such as English.

It is important to note that for indigenous communities in Colombia, no works related to sentiment analysis or text analytics were found. Furthermore, at a global level, there is very little research that addresses these issues for this type of population.

2.2 Sentiment analysis and news

In sentiment analysis, the broadest study is found when the corpus refers to social networks, especially Twitter [2]. Due to its microblog structure and the generation of opinions in real-time, it is suitable for this type of study, [21,23,1].

There are several works whose focus is to establish a methodology for the analysis of feelings through journalistic documents or news. Godbole *et al.* [13] explore positive or negative connotations in sports, weather, and stock market news. Raina et *al.* [22] emphasize the computational technique called Sentic Computing[3]. Balahur *et al.* [3] analyze different points of view that can be given when analyzing a news item.

On the other hand, there are practical cases also prioritizing the search for documents related to the news. Bautin *et al.* [5] address the problem of language, translations and their effectiveness in this process; Li *et al.* [17] perform financial sentiment analysis for the Hong Kong Stock Exchange based on economic news and indicators; Kaya *et al.* [15] focuses on political journalism in Turkey.

[1] https://info.leximancer.com/
[2] https://twitter.com/
[3] https://sentic.net/computing/

2.3 Timelines and summaries

The first works focus on the methodological and statistical basis of the generation processes of time entities and summary. Swan *et al.* [25] propose a statistical model to find temporal characteristics in a text . Mani *et al.* [18] propose an experimental model to determine the chronology of events and [26] generate a visualization of events in a timeline. Allan et al. generate an experimental methodology to manage a corpus and propose abstract structures. Chieu *et al.* [9] propose a framework that assumes important events within a text collection and orders them chronologically.

The most recent jobs take a corpus of news and search for timelines and summaries, often using *machine learning* techniques. However, they have a different approach depending on the case. Yan *et al.* [28] focus on finding the most relevant news under the *web mining* technique. Kessler *et al.* does not focus on events but exclusively on the normalization of temporal entities and [27] relies on news headlines as a fundamental element of the search. Chen *et al.* [8] focuses on time series and uses neural networks to generate more effective summaries.

3 Corpus determination

Two indigenous populations were chosen for the present study: Inga and Aw'a. Both communities are related to the Regional Indigenous Council of Cauca[4]. From each population, particular categories of information are prioritized, these are:

- Official documents and/or issued by territorial entities that characterize the population;
- The news generated by three emblematic media in Colombia (Semana magazine, and El Tiempo and El Espectador newspapers) in a range of five years (2015-2020);
- The sentences that the Constitutional Court generated for these communities in the last 30 years (1990-2020). The characterization documents indicate the demographic, social, and political factors of the populations.

3.1 Characterization

These documents were obtained with automatic extraction, which through an automated download process the PDF documents that met the search parameters were selected. This process consisted of searching for keywords, such as the name of the population and the word “characterization”. For this process, the wget[5] Python library was used.

[4] El Consejo Regional Indígena del Cauca (CRIC) it is an association of indigenous councils in Cauca, Colombia.

[5] https://pypi.org/project/wget/

Subsequently, a manual debugging of redundant or irrelevant documents was performed. In total, ten documents were obtained for each population in PDF format that contains the characterization information.

3.2 News

The news was obtained through a web scraping process. This process consists of directly searching Google using the Python google-search library[6], filtering by three parameters: the date range (2015-2020), the keywords (here the name of the community and the word "indigenous" has been used), and the information sources that correspond to the URLs of the aforementioned media. In this way, only the text corresponding to each news is obtained and they are saved in files .txt. As a result, 133 documents related to the Aw'a community and 139 to the Inga community were obtained, distributed as shown in Table 1.

Table 1: News Corpus

Indigenous community	Source	Number of documents	Number of words	Number of characters
Awá	El Espectador	100	78150	496754
Awá	El Tiempo	3	2140	13362
Awá	Revista Semana	30	38725	239875
Inga	El Espectador	79	80587	511989
Inga	El Tiempo	41	35111	217731
Inga	Revista Semana	19	19559	122689

3.3 Constitutional Court Sentences

To obtain this corpus, it was necessary to do a manual search on the website of the Constitutional Court of Colombia. The page allows searching by year and by keyword. In this way, ten documents were obtained from each town.

4 Corpus General analysis

For the general analysis of the corpus, the characterization documents, news, and sentences were loaded in a single DataFrame[8]. One script was run for the PDF documents and another for the .txt files. The text was encoded in the

[6] https://pypi.org/project/google-search/
[7] http://www.corteconstitucional.gov.co
[8] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html

UTF-8 format and special characters that could interfere in the information analysis process were removed.

After extracting the text from the documents, *tokenization* process was carried out. This process consists of extracting the words from the text in different types of entities (the *tokens*). These tokens are exempt from punctuation marks and *stopwords*. The latter refers to a set of the most common words in the language that, due to their high repetition, are not important [19]. As a next step, *bigrams* and *trigrams* were created, which are pairs and trios of words respectively, that appear consecutively on several occasions, which is why it is determined that they belong to the same context[19].

The last process carried out in the text is the *Lemmatization*. This is a "process by which the words of a text that belong to the same inflectional or derivative paradigm are taken to a normal form that represents the whole class" [4].. With this, words can be grouped according to their meaning even if they have different forms or conjugations. With the information *lemmatized*, the document-terms matrix is created. With the stemmed information, the document-terms matrix is created. The matrix indicates the number of times the words are repeated. Finally, this is the first way to measure the importance of words [19].

To determine information topics, the *Latent Dirichlet Allocation* (LDA) algorithm was used. This "is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar" [6].

To solve the problem of determining the optimal amount of topics to evaluate the model, the *LdaMallet*[9] library was used. Consequently, the model is executed with different parameters, and the one with the highest coherence range is chosen. This determines the number of themes that will be used. Figure 1 shows the coherence of the model against the number of subjects evaluated.

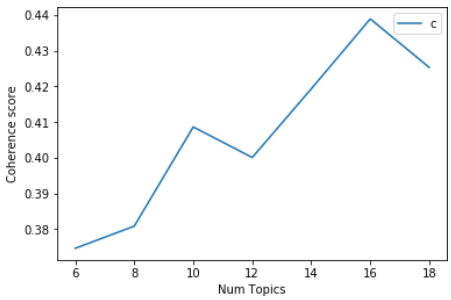


Fig. 1: Coherence of the model vs Number of topics

The highest coherence is found at number 16 with a coherence coefficient of 0.439. However, and after evaluating models of 16 topics, it is possible to perceive the redundancy of information and several topics without

[9] http://mallet.cs.umass.edu/

representation within the documents. Therefore, the experimentation showed that the ideal number is close to 10. Figure 2 shows the distribution of the topics concerning the information extracted and gives a first idea of the information present in each topic. Table 2 shows the number of documents per topic.

Table 2: Documents by topic

Topic	Number of documents
9	79
2	55
7	33
5	25
6	21
1	20
4	18
0	18
3	17
8	16

When analyzing the word cloud in Figure 2 and several of the documents by topic, the following logical distribution by topic is concluded as shown in Table 3. This selection of generalities is the summary of a study of each article classified by topic. The identification of topics allows finding similarities between documents, review the variety of topics that are addressed in the texts, and get a general idea of the information contained in the corpus.

5 Sentiment Analysis on News

For this section, it is intended to determine through the automatic reading of the news corpus and subsequent sentiment analysis, if positive or negative aspects of these indigenous communities were reflected in the media. For an initial reading, word clouds were generated, which allows us to identify the themes that appear repeatedly and give us a first idea of what the news is talking about. Figure 3a shows the word cloud generated for the Inga community, and Figure 3b the corresponding one for the Aw'a.

Similarities are seen between the word clouds in Figure 3. For example, the terms *indígena* (*indigenous), comunidades (communities), territorio (territory)*, and pueblos (*peoples)* appear in both clouds, as each one highlights the name of its community. However, in the differences, it can be noted that the Inga community presents words with a positive connotation such as, for example, *vida* (*life), Universidad (university), paz (peace)*. On the other hand, the Aw'a community has words like *armados* (*armed), víctimas (victims), FARC (Revolutionary Armed Forces of Colombia)*, which in principle would indicate that the texts referring to the Aw'a have more negative aspects than those of the Inga.

Fig. 2: Total corpus word cloud

Table 3: Documents by topic

Topic	Generality of topics
0	Territory, demographic aspects, economic activities.
1	Science, university professionals and different academics fields
2	Productive projects and new ways of producing in the face of adversity
3	Violence, armed groups, recruitment of minors, threats
4	Territorial disputes, claim for environmental rights, indiscriminate mining
5	Sentences that protect indigenous peoples, their subsistence and ethnic and/or cultural identity
6	Everything related to the land, either as a territory and its geographical characteristics or agricultural, fauna and flora themes.
7	Characterization of the peoples and geographical aspects
8	Social aspects, religions, festivals and rituals
9	Oral Tradition, Music, festivals, peace initiatives and cultural activities.

(a) Words cloud news Awá (b) Words cloud news Inga

Fig. 3: Top topics news indigenous communities

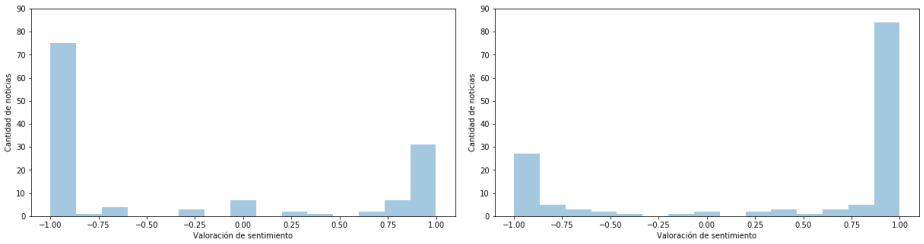
To determine if the news covers a positive or negative context, the Nltk.sentiment[10] package was used. This program allows assigning a numerical value to the general sentiment of each text, being -1 the maximum negative, and 1 the maximum positive. Since this package currently only works for the English language, the translation of the text was previously carried out with Google

[10] https://www.nltk.org/api/nltk.sentiment.html#nltk-sentiment-package

Translate API[11] Table 4 confirms what is shown in the word clouds. In the case of the Aw'a, most of the documents are negative, a trend that is the opposite in the case of the Inga.

Table 4: Sentiment evaluation news

	Positives	**Negatives**
Awá	54	79
Inga	100	39



(a) Awá community (b) Inga community

Fig. 4: Sentiment distribution news

An additional analysis was carried out comparing the distributions in Figure 4. Once again, it is observed that the documents of the Inga community have a more preponderant positive connotation than those of the Aw´a community that have a greater tendency to a negative aspect.

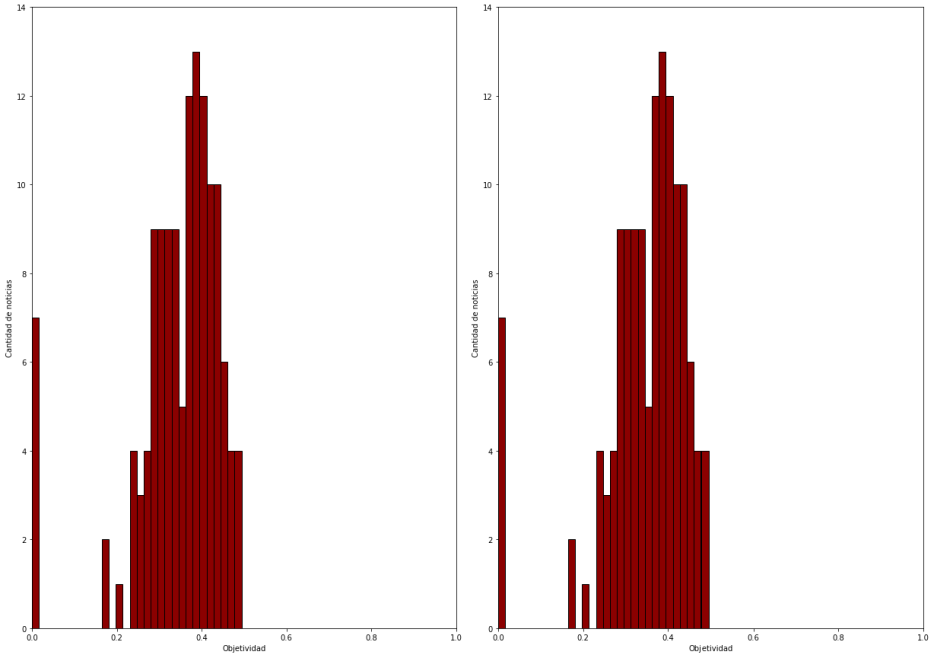
On the other hand, the objectivity of the news corpus is analyzed, to determine if the narrative is closer to an opinion or a fact. For this purpose, an objectivity score was calculated using the Textblob[12] library.

Values closer to 0 indicate high subjectivity in the text, and values closer to 1 high objectivity. Figure 5 shows the objectivity values obtained for the Inga and Aw'a population. It is observed that most of the news for both populations is between 0 and 0.5, which indicates that there is a tendency towards subjectivity.

Finally, it is analyzed if the feeling is correlated with subjectivity. Figure 6 shows the relationship between the two variables for each of the populations under study. According to the charts, for the Aw'a people, the Pearson correlation coefficient is 0.88, and for the Inga people, it is 0.87.

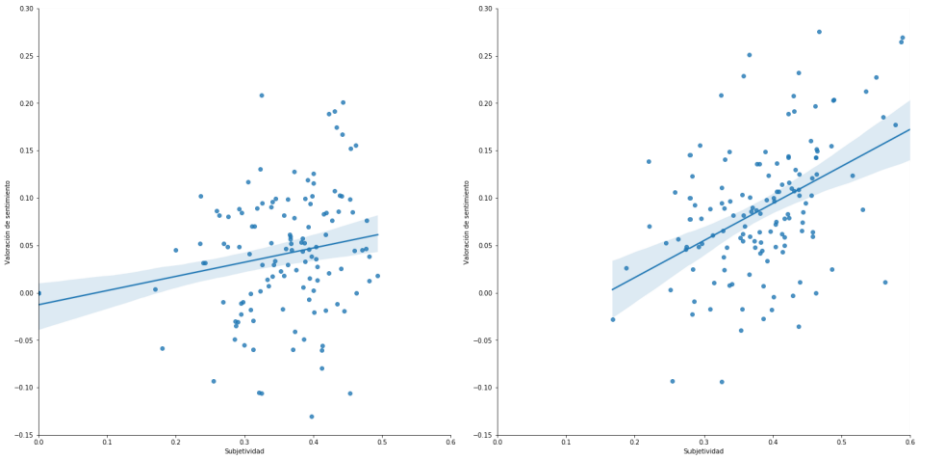
[11] https://cloud.google.com/translate/docs/basic/translating-text#translate translate text-python

[12] https://textblob.readthedocs.io/en/dev/



(a) Awá People (b) Inga People

Fig. 5: Distribution of news objectivity



(a) Awá People (b) Inga People

Fig. 6: Correlation objectivity feeling

This indicates that they are highly correlated variables in both cases.

Consequently, it can be concluded that the more positive a new is, the closer is to a fact, and the more negative it is, the closer to an opinion is.

The work carried out on the news indicates that in the period studied, the Aw'a community presents news with a negative context. This may be because it is highly affected by armed groups and the plantation of illicit crops. On the other hand, for the Inga people, the media have highlighted their productive, cultural, and artistic projects, causing that the news tends more to a positive aspect. However, it was found that the news from Colombian media, for the peoples' understudy, tends to be subjective and this subjectivity presents a direct correlation with the sentiment analyzed.

6 Automatic generation of Court Sentences timelines

Using the corpus of sentences of the Constitutional Court, it is intended to extract the important events within a sentence. For this purpose, the text of the PDF Sentence T-630/16[13] is extracted and debugged. This sentence deals with a legal dispute between the indigenous communities and the oil company Gran Tierra Energy Colombia LTD, which seeks to protect the right to prior consultation. It was chosen due to the complexity of the case and the length of the document. Important events are determined by the time in which they happened. However, since the nomenclature of a date can be presented in various ways in a text, it is necessary to identify various date patterns and time intervals, which is done with a self-made Python algorithm, described below.

To find a date written in a non-traditional format, such as: "the early morning of January 30, 2020" or "by mid-2015" the first thing that is searched in the text are years, between 1900 and 2099 that have the peculiarity that they only present four digits. Once the year has been found, in a range of next and previous characters the name of a month or a possible abbreviation is searched, for instance, “January” or “Jan”. Finally, a digit with a maximum of two digits is sought that indicates a particular day, its respective written form or its ordinal representation, for example, "1", "one" or "first". The text may refer to a time range (that is, a range of dates) rather than a single date. The procedure to determine an interval is as follows:

- If the text mentions only one year, for example: “a tragedy happened in 2018”, the date range is the entire year in mention from January 1 to December 31.
- If the text contains a year accompanied by the month, for example: "The letter was signed in January 2017", the range of dates is all that comprises said month in question, in this case, from January 1 to January 31 from 2017

[13] https://www.corteconstitucional.gov.co/relatoria/2016/t-630-16.htm

- If the text contains the full date, for example: “Everything started on January 15, 2019” the date is complete and the range is a whole day. In this case, the interval is from 00:00 hours on January 15 to 11:59 on the same day.

Once the date is obtained, it is proceeded to try to identify the event that happened at that time. To do this, the entire sentence that is enclosed in that text is captured. With the help of the spaCy[14] library, the analysis of the grammatical components and recognition of entities of the sentence is carried out to identify the verbs, the subjects, and the entities that the sentence has. By extracting an example sentence we can make a visual analysis of the entities recognized by Spacy as shown in Figure 7.

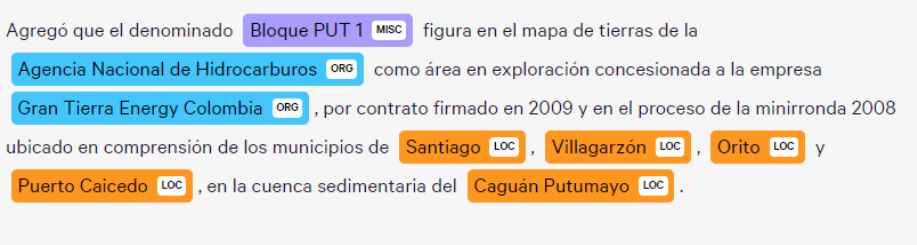


Fig. 7: Entity recognition example

The entities detected are the following:[15]

- LOC: Places, mountain ranges, bodies of water
- ORG: Designated as a corporate, governmental, or other organization entity
- MISC: Various entities, for example, events, nationalities, products or works of art

Regarding the grammatical composition, using another example sentence, the distribution presented in Figure 8 is generated.

The characteristics detected are the following:[16]

- NOUN: Subjects.
- VERB: Actions.

Finally, the sentence that is obtained is generally very long, so each sentence is summarized to visualize the data.

A function is created in Python that allows to make a summary of a paragraph of the text, starting from the original text. This function generates a frequency matrix of terms, excluding stopwords. Subsequently,

[14] https://spacy.io/
[15] https://spacy.io/api/annotation
[16] https://spacy.io/usage/linguistic-features

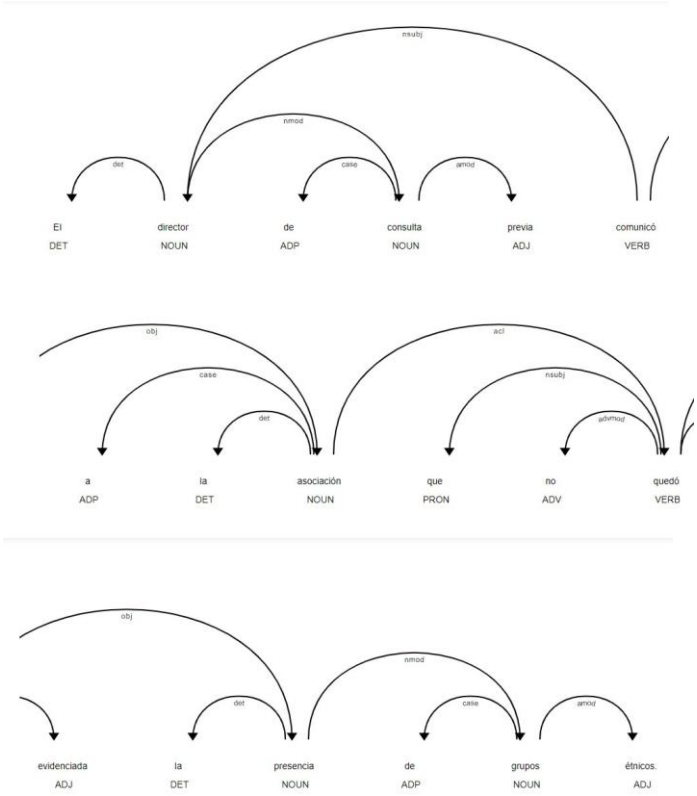


Fig. 8: Example grammatical composition

a percentage weight is assigned to each word and thus each sentence of the text receives a value depending on the words they have. Phrases that exceed a higher weighted than 0.5 of the calculated will be taken into the summary, otherwhise will be ignored. Finally, the information collected was the one shown in Table 5.

Table 5: Data collected Time line

Data	Datatype
Event	Int
Initial date	Datetime
Final date	Datetime
Text	String
Entities	String
Actions	String
Nouns	String
Summary	String

Fig. 9: Time line 1

Fig. 10: Time line 2

Figures 9 and 10 show part of the timeline resulting from events over time. Events automatically extracted from the sentence are reflected in the timeline. These were ordered by time and the summary assigned to that instant of time was used as a label. About 100 events were captured, but for practical purposes, only the first 10 events are plotted in chronological order from smallest to largest.

7 Conclusions

This article has explored the potential of text mining to characterize various aspects of texts related to two indigenous populations in southern Colombia. Through the analysis of topics carried out on the corpus, it was observed that these have a significant diversity of themes due to their cultural wealth and their processes of fight and resistance. Although it is evident that violence caused by factors such as the armed conflict is an issue that affects indigenous people significantly, their productive projects and culture are also topics that stand out. It can also be inferred that through the text analytics techniques used in this article, populations similar to indigenous ones can be studied.

Through the sentiment analysis and objectivity, the impact of violence in different populations can be compared. The negative connotations indicate the significant presence of violence in the narrative of the texts analyzed. The positive connotations indicate productive projects, zones of peace, and cultural manifestations. The analysis can also be used to investigate thoroughly media coverage of indigenous peoples.

The generation of events within documents could frame a guideline to achieve effective summaries of legal documents. Time entities would be able to establish a logical sequence of events within a text concerning time, unlike conventional analytical summaries. This is framed within a temporal process, which would allow us to change the order of the narrative and generate new data framed within a timeline.

Text mining for the Spanish language has a long way to go. The existing analytics solutions and those used in this study are well established in the English version. However, for Spanish, it is necessary to use either to a translation, which implies a considerable loss of information or to libraries that offer Multilanguage support such as Spacy. However, it should be noted that these libraries work in a reduced way for Spanish compared to their original version in English

.

Future research can be framed from two points. On the one hand, predictive analysis can be performed for the journalistic line corpus. This attempts, through the analysis of timelines, to predict the sentiment trend that could occur with respect to a corpus classified by dates. In this way, it can predict whether the next news tends to be positive or negative. Consequently, human rights early warning can be proactively issued for the protection of these communities. On the other hand, it is possible to consider incorporating machine learning techniques to improve the search algorithm for events based on temporal events. Likewise, since the practice was clearly focused on decisions of the Constitutional Court, it can be focused on other types of long documents that require this type of summary by schedule. Exploring interactive visualizations for timelines also has the potential to be a promising research direction.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.J.: Sentiment analysis of twitter data. In: Proceedings of the workshop on language in social media (LSM 2011). pp. 30–38 (2011)
2. Allan, J., Gupta, R., Khandelwal, V.: Temporal summaries of new topics. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 10–18 (2001)
3. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment analysis in the news. arXiv preprint arXiv:1309.6202 (2013)
4. Bassi A., A.: Lematizacion basada en análisis no supervisado de corpus (aug 2020), https://users.dcc.uchile.cl/~abassi/ecos/lema.html
5. Bautin, M., Vijayarenu, L., Skiena, S.: International sentiment analysis for news and blogs. In: ICWSM (2008)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
7. del Cauca, C.R.I.: Denuncia pública sobre las falsas acusaciones hacia la minga por parte de los medios de comunicación locales, regionales y nacionales (March 2019), https://www.cric-colombia.org/portal/denuncia-publica-sobre- las-falsas-acusaciones-hacia-la-minga-por-parte-de-los-medios-de- comunicacion-locales-regionales-y-nacionales/
8. Chen, X., Chan, Z., Gao, S., Yu, M.H., Zhao, D., Yan, R.: Learning towards abstractive timeline summarization. In: IJCAI. pp. 4939–4945 (2019)
9. Chieu, H.L., Lee, Y.K.: Query based event extraction along a timeline. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 425–432 (2004)
10. de Colombia, C.C.: Sentencia t-500/16 acción de tutela contra medios de comunicación (2016), https://www.corteconstitucional.gov.co/relatoria/2016/t-500-16.htm
11. Colquhoun, S., Dockery, A.M.: The link between indigenous culture and wellbeing: Qualitative evidence for australian aboriginal peoples (2012)
12. Doyle, M.M.: Acceso y participación de los pueblos indígenas en el sistema de medios de argentina. Disertaciones: Anuario electrónico de estudios en Comunicacion Social **11**(2), 30–49 (2018)
13. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. Icwsm **7**(21), 219–222 (2007)
14. Hernández Delgado, E.: La resistencia civil de los indígenas del cauca. Papel político **11**(1), 177–220 (2006)
15. Kaya, M., Fidan, G., Toroslu, I.H.: Sentiment analysis of turkish political news. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. vol. 1, pp. 174–180. IEEE (2012)
16. Kessler, R., Tannier, X., Hagege, C., Moriceau, V., Bittar, A.: Finding salient dates for building thematic timelines. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 730–739 (2012)
17. Li, X., Xie, H., Chen, L., Wang, J., Deng, X.: News impact on stock price return via sentiment analysis. Knowledge-Based Systems **69**, 14–23 (2014)
18. Mani, I., Wilson, G.: Robust temporal processing of news. In: Proceedings of the 38th annual meeting of the association for computational linguistics. pp. 69–76 (2000)

19. Manning, C.D., Schütze, H., Raghavan, P.: Introduction to information retrieval. Cambridge university press (2008)
20. Ortega, M.C.R.: El cauca lleva más de medio siglo azotado por la violencia: así es el territorio colombiano de las dos masacres esta semana (11 2019), https://cnnespanol.cnn.com/2019/11/01/cauca-masacres-indigenas-colombia-violencia-conflicto-armado/
21. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. vol. 10, pp. 1320–1326 (2010)
22. Raina, P.: Sentiment analysis in news articles using sentic computing. In: 2013 IEEE 13th International Conference on Data Mining Workshops. pp. 959–962. IEEE (2013)
23. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: International semantic web conference. pp. 508–524. Springer (2012)
24. Shah, S.R., Kaushik, A.: Sentiment analysis on indian indigenous languages: a review on multilingual opinion mining. arXiv preprint arXiv:1911.12848 (2019)
25. Swan, R., Allan, J.: Automatic generation of overview timelines. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 49–56 (2000)
26. Swan, R., Allan, J.: Timemine (demonstration session) visualizing automatically constructed timelines. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. p. 393 (2000)
27. Tran, G., Alrifai, M., Herder, E.: Timeline summarization from relevant headlines. In: European Conference on Information Retrieval. pp. 245–256. Springer (2015)
28. Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., Zhang, Y.: Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 745–754 (2011)