



## Identificación de temas del blog de noticias en idioma español

### Topic identification from news blog in Spanish language

#### Autores

✉ **Lizbeth Pacheco-Guevara**



✉ **\*Ruth Reátegui**



✉ **Priscila Valdiviezo-Díaz**



<sup>1</sup> Universidad Técnica Particular de Loja,  
Loja, Ecuador.

\* Autor para correspondencia

#### Resumen

Actualmente existe una gran cantidad de noticias en formato digital que necesitan ser clasificadas o etiquetadas automáticamente según su contenido. La Asignación Latente de Dirichlet (LDA por sus siglas en inglés) es una técnica no supervisada que crea automáticamente tópicos a partir de las palabras de los documentos. El presente trabajo tiene como objetivo aplicar LDA para analizar y extraer temas de noticias digitales en español. Se recogieron un total de 198 noticias digitales de un blog de noticias universitario. Se realizó un preprocesamiento de los datos y su representación en espacios vectoriales y se seleccionaron los valores de  $k$  en base a la métrica de coherencia. Con una matriz frecuencia de términos – frecuencia inversa del documento (TF\_IDF por sus siglas en inglés) y una combinación de unigramas y bigramas se obtienen temas con una variedad de términos y tópicos relacionados con actividades universitarias como programas de estudio, investigación, proyectos de innovación y responsabilidad social. Además, con el proceso de validación manual, los términos de los temas se corresponden con los hashtags escritos por los profesionales de la comunicación.

**Palabras clave:** LDA, modelado de tópicos, noticias, blog.

**Comó citar el artículo:** Pacheco-Guevara, L., Reátegui, R. y Valdiviezo-Díaz, P. (2022). Topic identification from news blog in Spanish language. *Informática y Sistemas: Revista de Tecnologías de la Informática y las Comunicaciones*. 6, (1), 31–37. DOI:<https://doi.org/10.33936/isrtic.v6i1.4514>

Enviado: 18/03/2022  
Aceptado: 24/04/2022  
Publicado: 27/05/2022

#### Abstract

Currently, exist a large amount of news in a digital format needs to be classified or labeled automatically according to its content. Latent Dirichlet Allocation (LDA) is an unsupervised technique that automatically creates topics based on words in documents. The present work aims to apply LDA in order to analyze and extract topics from digital news in the Spanish language. A total of 198 digital news were collected from a university news blog. A data pre-processing and representation in vector spaces were carried out and  $k$  values were selected based on the coherence metric. A term frequency – inverse document frequency (TF\_IDF) matrix and a combination of unigrams and bigrams produce topics with a variety of terms and topics related to university activities like study programs, research, projects for innovation, and social responsibility. Furthermore, with the manual validation process, terms in topics correspond with hashtags written by the communication professionals.

**Keywords:** LDA; topic modeling; news, blog.



## 1. Introduction

Currently it exists a large amount of news in a digital format. Text mining and natural language processing (NLP) allow us for transforming and analyzing unstructured information such as digital news. Topic modeling extracts the main topics from a collection of documents; this task can be done with the Latent Dirichlet Allocation (LDA) model.

LDA is a generative probabilistic model where documents are represented as random mixture over latent topics, and each topic is represented by a distribution of words (Blei, Ng, & Jordan, 2003). LDA is based on three levels: words, topics and documents.

In order to identify topics from news, some researches have been done. In (Xu, Meng, Chen, Qiu, Wang, & Yao, 2019) the authors proposed a method to analyze the evolution of news topics over time. They applied LDA method to extract topics from news, with LDA they obtained more reliable results than K-means clustering. Moreover, Misztal-Radecka (2018) worked in create user profiles based on news articles. This author used LDA and Word2Vec to represent information and demonstrated that Word2Vec works better in comparing small text like titles. However, for a classification task, a combination of topic identification and word embedding had a better performance.

Furthermore, in (Larsen & Thorsrud, 2019) the authors applied LDA models to analyze the role of business newspaper in predicting and explaining economic fluctuations. Moreover, Oliveira Capela & Ramirez-Marquez (2019) proposed a method to detect and to compare identities of cities using news articles from thirty-six cities from the United States of America. They used LDA to group words in topics. Also, Xu, Yeo, Hwang & Kim (2020) used LDA and network analysis to explore social media discourse on hyper-connected society base on news articles (Xu, Yeo, Hwang & Kim, 2020).

In addition, in (Guangce & Lei, 2021) the authors used LDA and the Apriori algorithm to improve media monitoring reports based on news articles. LDA allowed to identify topics, meanwhile Apriori allowed to discover relationships of topics of words. A recent work, focused on COVID-19, applied LDA in traditional and social media news to characterize and to compare early coverage of COVID-19 (Chipidza, Akbaripourdibazar, Gwanzura & Gatto, 2021).

Considering the importance of topics identification from unstructured documents, the present work aims to automatically classify and to assign topics to a corpus of news obtained from a university in Ecuador.

This paper is organized as follows: the section 2 presents the methodology for topic modeling; section 3 presents results and discussion, and finally, conclusions in section 4.

## 2. Methodology

In the reviewed literature, it was not found a predetermined methodology for topic modeling; in general, it is defined according to the type of data to analyze; however, some important steps mentioned in some related works were taken. Figure 1 shows five important steps mentioned in some related works to obtain a LDA model.

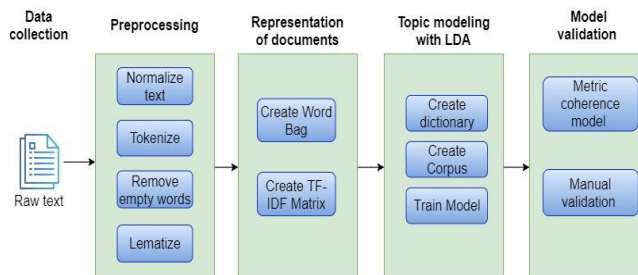


Figure 1. Flow for obtaining a LDA model.

Source: Authors

### 2.1. Data collection

In this step the data is gathering. In our case the data consists of digital news obtained from a news blog of a university in Ecuador. The university considered as a case of study has a news blog with updated information on the most relevant events of different academic activities that take place inside and outside of the institution. The 200 links were collected from October 2019 to May 2020, which are initially stored in a xml format. Next, the links were analyzed to verify they are working correctly and to eliminate duplicate information. Finally, 198 news were stored in a txt format to be analyzed by LDA. Figure 2 shows the news's HTML structure, it consists of a title and several paragraphs that make up the body of the news. It was verified that all news in the corpus have the HTML structure.



Figure 2. General structure of the news blog.

Source: Authors

## 2.2. Preprocessing

The preprocessing ensures data quality that influences LDA performance. Figure 3 shows the steps following in the preprocessing task: text normalization, tokenization, elimination of stop words and lemmatization.

**Text normalization:** It consists of removing numbers, alphanumeric values, special characters and punctuation marks (\*, /, %, \$, #, “”, @, °), among others. Also, in our cases letters were transformed from uppercase to lowercase and dates were eliminated. Denny and Spirling (2017) mention that numbers have relevance according to the domain of analysis. In this study numbers are not relevant because the objective is to identify topics that could show the main activities of the university in its different departments. Also, words in topics could help to select hashtags for each news, and as we can see in our dataset number are not used as hashtags, unless they are a compound word like COVID-19.

**Tokenization:** This step is used to separate words contained in each news.

**Elimination of Stop Words:** This step is used to select the most important words in the news. We used the list of stop words defined in Python spaCy library for the Spanish language.

**Lemmatization:** Our corpus is in the Spanish language; therefore, when processing text in Spanish it is necessary to rely on libraries with pre-trained statistical models for this language. In this work, the es\_core\_news\_sm model was used, which basically performs multitask work in Spanish trained in UD Spanish AnCor and WikiNER. It assigns token vectors, POS tags, dependency parsing, and context-specific named entities (Kim, Seo, Cho, & Kang, 2019). In addition, we used the token.lemma tag to obtain the word in its base form and the token\_pos tag to return the word lemmatized by nouns, verbs, and adjectives.

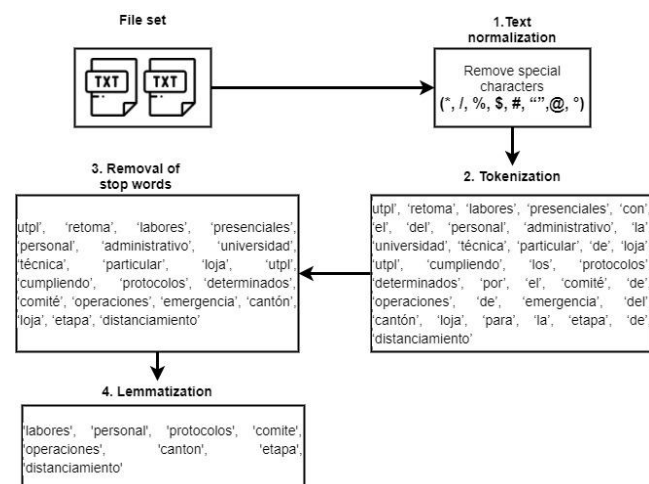


Figure 3. Preprocessing steps for Spanish news documents.  
Source: Authors

## 2.3. Document representation

Additionally, Bags of words and Inverse document frequency matrix are two of the most widely used methods to represent documents with matrices.

**Bag of words (BoW)** counts the occurrence of each word and annotates them into a vector. One of the first tasks for the preparation of the model is the vectorization of the corpus.

The Gensim function called doc2bow, document to document representation, was used to transform words within news into a tuple for numerical representation (Oliveira Capela & Ramirez-Marquez, 2019). To create the dictionary, we applied the filters and default values no\_below = 5, which keeps the tokens contained in at least 5 documents, and no\_above = 0.5, which keeps the tokens above 0.5 of the total size of the corpus. Finally, the dictionary consists of 1861 words.

**Inverse document frequency matrix (TF-IDF)** that gives each word a weight according to the frequency in a document (TF), and the rarity of its occurrence in the corpus (IDF). In our case, terms with a scarcity greater than 0.99 and terms with a very low frequency were eliminated. Through inverse term-frequency analysis of documents, the most relevant terms within the corpus were identified.

Figure 4 shows an example of the result of the two matrices.

| BoW                   | TF-IDF                  |
|-----------------------|-------------------------|
| [[ 'actividades', 6], | [[ 'actividades', 0.2], |
| ['afecciones', 1],    | ['afecciones', 0.14],   |
| ['asistencia', 1],    | ['asistencia', 0.11],   |
| ['cambio', 1],        | ['cambio', 0.07],       |
| ['campus', 2],        | ['campus', 0.17],       |
| ['canales', 1]]       | ['canales', 0.11]]      |

Figure 4. BoW and TF-IDF representation  
Source: Authors

## 2.4. Topic Modelling

The topic modeling step with LDA requires inputs for specific hyper-parameters such as alpha( $\alpha$ ), beta( $\beta$ ), and the number of topics ( $k$ ). Alpha refers to the distribution of documents per theme and beta refers to the distribution of words per topic (Blei, Ng, & Jordan, 2003, Xu, Meng, Chen, Qiu, Wang, & Yao, 2019).

An exhaustive search for the number of topics  $k$  was performed and the hyperparameter  $\alpha$  was optimized for the LDA algorithm. The number of topics was explored from  $k = 2$  to  $k = 40$ .

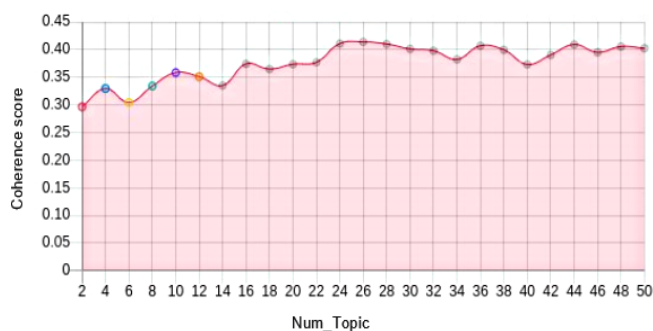
## 2.5. Model Validation

The main challenge when modeling topics with LDA is to define the appropriate number of topics ( $k$ ) to represent the whole corpus and the interpretability of the topics (Misztal-Radecka, 2018, Wang, Feng & Dai, 2018). Therefore, the coherence metric developed by Mimno & Wallach (2011) was applied to validate

the tuples  $(k, \alpha)$ .

The coherence metric takes into account the co-occurrence statistics of words collected from the corpus. It calculates the benchmark score in a range from 0 to 1, higher values represent better subject quality (Oliveira Capela & Ramirez-Marquez, 2019).

Alpha values distributed between 0 and 1 were tested. A text is considered coherent if all or most of the main words of the text are related to each other (Buenano-Fernandez, Gonzalez, Gil & Lujan-Mora, 2020). Figure 5 shows the results of the coherence metric for different numbers of topics. According to the behavior of the coherence values, the maximum value occurs for  $k = 10$ ,  $\alpha = 0.01$  and  $\beta = 0.1$  with a coherence of 0.35, the model was carried out with these parameters.



**Figure 5.** Hyper-parameter optimization for the number of topics  $k$  and LDA parameter  $\alpha$ .

Source: Authors

Additionally, most of the news has some hashtags written by communication professionals. Considering these hashtags, we made a manual validation of the result through a comparison between the hashtags and terms of the topic assigned to each news. In the next section, we show the topics identified and the comparison as a manual validation.

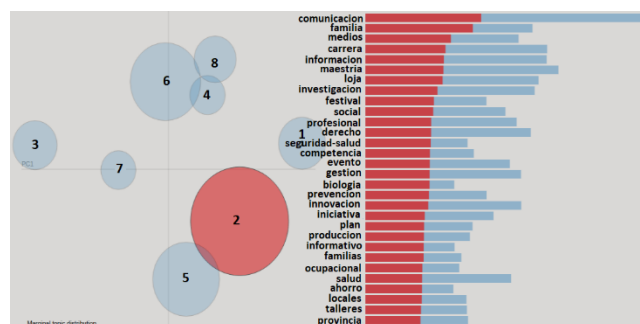
### 3. Results and Discussion

Once the reference values for  $k$  were obtained, some experiments were made considering the BoW and TF-IDF matrixes and different  $k$  values from 8 to 12. Matrices with TF-IDF gave us the best coherence results; therefore, we decide to explore a visual representation of the topics using the pyLDAvis with the following TF-IDF matrixes:

- Matrix TF-IDF,  $K = 8$ , with unigrams and bigrams
- Matrix TF-IDF,  $K = 10$ , with unigrams and bigrams

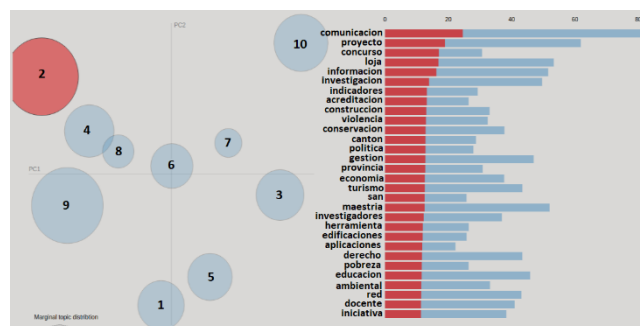
The pyLDAvis package represents topics by circles in 4 quadrants.

As long as the circles are more dispersed between the quadrants and between each other, we have a good model. Also, this graph allows the option to visualize the terms contained in a topic, this appears on the right side in red color, and the terms are ordered based on their relevance. Figures 6 and 7 shows the pyLDAvis results, when  $k=8$  exists some topics overlapped, while when  $k=10$  presents a good distribution of topics.



**Figure 6.** pyLDAvis graph with TF-IDF and  $k=8$

Source: Authors



**Figure 7.** pyLDAvis graph with TF-IDF and  $k=10$

Source: Author

Figure 7 shows the predominant topic 2 with terms: “comunicacion”, “proyecto”, “concurso”, “loja”, “informacion”, “investigacion”, “indicadores”, “acreditacion”, “construccion”, “violencia”; followed by topic 9 with the terms: “proyecto”, “salud-mental”, “cultural”, “conservacion”, “comunicacion”, “gastronomia”, “especies”, “museo”, “mujer”, “virtual”, and topic 3 with the terms: “derecho-penal”, “radio”, “covid”, “prendho”, “crisis”, “pandemia”, “personal”, “familia”, “graduados” and “instituto”.

Considering that the value of  $k=10$  has the best topics distribution, next, we will present in Table 1 the topics obtained with TF-IDF matrix when  $k=10$ . Topics listed from 0 to 9 correspond with topics 1 to 10 in pyLDAvis Figure 7. The number before each term represents the probability of the term belong to the topic.

All the texts analyzed were in a university environment; therefore, the topics are related to the activities developed in an academic field. Topics 5, 7, and 9 refers to different carriers that the UTPL offers. Topic 4 refers to master programs, topic 6 mentions two modalities of study such as open university and continuing education. Topic 3 refers to a family event as well as studies and researches. Topics 0, 1, 2 y 8 presents words related to studies, projects and researches developed by the UTPL.

In Table 2, we verified that the topics assigned by the model are related to the text within the news. This table shows some news of the corpus, columns Num\_News represents the news number,

Dominant\_Topic refers to the most relevant topic assigned to news, Term\_in\_Topics represents the terms in the dominant topic, and Terms\_in\_New presents some terms within the news.

As an example, news number 85 and 119 refer to “concurso”, news 187 refers to “construccion”. Also, news 13 refers to “conservación”, news 7 refers to “gastronomía”, and news 15 and 131 refers to “cultural”.

In general, after analyzing the results, the TF-IDF topics include a variety of terms facilitating and making a distinction among topics.

**Table 1.** Topics obtained with the TF-IDF matrix with k=10  
Source: Authors

| Topics  | Terms                                                                                                                                                                                                                        |
|---------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Topic 0 | 0.003**"paz" + 0.002**"campus" + 0.002**"comunicacion" + 0.002**"violencia" + 0.002**"educacion-superior" + 0.002**"ninios" + 0.002**"investigacion" + 0.002**"red" + 0.002**"mejora" + 0.002**"educacion"                   |
| Topic 1 | 0.003**"comunicacion" + 0.003**"proyecto" + 0.002**"concurso" + 0.002**"loja" + 0.002**"informacion" + 0.002**"investigacion" + 0.002**"indicadores" + 0.002**"acreditacion" + 0.002**"construccion" + 0.002**"violencia"    |
| Topic 2 | 0.002**"estudios" + 0.002**"mujer" + 0.002**"educacion" + 0.002**"universidad" + 0.002**"feria" + 0.002**"innovacion" + 0.002**"prueba" + 0.002**"museo" + 0.002**"sur" + 0.002**"software"                                  |
| Topic 3 | 0.005**"familia" + 0.003**"comunicacion" + 0.002**"radio" + 0.002**"congreso" + 0.002**"ods" + 0.002**"educativa" + 0.002**"sector" + 0.002**"medios" + 0.002**"pobreza" + 0.002**"cambios"                                  |
| Topic 4 | 0.004**"maestria" + 0.003**"festival" + 0.003**"suelo" + 0.003**"comunicacion" + 0.003**"paz" + 0.002**"mencion" + 0.002**"informacion" + 0.002**"responsabilidad-social" + 0.002**"laboratorio" + 0.002**"equipo"           |
| Topic 5 | 0.003**"derecho" + 0.003**"mujeres" + 0.003**"ods" + 0.002**"artes" + 0.002**"aplicacion" + 0.002**"mayo" + 0.002**"loja" + 0.002**"muerte" + 0.002**"donaciones" + 0.002**"carrera"                                         |
| Topic 6 | 0.003**"distancia" + 0.003**"ambiental" + 0.002**"virtual" + 0.002**"secretario" + 0.002**"carreras" + 0.002**"ayudar" + 0.002**"ideas" + 0.002**"educacion-continua" + 0.002**"seminario" + 0.002**"dolor"                  |
| Topic 7 | 0.003**"turismo" + 0.003**"personas-discapacidad" + 0.003**"comunicacion" + 0.002**"social" + 0.002**"psicologia" + 0.002**"convenio" + 0.002**"maestria" + 0.002**"libros" + 0.002**"experiencia" + 0.002**"reconocimiento" |
| Topic 8 | 0.003**"proyecto" + 0.003**"salud-mental" + 0.003**"cultural" + 0.003**"conservacion" + 0.003**"comunicacion" + 0.002**"gastronomia" + 0.002**"especies" + 0.002**"museo" + 0.002**"mujer" + 0.002**"virtual"                |
| Topic 9 | 0.004**"derecho-penal" + 0.003**"radio" + 0.003**"covid" + 0.003**"prendho" + 0.002**"crisis" + 0.002**"pandemia" + 0.002**"personal" + 0.002**"familia" + 0.002**"graduados" + 0.002**"instituto"                           |

## 2.1. Manual validation

Coherence value is a metric that allow us to validate the model, but considering that the university news blog has hashtags placed at the end of each news, we decided to make a comparison of the hashtags and the topics terms obtained in each news. Table 3 shows this comparison, where the coincidences are highlighted in two colors: pink color for exact terms and green color for similar terms. For example, the relevant topics for news number 1 are 9, 0, and 1; in topic 9 the terms “covid” and “pandemia” are related to “semaforo amarillo”. In Ecuador due to the covid-19 pandemic, a traffic light system advises about the level of

contagion. In topic 0 the term “educacion” appears highlighted in pink because it coincides in both columns. In news number 6 the relevant topics are 5 and 1, in this case, the word “donacion” are highlighted with pink, and the words “loja” and “utpl” with green because the principal campus of the university analyzed is in Loja. In this way, it can be concluded that the topics identified with LDA are quite good since most of them are related to the hashtags and the words within the news. We must to remark that all the news did not have hashtags; therefore, we did not calculate a numeric representation of the coincidence between hashtags and the terms of the topics.

**Table 2.** Examples of topics assigned to documents  
Source: Authors

| Num_<br>News | Dom -<br>inant_<br>Topic | Terms_in_Topics                                                                                                        | Terms_in_News                                                                                                                                                                                                                      |
|--------------|--------------------------|------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 85           | 1                        | comunicacion, indicadores, informacion, acreditacion, loja, concurso, proyecto, violencia, investigacion, construccion | utpl, destaca, concurso, arquitectura, urbanistica, quitoel, noviembre, teatro, sucre, quito, sede, premiacion, finalistas, concurso, arquitectura, diseno, plan, especial, ...                                                    |
| 119          | 1                        | comunicacion, indicadores, informacion, acreditacion, loja, concurso, proyecto, violencia, investigacion, construccion | estudiantes, utpl, obtienen, competencia, nacional, pasteleriaen, competencia, nacional, pasteleria, expo, sweet, desarrollo, centro-convenciones, metropolitano, quito, rosalia-arteaga, veronica, abad, estudiantes, carrera,... |
| 187          | 1                        | comunicacion, indicadores, informacion, acreditacion, loja, concurso, proyecto, violencia, investigacion, construccion | constructores, fortalecen, conocimientos, edificaciones, lojapara, garantizar, calidad, construccion, obras, civiles, loja, universidad-tecnica, particular-loja, utpl,...                                                         |
| 13           | 8                        | comunicacion, especies, virtual, salud-mental, proyecto, conservacion, museo, cultural, gastronomia, mujer             | turismo, aporta, conservacion, medioambienteel, turismo, industria, peso, economia, mundial, convertido, sector, crecimiento, mundo, emplea, millones, personas, nivel, global, datos, organizacion-mundial, ...                   |
| 7            | 8                        | comunicacion, especies, virtual, salud-mental, proyecto, conservacion, museo, cultural, gastronomia, mujer             | copa, culinaria, utpl, reto, expone, talento, jovenesla, gastronomia, profesion, innovadora, emplea, atributos, creatividad, destrezas, motrices, elaborar, postre, sofisticado, ...                                               |
| 15           | 8                        | comunicacion, especies, virtual, salud-mental, proyecto, conservacion, museo, cultural, gastronomia, mujer             | estudiantes, utpl, innovan, museo-matilde, hidalgo, lojael, impacto, figura, ecuatoriana, matilde-hidalgo, mujer, latinoamerica, votar, eleccion, nacional, motivo, ...                                                            |
| 131          | 8                        | comunicacion, especies, virtual, salud-mental, proyecto, conservacion, museo, cultural, gastronomia, mujer             | conferencias, virtuales, arte, cultura, desarrollan, difundir, proyectos, estrategias, fortalezcan, sector, cultural, ecuador, ...                                                                                                 |

#### 4. Conclusions

When analyzing the corpus of digital news with the LDA model, a better topic coherence was obtained using the TF-IDF matrix constructed with unigrams, bigrams, words lemmatized by nouns, adjectives and verbs, and with a value of  $k=10$ . Considering the news correspond to an academic environment, the topics obtained refer to a greater visibility of the work and activities carried out by the communication, tourism and law careers. In addition, considering that the university has two modalities of study most of the news focused on the open modality. Furthermore, some topics show the interest of the university in research and social activities that involve the promotion of a culture of peace and non-violence, social responsibility, health and so on.

In the manual validation process many words or terms in topics assigned to news correspond with news hashtags, then we pretend to implement a web application to recommend hashtags according to the news. Also, for future work, others topic modeling algorithms could be applied in order to improve the results.

#### Authors' contribution

**Lizbeth Pacheco-Guevara:** Methodology, experimentation, results, article editing. **Ruth Reátegui:** Introduction, conclusions, supervision, article editing. **Priscila Valdiviezo-Díaz:** Conclusions, supervision, article editing.

**Table 3.** Comparison between the topics emitted by LDA and hashtags given from the news  
Source: Authors

| Num_ News | Terms_in_Topics                                                                                                                 | Hashtags_News                                                      |
|-----------|---------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------|
| 1         | Topic 9: derecho-penal, radio, covid, prendho, crisis, pandemia, personal, familia, graduados, instituto                        |                                                                    |
|           | Topic 0: paz, campus, comunicacion, violencia, educacion-superior, ninos, investigacion, red, mejora, educacion                 | #Retorno, #semáforo amarillo<br>#educacion                         |
|           | Topic 1: comunicacion, proyecto, concurso, loja, informacion, investigacion, indicadores, acreditacion, construccion, violencia |                                                                    |
| 6         | Topic 5: derecho, mujeres, ods, artes, aplicacion, mayo, loja, muerte, donaciones, carrera                                      | #donacion<br>#campanaDalesuna patita                               |
|           | Topic 1: comunicacion, proyecto, concurso, loja, informacion, investigacion, indicadores, acreditacion, construccion, violencia | #relacionespublicas utpl<br>#utpl solidaria"                       |
|           |                                                                                                                                 |                                                                    |
| 11        | Topic 9: derecho-penal, radio, covid, prendho, crisis, pandemia, personal, familia, graduados, instituto                        | "teleasistencia médica<br>#COVID-19<br>#asistencia médica gratuita |
|           | Topic 1: comunicacion, proyecto, concurso, loja, informacion, investigacion, indicadores, acreditacion, construccion, violencia | #Loja<br>#utpl<br>#Hospital UTPL"                                  |
|           |                                                                                                                                 |                                                                    |

### Conflicts of interest

The authors declare no conflict of interest

### Bibliographic references

- Blei, D.M., Ng, A.Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Buenano-Fernandez, D., Gonzalez, M., Gil, D., & Lujan-Mora, S. (2020). Text Mining of Open-Ended Questions in

Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access*, 8, 35318–35330. doi: 10.1109/ACCESS.2020.2974983.

- Chipidza, W., Akbaripourdebazar, E., Gwanzura, T., & Gatto, N.M. (2021). A topic analysis of traditional and social media news coverage of the early COVID-19 pandemic and implications for public health communication. *Disaster Medicine and Public Health Preparedness*, 3, 1-8. doi:10.1017/dmp.2021.65
- Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189. doi:10.1017/pan.2017.44
- Guangce, R., and Lei, X. (2021). Knowledge discovery of news text based on artificial intelligence. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(1), 2021, doi:10.1145/3418062
- Kim, D., Seo, D., Cho, S., & Kang, P. (2019) Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Inf. Sci. (Ny)*, 477, 15–29. doi: 10.1016/j.ins.2018.10.006.
- Larsen, V.H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1), 203–218, doi: 10.1016/j.jeconom.2018.11.013
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. EMNLP 2011 - Conf. Empir. *Methods Nat. Lang. Process. Proc. Conf.*, 2, 262–272.
- Misztal-Radecka, J. (2018). Building semantic user profile for polish web news portal. *Computer Science*, 19(3), 307–332, 2018.
- Oliveira Capela F. de, & Ramirez-Marquez, J. E. (2019). Detecting urban identity perception via newspaper topic modeling. *Cities*, 93, 72–83. doi: 10.1016/j.cities.2019.04.009.
- Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., and Yao, H. (2019). Research on Topic Detection and Tracking for Online News Texts. *IEEE Access*, 7, 58407–58418, doi: 10.1109/ACCESS.2019.2914097.
- Xu, L., Yeo, H., Hwang, H., & Kim, K.O. (2020). 5G service and discourses on hyper-connected society in south Korea: Text mining of online news. *Advances in Intelligent Systems and Computing*, 1120, 892-897, doi:10.1007/978-3-030-39442-4\_68
- Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electron. Commer. Res. Appl.*, 29, 142–156. doi: 10.1016/j.elerap.2018.04.003.

