Automatic social media news classification: a topic modeling approach

Clasificación automática de noticias en redes sociales: una aproximación desde el modelado de tópicos

Daniel Amador[1], Carlos Gamboa-Venegas[2],
Ernesto García[3], Andrés Segura-Castillo[4]

Amador, D.; Gamboa-Venegas, C.; García, E.; Segura-Castillo, A. Automatic social media news classification: a topic modeling approach. *Tecnología en Marcha*. Vol. 35, special issue. IEEE International Conference on Bioinspired Processing. December, 2022. Pág. 4-13.

https://doi.org/10.18845/tm.v35i9.6477

[1] Centro Nacional de Alta Tecnología. E-mail: damador@cenat.ac.cr
https://orcid.org/0000-0003-1197-4313
[2] Centro Nacional de Alta Tecnología. E-mail: cgamboa@cenat.ac.cr
https://orcid.org/0000-0001-9712-0575
[3] Universidad de Costa Rica. E-mail: luis.garciaestrada@ucr.ac.cr
[4] Universidad Estatal a Distancia. E-mail: asegurac@uned.ac.cr
https://orcid.org/0000-0001-5647-1176

Keywords

Automatic news classification; social media; topic modeling.

Abstract

Social media has modified the way that people access news and debate about public issues. Although access to a myriad of data sources can be considered an advantage, some new challenges have emerged, as issues about content legitimacy and veracity start to prevail among users. That transformation of the public sphere propels problematic situations, such as misinformation and fake news. To understand what type of information is being published, it is possible to categorize news automatically using computational tools. Thereby, this short paper presents a platform to retrieve and analyze news, along with promising results towards automatic news classification using a topic modeling approach, which should help audiences to identify news content easier and discusses possible routes to improve the situation in the near future.

Palabras clave

Clasificación automática de noticias; Redes sociales; Modelado de tópicos.

Resumen

Las redes sociales virtuales han modificado significativamente la forma en la que las personas acceden a contenido noticioso y, por ende, el debate en la esfera pública. Aunque el acceso a múltiples y diversas fuentes puede considerarse una ventaja, a su vez genera situaciones problemáticas relacionadas con la legitimidad y veracidad del contenido circulante, por ejemplo, desinformación y noticias falsas. Para lograr entender qué tipo de información se está publicando, se puede llevar a cabo una categorización de las noticias por tema, con ayuda herramientas computacionales para realizar este proceso de forma automática. Así, este artículo corto presenta una plataforma para recuperar y analizar noticias, así como resultados prometedores del uso de modelado de tópicos para la clasificación automática de contenido noticioso, en aras de facilitar a la audiencia la categorización del contenido. Asimismo, discute las rutas posibles a seguir para mejorar la propuesta a futuro.

Introduction

Social media has modified the ways people consume information and engage in political participation [1, 2, 3]. Although easy access to a myriad of data sources is a considerable advantage, it has brought some new challenges to the table. One of them has to deal with the amount of news that circulates and how to automatically and accurately classify them. For instance, [4], has pointed out that, given the vast amount of news circulating on social media, it has become more difficult for readers to identify topics according to their content, therefore, making fake news proliferation more possible and thus an issue. Research from the London School of Economics [5] insists that this situation has created a “trust crisis” that leads to a permanent social uncertainty. Moreover, Livingstone [5] argues that politicians, institutions, platforms, news media and activists have not found a viable way to mitigate the issue.

From another perspective, Waisbord [6] points out that the main challenge we have to deal with, is the transformation of the public sphere. Traditional news media legitimacy is at stake, the proliferation of digital spaces to express opinions publicly without any previous editorial filter contributes to a wider debate, but at the same time poses a problem for content validity and veracity. At the end of the day, behind issues such as massive media content and fake news, there is a dispute between different ideas about legitimacy and power, an "epistemic democracy" has emerged.

The relevance of understanding and mitigating this situation of interest is clear for the scientific community and contributions from computational perspectives to this matter are highly expected. Therefore, a topic modeling approach is proposed as a means to contribute to automatic social media news content detection, thus reducing the uncertainty for the user. The article presents the bases of topic modeling, then gives a description of the implemented approach, shows some preliminary results and finally discusses possible routes for future development and improvement.

Topic modeling

Amongst modern techniques of computational and automatic processing applied to digital humanities and social sciences, topic modeling is an unsupervised machine learning methodology that provides a suitable approach to problems that involve clustering, semantics and discourse analysis. In broad terms, a topic model algorithm receives an input of unclassified texts and outputs a set of topics that are common throughout a corpus. Additionally, it provides a net weight that reflects the degree to which the text belongs to a possible topic.

The topic modeling algorithm to be discussed in this article is a Latent Dirichlet Allocation (LDA). This model was proposed as a bioinspired approach to study population genetics by Prichard, Stephens and Donnelly in 2000, and later applied to computational science by Blei, NG and Jordan in 2003.

According to Blei [7] there are two key assumptions that an LDA model requires in order to be successful:

1. All documents in the corpus share similar word patterns that are called topics.
2. All the documents in the corpus can be allocated within any of this limited amount of topics.

Biel [7] states that the generative process of LDA begins once a corpus is feeded into the system. First, the process chooses an amount of topics, given as a parameter, to distribute the documents. The distribution distributes the contents among the topics using a Dirichlet distribution and then a multinomial one. Afterwards, each document is assigned with a weight to describe what topics have major presence in it. Finally, for every word composing the document the system will choose a topic assignment and will generate a new document with a given probabilistic weight of similarity towards the previous document inputted. As stated by Blei, NG and Jordan (2003), given the parameters α and β, the combined distribution of a set of topics θ, a set of N topics z, and a set of N words w is:

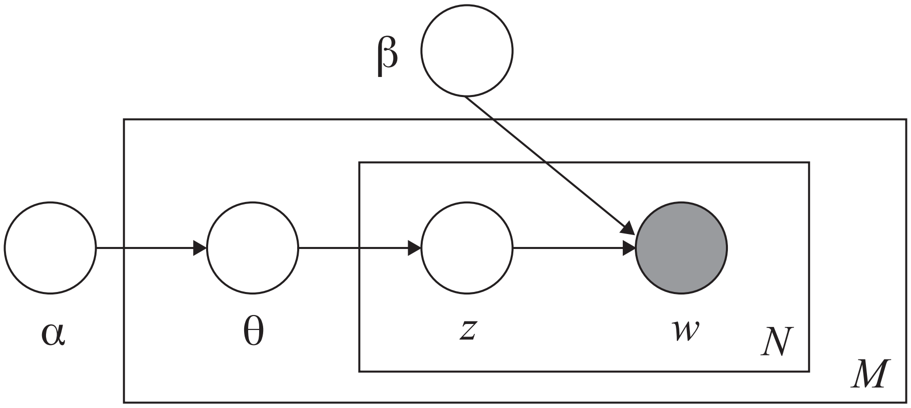

Figure 1. Graphical model representation of LDA retrieved from [8].

Blei, NG and Jordan [8], indicate that the boxes are "plates" representing replicates. The outerplate represents documents, while the inner plate represents the repeated choice of words within a document.

In the wider sense, an LDA model defines topics by two major sources of probabilistic outcomes within an operation: Dirichlet distribution and multinomial distribution. The Dirichlet distributions relate documents-topics and topics-words. These topics are decided by the user to further be computed by the system which will then cluster the words by topic related to the training corpus provided.

Currently, for ease of researching teams in the realm of semantic analysis and Natural Language Processing (NLP), there are open sourced libraries that can be implemented to projects. In the present article GenSim by Řehůřek and Sojka [9] was utilized. GenSim is a pure Python library that provides a memory-efficient and scalable out-of-the-box solution.

Methodology: Implemented approach

Based on previous work from Soto-Rojas et al. [10], Facebook posts from local news pages were automatically retrieved and stored on a database. It is important to mention that the platform that allowed our data retrieval, also includes an interface to request information for public consulting purposes, a Content Management System for expert content validation and administration and some visualization tools to explore the contents available on the platform. Having the benefit of such a platform, allowed us to explore the implementation of a Topic Modeling Approach to automate the classification of the news. At the moment, only 10% of the data, i.e, 91725 posts, have been manually tagged by experts into several categories. The pages from local media selected for this study are: La Nación (the reference newspaper in the country), CR Hoy (the most popular online newspaper), Telenoticias (a long time running news TV program), Repretel (a high rating news TV program), Semanario Universidad (an academic newspaper), El Financiero (a theme based newspaper, mostly focusing on economical issues), Noticias Monumental (a well known news radio program), Prensa Libre (the oldest newspaper), Diario La Extra (a tabloid newspaper) and Amelia Rueda (a popular news radio program). We have collected all 91795 posts for the topic modeling experimentation.

Figure 2 shows a general view of the complete solution, including the steps for data retrieval, storage and the Application Programmers Interface available for public request. Moreover, the process for topic modeling with the visual tools is also included.

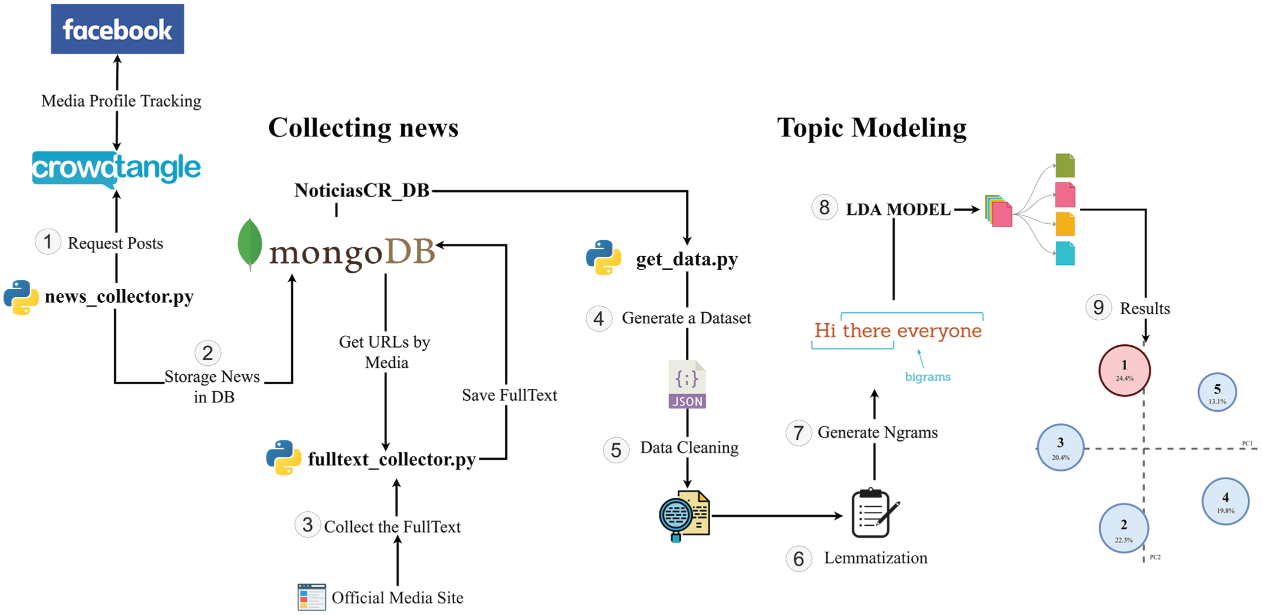

Figure 2. Diagram of the current infrastructure, with the complete execution pipeline of the solution, ending with the analysis of news using Top Modeling with LDA.

Data retrieval and storage

To store retrieved data we selected MongoDB, a document-based database that allows storage of data with different structures with high query efficiency. This database model makes it possible to have different collections in the same storage space. The database is hosted in the Kabre supercomputer at Centro Nacional de Alta Tecnología (CeNAT), and currently stores more than 1 million posts from Costa Rican news sources on Facebook, ranging from 2018 to the moment.

It is very important to mention that we obtained legal access to Facebook post content thanks to an agreement with CrowdTangle, their official source for academic purposes. Thus, we are able to retrieve the posts with all its contents and several aggregated statistics that show their respected engagement. No user data is ever manipulated or stored and the agreement strictly prevents non aggregated data publication [11].

Collecting news takes two steps. The first one gathers posts and their provided data from Facebook profiles of Costa Rican news media and stores them in the database. The second part takes the external link provided with the retrieved data and extracts the complete text from the original news source at their respective official website.

However, using CrowdTangle as a tool to consult Facebook posts, has its limitations. There is a rate limit for requests, when this is overpassed, indexes are restarted, thus causing duplication of data. Usually, it is possible to extract 10,000 entries before this happens. To avoid this issue, the implemented script gathers news daily only.

Then, to feed the database, a Python script is in charge of querying the list of news pages accessed through CrowdTangle, a library called Pytangle allowed to obtain the publications by a date range, in this module the filtering is applied to prevent the rate limit of the requests from being exceeded. In addition, a log was added to prevent duplicity of data by saving the last date recorded in the database, when an error occurs in the collection, the last date is loaded and the execution of the collector is restarted.

News texts retrieval

Each collected post from Facebook has a link to the original news source. Next in the process, that link is used to extract the original full text from the news media publication. Since every website is created with a different technology, several scripts were implemented to extract the text from each one, these programs were unified as a single scrapper. When the program starts, it runs nine scripts, each connecting to the database to add contents to the field "fullText". Each execution retrieves 200 news from each news source.

Additionally to the nine scrappers, the tool includes four scripts written in Python. Two containing information about configuration and connection to the database. And two more with the structure of the data to be read and the main code with the necessary flow to run each scrapper separately.

Topic modeling implementation

With all the posts and full content stored in the database, we proceeded with an exploration of the more topics prevalent in a specific set of news. To do so, an LDA model was used to perform the experiments with 55815 news from four evenly distributed previously tagged categories.

For the controlled experiments, a pipeline was implemented that is responsible for data cleaning, lemmatization, N-gram creation, corpus generation and subsequently the application of the LDA model. Figure 3 illustrates the steps of the data processing pipeline. Using this pipeline, the following cases were executed: identification of 5 topics, identification of 7 topics and identification of 10 topics.

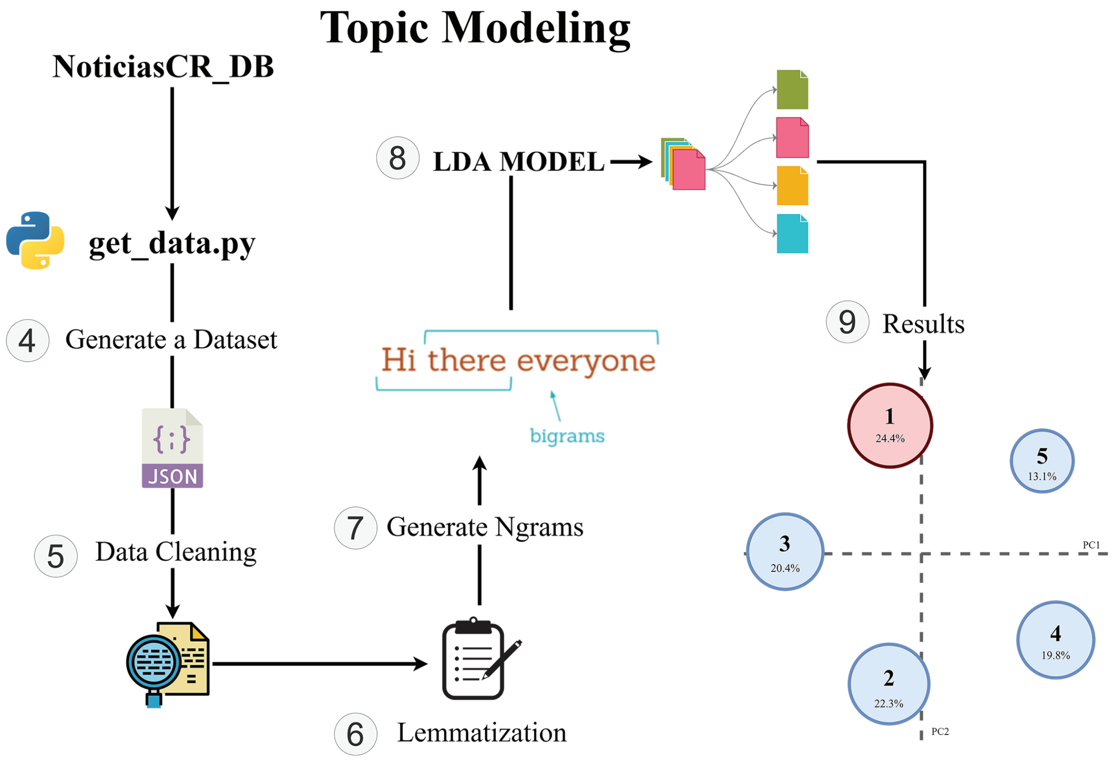

Figure 3. Diagram of the LDA pipeline, with the complete execution of data processing.

For each of the cases a series of data preprocessing steps must be followed. The first step is the selection of the data from the dataset, as mentioned above, an attempt was made to balance the data by selecting a balanced amount of news associated with each topic. With the dataset selected, the second step is the cleaning of the data, where we applied the elimination of stopwords, accents and converted all texts to lowercase. Approximately 1700 words were used as stopwords, which were removed from the texts since they have no significant relevance for the analysis. At the end of the data cleaning, the third step is the generation of the lemmatization, which consists of converting all conjugated words into their base form (e.g. "Robó", "Robando", "Robar") and thus avoid duplication of the meaning of the same word. When the lemmatization is finished, the N frames are created, which correspond to the association of N number of words, in this case, bigrams and trigrams were used. Once the N frames are generated, a corpus is generated with all the word unions already filtered and cleaned to be processed by the model. Finally, the LDA model is applied using the gensim library, which contains a series of models for the study of natural language processing.

The LDA model receives the following elements as parameters: corpus, id2word, num_topics, random_state, update_every, chunksize, passes, alpha. The corpus represents the sparse shape matrix (num_documents, num_terms) and is generated prior to execution, as well as, the id2word which serves to determine the size of the vocabulary. The number of topics was set according to each experiment, so three separate runs were made with 5, 7 and 10 numbers of topics to be identified. For random state a value of 100 was assigned and a value of 1 was added to the update every which represents the number of documents to use in each update, the chunksize which represents the number of documents to use in each training chunk was defined as 100, and the passes which is the number of passes through the corpus during training was set to 10. Finally, the alpha was defined as ***auto*** which means that the learning strategy is based on the asymmetric prior from the corpus.

At the end of the model training, a result is obtained with the different topics and the most relevant words for each one of them. To visualize these results, a library called pyLDAvis [12] was used to generate a diagram of the relevance of the words and the incidence of the topics in the total number of news items in the dataset. The obtained visualizations are presented in the next section.

Evaluation

To evaluate the model we use the coherence metric [13]. For the experiments we know that words will be grouped into a *topic*. This *topic* group is said to be coherent when the words of the group support each other in the same semantic meaning. For this reason a topic group can be interpreted as such in a context that embraces all or most of the words.

Complimentary, expert criteria were consulted. In terms of the expert criteria, elicited to determine a gold standard to classify the models, the researchers utilized The Semantic Fields Theory (Trier. 1931 as cited by Gliozzo [14]). This theory establishes that any language lexicon is structured into *Semantic Fields,* which is defined as semantic relations among concepts belonging to the same field. In other words, semantic relations between words of the same field have a very close semantic relationship while words that do not belong to the same field are very distant between each other. This can be exemplified by giving a lexicon of words: bird, dog, space, cat, math; where we can identify that bird, dog and cat belong to an "animal's field" while the other entries could belong to a "scientific field".

Based on this theory, a consulted expert proposed a classification, when possible, to the topics outputted by the model. As mentioned before, this will serve as a human inputted gold standard from the field of discourse analysis to cross check the performance of the model.

Results

In this section we show preliminary results and visualization of the LDA model with intertopic distance maps. The goals is to explore and analyze if the topics generated by the model can represent a similar distribution of subjects already labeled in the news of the complete dataset. The intertopic distance map is a visualization of the topics in a two-dimensional space. Each topic is represented by a circle, and its size is proportional to the amount of tokens (words) that belong to each topic across the dictionary.

Figure 4 illustrates the results of LDA with 5 topics. Topic 1 has the highest percentage and includes in its top 5 words like: *país, mundo, precio, nivel, nacional.* Following with words such as *mercado, mundial, internacional* in the top 10. This might be evidence that the subject in question is related with International and National news, and maybe economy and market. The smaller topic is number 5 which includes words like *educativo, proyecto, educación, nacional.* Suggesting that the education topic is the least covered in the dataset.

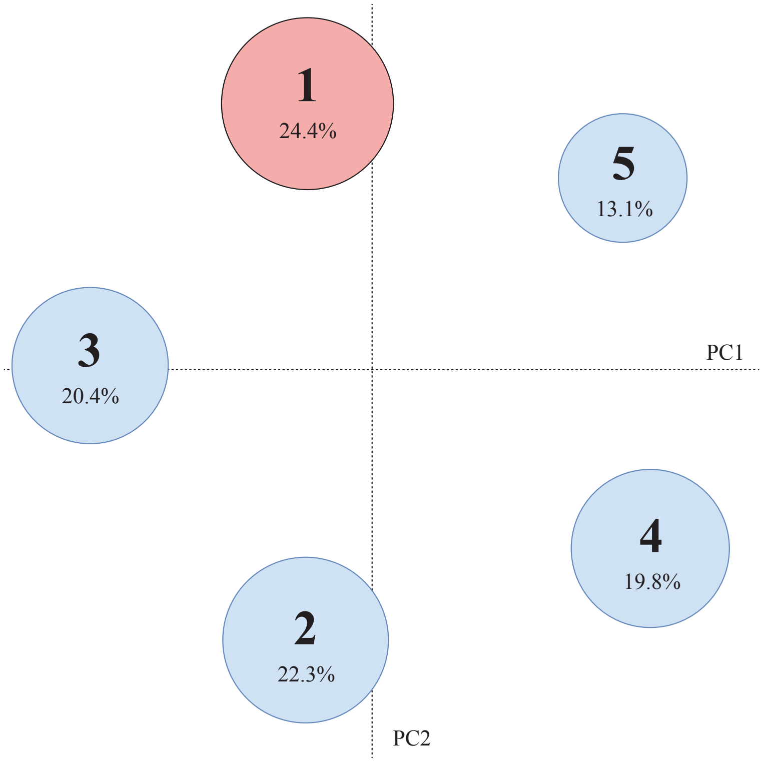

Figure 4. Intertopic Distance Map of LDA with 5 topics, and the percentage of tokens per group.

After running the three experiments, the coherence metric is returned by the algorithm, resulting in 79% for the 5 topics, 70% for the 7 topics and 64% for the 10 topics. This means that the 5 and 7 topic models scored high enough to be considered good candidates for the classification. At this point, expert criteria is needed to determine which model fits best for the purposes of the research.

In table 1, the expert criteria for the three different models are identified, if the words of the topic have not a clear semantic relationship then it will be tagged as unclassifiable.

Table 1. Expert criteria consulted for topic model experiments with 5, 7 and 10 topics. Topics numbers are independent, they do not indicate any relation among experiments.

	Semantic Field		
Topic #	5 topics	7 topics	10 topics
1	International	Unclassifiable	Unclassifiable
2	Health	Unclassifiable	Unclassifiable
3	Unclassifiable	International	International
4	Crime and judiciary	Government affairs	Government affairs
5	Education	Health	Health
6		Unclassifiable	Unclassifiable
7		Crime and judiciary	Crime and judiciary
8			Unclassifiable
9			Unclassifiable
10			Unclassifiable

By expert criteria suggestion, the model that best fits the semantic content of the data is the 5 topic one.

Discussion and conclusions

Results show an effective topic modeling approach towards automatic news classification, which could help news media to classify the topic without depending on an editorial intervention. Given the vast amount of news that circulates on social media, that would also mean a significant advantage for audiences, who would have an accurate classification of their content of interest in advance.

Furthermore, this approach could contribute to the reduction of the amount of time social media researchers have to invest for tagging purposes during their projects. To our knowledge, and particularly in Costa Rica, this would be the first tool available for that goal.

On the other hand, there are still some limitations to work around. First, Costa Rican news outlets on social media tend to focus on crime and judiciary issues, which could possibly create a bias for future automatic classification. As our database grows, more of these cases are included, therefore, the proposed topic modeling approach still needs to adjust the algorithm to efficiently avoid such an issue without losing accuracy.

Secondly, our gold standard approach derived from an expert consultation has several limitations. Without a clear quantitative parameter to classify the topics based on the semantic field they belong to, there exists room for error in the decision the expert took for classifying the topic. Furthermore, the distance between words that belong to a certain topic is very narrow for this article as they can be different parts-of-speech that, in the right context, could be part of any topic.

In the near future, it would be interesting to compare the approach with another unsupervised classification paradigm, for instance, convolutional neural networks. Given the amount of data available, it would be a feasible and interesting comparison. As a result, a mixed approach could be developed as a contribution to the field.

Finally, it is important to bear in mind that social media news consumption is a complex and evolving situation of interest. It unfolds in a specific context, therefore, special care needs to be taken before any generalization of findings. We have presented an effective topic modeling approach for the Costa Rican case, but other contexts might require particular adjustments that are out of our scope at the moment.

References

[1] T. Highfield, “Social media and everyday politics”, Cambridge: Polity Press, 2016.

[2] H. Margetts, P. John, S. Hale, & T. Yasseri, “Political turbulence: How social media shape collective action”, Princeton: Princeton University Press, 2016.

[3] N. Newman, et al, “Reuters Institute Digital News Report 2019”, Oxford: Reuters Institute, 2019.

[4] A. Marwick, “Why Do People Share Fake News? A Sociotechnical Model of Media Effects”, Georgetown Law Technology Review, 2(2), pp:474-512, 2018

[5] S. Livingstone, “Tackling the Information Crisis: A Policy Framework for Media System Resilience”, Foreword. In LSE 2018. p: 2. London: LSE, 2018.

[6] S. Waisbord, “Truth is what happens to news: On journalism, fake news, and post-truth. Journalism Studies”, 19(13), pp:1866–1878, 2018.

[7] D. Blei, “Topic Modeling and Digital Humanities”. *Journal of Digital Humanities*, 2(1), pp: 8-11. 2021.

[8] D. M. Blei, A. Y. Ng & M. I. Jordan, “Latent dirichlet allocation”. *Journal of machine Learning research*, 3(Jan), pp: 993-1022, 2003.

[9] R. Řehůřek, & P. Sojka, “Gensim—statistical semantics in python” Retrieved from genism.org, [Accessed May. 2, 2011]

[10] C. Soto-Rojas, C. Gamboa-Venegas, A. Céspedes-Vindas, “MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media”. *Tecnología en Marcha. Edición especial 2020*. 6th Latin America High Performance Computing Conference (CARLA). pp: 18-24. 2020.

[11] CrowdTangle Team. CrowdTangle. Facebook, Menlo Park, California, United States. [List ID: 1510711], 2020.

[12] C. Sievert and K.Shirley. “LDAvis: A method for visualizing and interpreting topics”. *In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp: 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics, 2014.

[13] M. Röder, A. Both, and A. Hinneburg. 2015. “Exploring the Space of Topic Coherence Measures”. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (WSDM ‘15). Association for Computing Machinery, New York, NY, USA, pp: 399–408. 2015.

[14] A. Gliozzo, “Semantic domains and linguistic theory”. In Proceedings of the LREC 2006 workshop Toward Computational Models of Literary Analysis, Genova, Italy. 2006.