Comunicación y Sociedad
Departamento de Estudios de la Comunicación Social
Universidad de Guadalajara

Modeling and Diffusion of News Topics in Social Media: Features and Factors of the Emergence of News in a Twitter Informative Channel

Modelado y difusión de temas noticiosos en medios sociales: características y factores de la emergencia de noticias en un canal informativo de Twitter

DOI: https://doi.org/10.32870/cys.v2019i0.6437

Carlos Arcila Calderón[1]
http://orcid.org/0000-0002-2636-2849
Eduar Barbosa Caro[2]
https://orcid.org/0000-0003-0297-8224
Ignacio Aguaded[3]
https://orcid.org/0000-0002-0229-1118

This study aims to characterize the modeling and diffusion of news topics in social media and determine the factors that influenced them. With Big Data analysis methods, such as topic modeling and sentiment analysis, we analyzed one year of tweets from Colombian newspaper *El Tiempo*. We found that the appearance of long-term topics was related to the message's attributes. Theoretical implications and contributions considering Diffusion of Innovations' model are mentioned.
Keywords: Twitter, dissemination of news, social media, topic modeling, Big Data.

Este estudio busca caracterizar el modelado y difusión de temas noticiosos en medios sociales y determinar los factores que influyen en su aparición. Con técnicas de Big Data, como el modelado de temas y el análisis de sentimientos, se analizó un año de tuits del periódico colombiano El Tiempo*, encontrando que la aparición de temas en el largo plazo se relaciona con atributos del mensaje. Se mencionan implicaciones teóricas y contribuciones para otros modelos a la luz del modelo de Difusión de Innovaciones.*
Palabras clave: Twitter, difusión de noticias, medios sociales, modelamiento de temas, Big Data.

How to cite:
Arcila Calderón, C., Barbosa Caro, E. & Aguaded, I. (2019). Modeling and Diffusion of News Topics in Social Media: Features and Factors of the Emergence of News in a Twitter Informative Channel. *Comunicación y Sociedad*, e6437. DOI: https://doi.org/10.32870/cys.v2019i0.6437

1 Universidad de Salamanca, Spain.
E-mail: carcila@usal.es
2 Universidad del Rosario, Colombia.
E-mail: eduar.barbosa@urosario.edu.co
3 Universidad de Huelva, Spain.
E-mail: aguaded@uhu.es
Submitted: 09/01/17. Accepted: 26/07/17. Published: 06/03/19.

Introduction

One of the most relevant aspects in the rapid and innovative evolution of the digital environment in the last decade has been the implementation of Internet tools to extend the communication networks, widening the media's target. Therefore, it follows that mass media from around the world have confidently opted for the use of interactive tools and social media to publish, inform and interact with their readers and audiences (Caballero, 2001; García de Torres et al., 2008; Lasorsa, Lewis & Holton, 2012; Said-Hung et al., 2013).

This paper presents the results of a large-scale quantitative study that addressed issues such as the use of the web in journalistic production (Micó, Canavilhas, Masip & Ruiz, 2008) and the efficiency and speed with which the mass media channels generate knowledge (Rogers, 2003). The main objective is to characterize the modeling and diffusion process of news topics in social media, using Big Data as a way of understanding the world in the age of information, where in order to produce more knowledge, more data is necessary (Mukherjee & Shaw, 2016; Ularu et al., 2012). Thus, this philosophy, that has permeated different areas of study, provides us with an adequate structure of thought to approach research questions that seemed unattainable before, although the metrics limit to some extent the scope of the analysis by the very nature of the data; that is, because of its great size and complexity.

For such reasons, this study used Topic Modeling (among other techniques of automated data analysis) and an entire year of publications (54 878 tweets) from the Twitter account of the Colombian newspaper *El Tiempo* (@ElTiempo).

The studies that have addressed the diffusion of news have been centered on the understanding of important news or world-class events (Greenberg, 1964; Henningham, 2000; Rogers & Seidel, 2002). In the 1960's, research pointed to the fact that a news story could take between one or two days to complete the dissemination process, even when having a great spread in the traditional media (Deutschmann & Danielson, 1960). This, without a doubt, has drastically changed with time, and especially with the advent of the Internet and social media.

The concept of news diffusion has been notably influenced by the work of Rogers (2003), for whom the news can be examined from the point of view of prominence, a concept that expresses the degree of importance of a news event as considered by individuals. For this reason, it is understood that out of all communication channels that are spread in the media spectrum, the public chooses and structures the news they consume. With the surge of social media, the study of news diffusion is becoming a field of study that has been developing dramatically as new forms of informing and communicating emerge. This level of complexity is due, as well, to the unpredictable nature of the occurrence of a news topic, combined with its rapid dissemination (Rogers, 2000).

The variety of the news disseminated through the social media makes it so that the study of the diffusion of news topics through long periods of time allows for the modeling of these specific topics within more general topics. In fact, most of the longer-term topics have been found to be labelled as "politics", "business" or "news", and that similar news topics tend to temporally organize themselves into topic chains (Kim & Oh, 2011). Kim & Oh (2011) found that a few "unique topics" or short-duration topics that appear when using the "topic modeling" tool are incoherent (although some of them could represent relevant events such as the death of someone famous or the increase of aviation safety). This is due to the fact that topic modeling is a method that learns topic structures from large collections of documents without human supervision (Arora et al., 2013), a necessary development taking into account that, with all the information available online, we have reached the point that it is humanly impossible to process it all (Blei, 2012).

Zhao et. al. (2011) addressed the issue of Twitter being just another news feed that was faster than traditional news media, tapping into the unsupervised topic modeling, a way of extracting topics (underlying semantic topics), using only the words that were found in a set of documents (Blei & McAuliffe, 2007). In this way, then, each tweet can be associated to a topic, and each topic to a specific category (Zhao et al., 2011). The previous discussions as well as the growing dynamics of news diffusion from traditional media using social media have led us to examine this process in the Twitter account of a media outlet of national scale and ask ourselves:

RQ1a: What are the *news topics* that emerge from the tweets published by the information channel in Twitter from the Colombian newspaper *El Tiempo* during a one-year period?

On the other hand, diverse research works have contributed to establishing the characteristics of the journalistic information transmitted through the Twitter channels and what properties of the messages are responsible for making the users be willing to follow the informational accounts of this social network (Argüelles & Muñoz, 2012; Lotan et al., 2011; Schultz & Sheffer, 2012; Stubbs, 2001; Ure & Parselis, 2013; Wasike, 2013). Besides the formal characteristics of the content found in social media (length, links, etc.), the *tone* or the *sentiment* expressed in the text has been one of the categories that can better characterize the content, as they allow us to elucidate automatically if the messages contain positive, negative or neutral sentiments in their structure (Leetaru, 2012).

There are different types of sentiment analysis (Feldman, 2013), but their general aim is to make a machine able to process and evaluate sentiments (Kechaou, Ammar & Alimi, 2013). Notwithstanding, although limited (Stieglitz & Dang-Xuan, 2013), their use has been propagated in the analysis of various types of content such as blogs, review sites, datasets and microblogging (Vinodhini & Chandrasekaran, 2012) and have diversified (Meena & Prabhakar, 2007; Turney, 2002), unifying it with topic modelling (Cai, Spangler, Chen & Zhang, 2010). The tone or sentiment within a message can be clearly identified as an innovative characteristic or property of the content. Knowing that the characteristics of the innovations are related to the process of diffusion in social media (Peslak, Ceccucci & Sendall, 2010) as well as to any news item or news topic (Rogers, 2003), we ask:

RQ1b: What were the characteristics or *innovative properties* of the news topics disseminated?

RQ2a: Did these characteristics influence the diffusion of news topics?

Social media has created a more complex ecosystem in terms of duration and distribution of news products (Newman, Dutton & Blank,

2012), and therefore debates have been opened on the emotional responses as a function of the diffusion of news (Ibrahim, Ye & Hoffner, 2008), the topics, entities and relationships exposed in news articles and social media (Kang, O'Donovan & Höllerer, 2012; Newman, Chemudugunta, Smyth & Steyvers, 2006), the temporal dynamics of the messages with respect to specific events (Jungherr, 2014) and the acceptance or rejection of the information (Emery, Szczypka, Abril, Kim & Vera, 2014).

One of the challenges in addressing these topics when one works with large scale data set from a long period of time comes from the identification of the diverse stages of the process. That includes the discovery of the topics, finding those that are similar and group them, finding short-term issues within the topics and identifying how these changes as a function of time (Kim & Oh, 2011). In this sense, this research study will try to learn:

RQ1c: How are the news topics disseminated through time?
RQ2b: What was the relationship between the time or moment of production of the message and the appearance of news topics?

Lastly, although there have been many studies on social media with analysis techniques of large quantities of data (Asfari, Hannachi, Bentayeb & Boussaid, 2013; Bogdanov, Busch, Moehlis, Singh & Szymanski, 2013; Gerber, 2014; Ghosh & Guha, 2013; Ferrari, Rosi, Mamei & Zambonelli, 2011; Michelson & Macskassy, 2010; Paul & Dredze, 2014), little has been done to address the importance of the message's authorship. In platforms such as Twitter, the channel could be referring to the author that emits the message, having as a possibility that the message was created by the account holder or by a third party that the account holder simply re-tweets (RT), which can have implications on the dynamics of the production of information. Therefore, it is necessary to determine:

RQ1d: What were the channels or types of message authorship that originated the news topics?
RQ2c: In what way did the channel or type of authorship affect the appearance of news topics?

Method

Sample and procedure

For this quantitative study, a total of 54 878 tweets compiled from February 1st, 2013 to February 1st, 2014 were gathered through an automated script using an Open Standard for Authorization (OAUTH) and the REST APIS of Twitter, which provide programmatic access to the tweets and their metadata in JSON format. All the tweets came from the Twitter account of the Colombian newspaper *El Tiempo* (@ElTiempo), one of the traditional Colombian news outlets, which serve as a reference in the Latin American context. For this task, two previously-trained coders downloaded and stored the complete message texts as well as the date and hour (created_at), user (user.screen_name), retweet (RT) and external link (URL).

Measurements

With the aim of collecting and processing the data, a series of variables were adapted ad hoc starting with the traditional model by Rogers (2003). This would allow us to describe a set of indicators inherent in the dynamics of a social network and establish relationships between them, to be able to answer the research questions posited. More specifically, the following variables were measured:

1. Innovation: the "innovation's content" (it refers to the news topic, its sub-categories were inductively extracted thanks to computer modeling) and the "innovation's properties" (it refers to the message characteristics or properties). On the other hand, the "tone" or "message sentiment" was analyzed. Its scale was measured with the following scores: Positive or P+ (+1), Negative or N+ (-1) and Neutral or NEU (0). Additionally, formal indicators such as number of characters, number of words, number of links and number of mentions per tweet were extracted.
2. Time: it refers to the moment of production of the tweet, meaning the day and the time of day the message was first published.
3. Channel: it refers to the authorship of the message. In this study, the channel was defined as the source from which the message

came from (authorship), meaning, if the message was produced by the own medium (0) or if it was a retweet (or RT) from another account (1).

Data analysis

For the analysis, Big Data analysis tools were used, such as automated text codifying, Sentiment Analysis and computational tools based on Machine Learning such as Unsupervised Topic Modeling, which allowed us to model the underlying topics in a set of unstructured data (Arcila-Calderón, Barbosa-Caro & Cabezuelo-Lorenzo, 2016). Also, the procedures that were expected to be used from the questions asked were rapidly and efficiently executed. These technologies allowed us to cover the totality of the collected without having to select a sample. In addition, with this the processes expected to answer the research questions (RQs) were executed quickly and efficiently. Due to the number of messages gathered and especially due to the analysis techniques used, these tools were essential for the performing of the study.

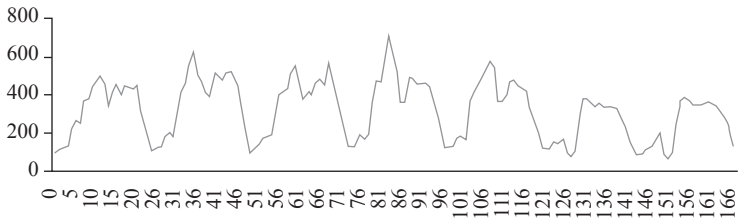
The software used to extract information from the data were: 1) Topic Modeling Tool (a tool developed by The Stanford Natural Language Processing Group which is based on Latent Dirichlet Allocation), which was used for the modeling of topics to obtain, from the same data, diverse groups of terms which were associated to specific topics; 2) Textalytics for Excel, for the automated analysis of sentiment (based on dictionaries); 3) Microsoft Excel, which was used for the counting of frequencies and automated codification; 4) SPSS Statistics and R, for cross-tabulation of variables, regressions and other statistical operations; 5) QDA Miner, a package that allow us to obtain the frequency of keywords in the corpus, the most used hashtags, the most used phrases and keywords in context.

FINDINGS

The number of tweets emitted by the newspaper *El Tiempo* in Twitter in the year 2013 (N= 54 878) was lowest in the months of June and August (2 544 and 4 133, respectively), and highest in the month of

February (5 266 messages) (January= 5 122, March= 4 414, April= 5 085, May= 4 722, July= 4 557, September= 4 280, October= 5 249, November= 4 769, December= 4 501). From June 9th to 20th, there were no tweets found due to a technical error in the application used to capture the messages from the @ElTiempo account, which explains the atypical results in this month.

FIGURE 1
TOTAL NUMBER OF TWEETS PER HOUR DURING THE WEEK.
CUMULATIVE FOR 2013 PROFILE @ELTIEMPO



Source: the authors.

The three days of the week with the highest number of tweets (emitted by @ElTiempo in 2013) were Thursday (n=9 155), Tuesday (n=9 114) and Wednesday (n=8 675). Figure 1 shows the visualization of the diffusion of tweets throughout an average week in 2013, considering that Monday starts at hour 0 and Sunday ends at hour 166. We can see that each day of the week had a similar behavior, with a peak of message emission around 10:00 in the morning, and a notable decrease starting at 8:00 in the evening. These results allowed us to clearly see that the diffusion process corresponded to the working days of the outlet.

To answer the question of which news topics emerged from the @ElTiempo Twitter channel in 2013 and how they were diffused (RQ1a), the underlying topics were modeled with unsupervised topic modeling, by which keywords were repeatedly clustered, and then the predominance of each topic in each message was determined:

- Topic 1: Venezuela/International. This topic encompasses news items that are focused on international topics, and the tweets that refer to Venezuela (politically and militarily), directly or indirectly. They possess keywords such as "Venezuela", "Chávez", "Maduro" and "General". An example of this topic is the following tweet: "El presidente Chávez está en Venezuela en una batalla para recuperarse" [President Chávez is in Venezuela in a battle to recover]: *@VillegasPoljakE a @WRadioColombia*.
- Topic 2: Sports/Entertainment/General. In this classification we find the tweets that address topics related to sports and the world of entertainment. There are also news items of less importance, although with less of a presence. Some of the keywords that can be highlighted are "mundial" [worldwide], "Colombia", "liga" [league], "final", "copa" [cup], "tiempo" [time], "partido" [match], "gol" [goal], "Falcao", "Medellín" and "colombiano" [Colombian]. A tweet that exemplifies this topic is: "Todos al acecho de Santa Fe en el grupo A de la Liga" [Everyone lurking Santa Fe in league group A].
- Topic 3: Politics/National Interest/Conflict. This topic groups all the tweets that refer to situations involving politics, war or armed conflict. It is delimited by keywords such as "Gobierno" [government], "FARC", "paro" [strike], "política" [politics], "paz" [peace], "proceso" [process], "ataque" [atack], "libertad" [freedom] and "Santos". A representative tweet of this grouping could be: "El balance de las Farc y el Gobierno de siete meses de #DiálogosDePaz" [The seven-month balance of Farc and the government].
- Topic 4: Residual topics. In this topic we find the tweets that could not be clearly associated with any of the other three groups.

The results showed 39.1% of the tweets were mainly related to the topic Politics/National Interest/Conflict, which is composed of themes such as armed conflict, violence, national events of great scale and incidents corresponding to political figures and governmental institutions. In second place, we found messages associated to the topic Sports/Entertainment/General (27.7%), followed by Venezuela/

International (10.3%) and Residual topics (23%). In this last classification, messages could refer to localized social issues, gender, "isolated" news or other topics that, due to their nature and frequency, could not be unequivocally associated with any of the major themes identified in the corpus. In this group we found examples like the following:

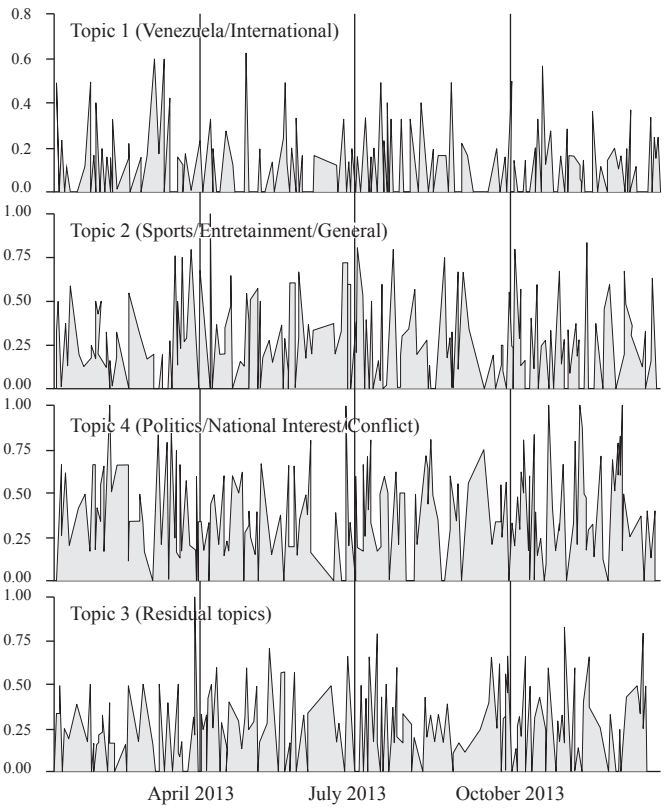
- Mujer inglesa está dispuesta a todo por convertirse en parapléjica [British woman is willing to do anything to become paraplegic].
- De que sirve enamorarse [What is the avail of falling in love].
- Si no dice matrimonio no lo vamos a aceptar: pareja LGBTI [We won't accept it if it doesn't say marriage].

Visualizations of how these topics were spread in time in the year 2013 (RQ1c) are shown in Figure 2. These time series clearly show that there is no underlying regularity in the appearance of topics, so we cannot speak of a seasonal series, which suggests that other characteristics –different to temporality itself– are responsible for the variation in each one of the four topics.

With the aim of describing innovative characteristics or properties of the news topics disseminated in the year 2013 (RQ1b), the tone of the message as well as its formal aspects were identified. In the first place, we found that an important number of the messages containing some of the four news topics did not possess indications of subjectivity. This means that they could be classified as neutral (39.2%). However, a clear sentiment in the content was detected in most of them, with 19.8% of the messages having a predominantly positive tone, and 19.7% having a predominantly negative tone.

Likewise, the formal characteristics of the messages were analyzed, and the words that were most used in the news topics were found. Among these, there is a clear predominance of the words "Colombia" (2 971 cases) and "Bogotá" (2 391 cases) in the corpus of the tweets. These were listed along with words that referred to conflict, delinquency-related events or violence ("paz" [peace], "gobierno" [government], "FARC", "presidente" [president], "policia" [police], "Santos"). It should be noted that the presence of the words "mundial" [worldwide]

FIGURE 2
TEMPORAL EVOLUTION OF TOPICS IN THE YEAR 2013



Source: the authors.

(723 cases) and "mujer" [woman] (482 cases) were also found among the most used words. When looking at the average number of characters (M= 93.91, SD= 21.25), we can see that the messages did not use the maximum number allowed at that moment (140 characters), using an average of 12 words per tweet (M= 11.65, SD= 3.98). The data reveal that most of the messages included links (M= 0.88, SD= 0.35) to other websites, and in lesser number, mentions to other users (M= 0.35, SD= 0.63) and labels or hashtags (M= 0.33, SD= 0.54).

An analysis of the mentions (@) found in the messages revealed that the most-mentioned profiles (@Portafolioco, @ElTiempo, @FUTBOLRED, @CityTV) belonged to the same publisher as the newspaper (Casa Editorial *El Tiempo*). Likewise, it was observed that the most-mentioned public figure accounts were @JUANMANSANTOS, @PETROGUSTAVO, @FALCAO and @BARACKOBAMA. To understand the type of authorship or channel used to disseminate the news topics (RQ1d), an analysis of the retweets (RT) was performed on the @ElTiempo profile. We found that 18.6% of the news topics disseminated all through the year 2013 originated from sources other than the account holder (retweets), while the percentage of their own tweets was 81.4%.

Next, with the objective of verifying the relationship between the appearance of the news topics and: a) the innovative characteristics or properties of the message referring to the tone as well as formal aspects (RQ2a); b) the time or moment of production (RQ2b); and c) the channel or authorship of the message (RQ2c), a series of linear regressions were performed. We verified the absence of multicollinearity (with tolerance values close to 1 and the variance inflation factors (VIF) below 5). The results are summarized in Table 1.

The analysis shows that only regression models for topics 2 ($F(4.54872)= 84.483$, $p< .000$), 3 ($F(4.54872)= 11.731$, $p< .000$) and 4($F(4.54872)= 44.187$, $p< .000$) can explain part of the variance of the appearance of news items as a whole, although that explained variance is low (6%, 1% and 3%, respectively). Notwithstanding the data reveal that the innovative properties of the message, meaning their formal characteristics and tone, were linked to the dissemination of topics 2, 3 and 4 with low but statistically significant coefficients. Similarly, we see that time or moment of production was a determinant factor in the same topics. In the case of message authorship (RT or not), we observed that it only explained the dissemination of news topics 3 and 4. Even though the influence of these factors is low, the fact that they appear to model the appearance of almost all the news topics is significant. This suggests that even if the link is weak, the diffusion of news topics is not fortuitous, but responds to these and other patterns that should be studied.

TABLE 1
LINEAR REGRESSIONS OF NEWS TOPIC

	Topic 1		Topic 2		Topic 3		Topic 4	
	B	β	B	β	B	β	B	β
Formal characteristics+	.000	-.004	-.002***	-.021***	.001	.007	.001***	.017***
Tone	.003	.008	.019***	.043***	-.016***	-.034***	-.005*	-.012*
Time	.000	-.004	.009***	.061***	-.006***	-.040***	-.002***	-.015***
RT or not	-.003	-.008	.003	.005	.011***	.019***	-.011***	-.021***

+ Average number of characters, number of words, number of links and number of mentions per tweet.
*p < .05; **p < .01; ***p < .001.
Source: The authors.

Discussion and conclusion

In this study we have examined the dissemination of news topics on Twitter based on the modeling of underlying patterns in a longitudinal corpus of informative messages of the social account of a communication outlet, which has allowed us to explore the variables that influence the appearance of themes. The search for patterns can be applied both in social media and in online media, making it possible to explore media agenda problems and, even, to manifest certain practices that are difficult to attend, such as "fake news" or false news, since the algorithms of detection of underlying topics can be modeled with different approaches.

According to the results of the study, the topic Politics/National Interest/Conflict was found in the largest part of the messages published in the Twitter account of the Colombian newspaper *El Tiempo* in the year 2013. If we examine the most important news events in Colombia in 2013 (such as the start of peace talks between Santos' Government and the Colombian Revolutionary Armed Forces (FARC-EP) at the end of 2012 and the elimination rounds of the FIFA World Cup Brazil 2014), we see that they correspond, besides the topics found, with keywords (peace, government, FARC, president, police, Santos, against, world and worldwide, among others).

The data suggests that there is a relationship between the most-mentioned profiles, the more frequent keywords and the main underlying news topics (topics 1 and 2) found with the topic modeling algorithm, as the results of all these items fit within the categories politics or sports.

When analyzing the number of characters as an innovative property, it is important to note that the tweets that use all of the space provided by the platform (140 characters) are merely 4.5%, a total of 2 495 cases. This could suggest that the tendency of the outlets is to publish messages that are relatively short, summarizing information using just a few words, and giving concise answers to other users of the platform. We could also highlight the fact that 71.4% of the tweets lacked mentions, and that 86.5% had links. The first could suggest a low degree of interaction with other users in the platform or the possibility that the other people/entities named in the messages did not have a Twitter account, and the second could point to a high rate of re-directing from this platform.

The range of published tweets varied between 4 000 and 5 000 (per month), evidencing a notable diminishing in the production of tweets on Saturdays (5 585) and Sundays (5 873) with respect to the other days of the week. This situation changed with the increase shown on Mondays (there were 2 136 more tweets than on Sunday), which could be linked to journalistic dynamics and the volume of information generated during the weekend. Previous literature has suggested that time lapses (as far as social and mental constructs) help to create a reality (Stubbs, 2001). Therefore, the days of the week could be associated with certain types of messages or emotions that are transmitted, by a person or, in this case, a medium of news diffusion (relaxation or sports associated to Sundays, leisure with the other weekend days, etc.).

On the other hand, the low frequency of retweets (18.6%) indicated low levels of content replication. This means that the newspaper generates scarce moments of visibility for messages from other accounts. While maintaining a high percentage of originality of its publications, the account studied redirected the users to their website or other online resources, turning the external websites into the frame of reference for reading their tweets. Then, even if their own messages are short, they provide more information that is hosted in other domains. Despite this, we can highlight that 70% of the tweets did not have hashtags. This limits in some way the inclusion of the messages in conversations different from those generated by the news diffusion medium.

Although the statistical models used only explained little variance found in the news topics, we can highlight that the innovative properties or characteristics and the time or moment of production of the message appeared to be significant predictors in 3 out of 4 news topics modeled. The channel or message authorship appeared to be a significant predictor in only two topics, one of them referring to political and conflict subjects, and the other to residual topics. This association could be due to the fact that generally, in order to report on political or conflict subjects, the sources (defined as channels in this study) are usually cited verbatim to back the information. This could invite us to think about using the messages from other profiles (i.e. political figures) to communicate this type of news events. However, as these data were not consistent enough, we believe that other complementary analyses are necessary.

Considering what was presented previously, we can argue that these results shed light, using the theory proposed by Rogers (2003) and the advanced computational methods, on some characteristic elements of the diffusion of news in social media. Also, the automated analysis of content methods used, topic modeling and the automated sentiment analysis, among others, were shown to be convenient for this kind of work, by becoming a theoretical-practical support and being helpful in the study of journalistic dynamics. Having found a similarity in numbers between messages classified as positive and negative, as well as a good part of the corpus labeled as neutral, we can infer that there is a distance in terms of linguistic subjectivity in the studied messages. This implies that the media outlet, at least in its Twitter profile, avoids adjectives when reporting events.

One of the limitations of the study was the lack of data from June 9th to June 20th, 2013, caused by a technical problem in the application that was used to compile the tweets. This could have caused a slight bias when observing that month's data, but we must consider that it is impossible to recover the messages that were lost in that timeline, unless the source (in this case *El Tiempo* newspaper) provides them. Another limitation was the scarcity of previous research on Latin America that could have had similar characteristics such as the size of the sample as well as the techniques used. This is the reason why the comparison of empirical data with other equivalent explorations was not possible.

Besides what was already mentioned, we can add that as Twitter is a platform where the messages have a short lifetime (as the oldest messages disappear as new tweets are published). Another limitation was the gathering of an archive that encompassed a longer timeframe, as this implies having open access to the content stored by the owners of the profiles. Lastly, and with respect to the theory by Rogers (2003) in our research, a limiting factor appeared when we could not include the social system as a category of analysis, due to the characteristics of Twitter and the perspective we used.

In future research studies, it would be relevant to ask about the inner workings of the process of adoption of these news topics in social media from the perspective of the subject, how short-term news topics appear and disappear as a function of time, and what models could be

created to try to explain the variance of other message characteristics from the messages found in platforms such as Twitter.

Bibliographic references

Arcila-Calderón, C., Barbosa-Caro, E. & Cabezuelo-Lorenzo, F. (2016). Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. *EPI, El Profesional de la Información, 25*(4), 623-631.

Argüelles, I. & Muñoz, A. (2012). An insight into Twitter: A corpus based contrastive study in English and Spanish. *Revista de Lingüística y Lenguas Aplicadas*, *7*, 37-50.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag. D., Wu, Y. & Zhu, M. (2013). Practical algorithm for topic modeling with Provable Guarantees. *30th International Conference on Machine Learning (ICML)*, *28*(2), 280-288. Atlanta, United States.

Asfari, O., Hannachi, L., Bentayeb, F. & Boussaid, O. (2013). Ontological topic modeling to extract Twitter users' topics of interest. *8th International Conference on Information Technology and Applications (ICITA)* (pp. 141-146). Sydney, Australia.

Blei, D. (2012). Topic models and digital humanities. *Journal of Digital Humanities*, *2*(1), 8-11.

Blei, D. & McAuliffe, J. (2007). Supervised topic models. *Neural Information Processing Systems*, *20*, 1-8.

Bogdanov, P., Busch, M., Moehlis, J., Singh, A. K. & Szymanski, B. K. (2013). The social media genome: Modeling individual topic-specific behavior in social media. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM* (pp. 236-242). Niagara Falls, Canada.

Caballero, U. (2001). Periódicos mexicanos en Internet. *Revista Universidad de Guadalajara, 22*(46).

Cai, K., Spangler, S., Chen, Y. & Zhang, L. (2010). Leveraging sentiment analysis for topic detection. *Web Intelligence and Agent Systems: An International Journal*, *8*, 291-302.

Deutschmann, P. & Danielson, W. (1960). Diffusion of knowledge of the major news story. *Journalism Quarterly*, *37*(3), 345-355.

Emery, S., Szczypka, G., Abril, E., Kim, Y. & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the tips campaign. *Journal of Communication*, *64*(2), 278-295.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89. DOI: https://doi.org/10.1145/2436256.2436274

Ferrari, L., Rosi, A., Mamei, M. & Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. *Proceedings of the 3rd ACM Sigspatial international workshop on location-based social network* (pp. 9-16). Chicago: ACM.

García de Torres, E., Rodrigues, J., Saiz, J., Albacar, H., Ruiz, S. & Martínez, S. (2008). Las herramientas 2.0 en los diarios españoles 2006-2008: tendencias. *prisma.com*, *7*, 193-222.

Gerber, M. (2014). Predicting crime using Twitter and Kernel Density estimation. *Decision Support Systems*, *61*, 115-125.

Ghosh, D. & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, *40*(2), 90-102.

Greenberg, B. (1964). Diffusion of news of the Kennedy assassination. *Public Opinion Quarterly*, *28*(2), 225-232.

Henningham, J. (2000). The death of Diana: An Australian news diffusion study. *Australian Journalism Review*, *22*(2), 23-33.

Ibrahim, A., Ye, J. & Hoffner, C. (2008). Diffusion of news of the shuttle Columbia disaster: The role of emotional responses and motives for interpersonal communication. *Communication Research Reports*, *25*(2), 91-101.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239-259. DOI: https://doi.org/10.1111/jcom.12087

Kang, B., O'Donovan, J. & Höllerer, T. (2012). Modeling topic specific credibility in Twitter. *IUI'12 Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (pp. 179-188). Lisbon, Portugal.

Kechaou, Z., Ammar, M. & Alimi, A. (2013). A multi-agent based system for sentiment analysis of user-generated content. *International Journal on Artificial Intelligence Tools*, *22*(2), 1-28.

Kim, D. & Oh, A. (2011). Topic chains for understanding a news corpus. *CICLING'11 Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing-Volume Part II* (pp. 163-176). Tokyo, Japan.

Lasorsa, D., Lewis, S. & Holton, A. (2012). Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism Studies*, *13*(1), 19-36.

Leetaru, K. (2012). *Data mining methods for the content analyst. An introduction to the computational analysis of content*. New York: Routledge.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. & Boyd, D. (2011). The revolutions were Tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, *5*, 1375-1405.

Meena, A. & Prabhakar, T. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In A. Amati, C. Carpineto & G. Romano (Ed.), *Advances in Information Retrieval. ECIR 2007. Lecture Notes in Computer Science* (vol. 4425). Berlin: Springer.

Michelson, M. & Macskassy, S. (2010). Discovering users' topics of interest on Twitter: A first look. *AND'10 Proceedings of the fourth workshop on analytics for noisy unstructured text data* (pp. 73-80). Toronto, Canada.

Micó, J. L., Canavilhas, J., Masip, P. & Ruiz, C. (2008). La ética en el ejercicio del periodismo: credibilidad y autorregulación en la era del periodismo en Internet. *Estudos em Comunicação*, *4*, 15-39.

Mukherjee, S. & Shaw, R. (2016). Big data. Concepts, applications, challenges and future scope. International *Journal of Advanced Research in Computer and Communication Engineering*, *5*(2), 66-74.

Newman, D., Chemudugunta, C., Smyth, P. & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *ISI'06 Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics* (pp. 93-104). San Diego, United States.

Newman, N., Dutton, W. & Blank, G. (2012). Social media in the changing ecology of news: The fourth and fifth states in Britain. *International Journal of Internet Science*, *7*(1), 6-22.

Paul, M. & Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS one*, *9*(8), e103408. DOI: https://doi.org/10.1371/journal.pone.0103408

Peslak, A., Ceccucci, W. & Sendall, P. (2010). An empirical study of social networking behavior using diffusion of innovation theory. *Conference on Information Systems Applied Research 2010 CONISAR Proceedings.* Nashville, United States.

Rogers, E. (2000). Reflections on news event diffusion research. *Journalism & Mass Communication Quarterly*, *77*(3), 561-576.

Rogers, E. (2003). *Diffusion of innovations*. New York: Free Press.

Rogers, E. & Seidel, N. (2002). Diffusion of news of the terrorist attacks of September 11, 2001. *Prometheus*, *20*(3), 209-219.

Said-Hung, E., Serrano, A., García, E., Calderín, M., Rost, A., Arcila, C., Yezers'ka, L., Edo, C., Rojano, M., Jerónimo, P. & Sánchez, J. (2013). Ibero-American online news managers' goals and handicaps in managing social media. *Television and New Media*, *4*(2).

Schultz, B. & Sheffer, M. L. (2012). New brand: The rise of the independent reporter through social media. *Online Journal of Communication and Media Technologies*, *2*(3), 93-112.

Stieglitz, S. & Dang-Xuan, L. (2013). Emotions and information diffusion in social media-Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217-247.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell: Oxford.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 417-424) Philadelphia, United States.

Ularu, E., Puican, F., Apostu, A. & Velicanu, M. (2012). Perspectives on big data and big data analytics. *Database Systems Journal*, *3*(4), 3-14.

Ure, M. & Parselis, M. (2013). Argentine media and journalists enhancing and polluting of communication on Twitter. *International Journal of Communication*, *7*, 1784-1800.

Vinodhini, G. & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(6), 282-292.

Wasike, B. (2013). Framing news in 140 characters: How social media editors frame the news and interact with audiences via Twitter. *Global Media Journal-Canadian Edition*, *6*(1), 5-23.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E., Yan, H. & Li, X. (2011). Comparing Twitter and traditional media using topic models. *Advances in Information Retrieval: 33rd European Conference on IR Research - ECIR, 2011*. Dublin, Ireland.