Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Method for Collecting Relevant Topics from Twitter supported by Big Data

To cite this article: Jesús Silva *et al* 2020 *J. Phys.: Conf. Ser.* **1432** 012094

Method for Collecting Relevant Topics from Twitter supported by Big Data

Jesús Silva[1], Alexa Senior Naveda[2], Ramiro Gamboa Suarez[3], Hugo Hernández Palma[4], William Niebles Núñez[5]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2] Universidad de la Costa, Barranquilla, Atlántico, Colombia
[3]Universidad Surcolombiana, Neiva, Huila, Colombia
[4] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[5]Universidad del Sucre, Sincelejo, Sucre, Colombia.

Email: jesussilvaUPC@gmail.com

Abstract. There is a fast increase of information and data generation in virtual environments due to microblogging sites such as Twitter, a social network that produces an average of 8,000 tweets per second, and up to 550 million tweets per day. That's why this and many other social networks are overloaded with content, making it difficult for users to identify information topics because of the large number of tweets related to different issues. Due to the uncertainty that harms users who created the content, this study proposes a method for inferring the most representative topics that occurred in a time period of 1 day through the selection of user profiles who are experts in sports and politics. It is calculated considering the number of times this topic was mentioned by experts in their timelines. This experiment included a dataset extracted from Twitter, which contains 10,750 tweets related to sports and 8,758 tweets related to politics. All tweets were obtained from user timelines selected by the researchers, who were considered experts in their respective subjects due to the content of their tweets. The results show that the effective selection of users, together with the index of relevance implemented for the topics, can help to more easily find important topics in both sport and politics.

1. Introduction

Microblogging services, especially Twitter, offer a dual function: they are microphones for the masses [1] and serve as sources of information that complements traditional media. This phenomenon has generated a new ecology in news consumption, in which 'Social Media' platforms are becoming increasingly important [2]. Specifically on Twitter, every time users log into their account, they come across a post feed that results from the combination of all the content produced by the users they are following. These users who are followed, named 'followees', can be friends, figures of local or regional relevance, companies, or news networks. As the number of 'followees' increases, the amount of information that the user must process also increases and a problem of information overload is generated [3].

Previous studies have extensively analyzed the problem of 'authorities' user' on Twitter [4,5], referring to the ability of a user to produce valuable information for his followers where the value is measured by metrics such as the number of re-tweets received to a given tweet or the amount of debate generated by it. However, news consumers facing the problem of information overload often have no problem identifying users who constitute 'authorities' on national issues.

The challenges these users face is more related to the ability to get a summary of their news feed where it is possible to automatically sort posts by relevance. In this paper, a method is introduced to convert a tweet feed into a set of topics with an associated score to identify its relevance. Topics in the context of natural language processing are a list of words that summarize an event or news series. The process begins with the selection of expert user profiles in each of the areas that would be analyzed taking Sports and Politics as case studies. Data filtering and pre-processing continues. The next stage is the collection of topics using a probabilistic model. Finally, as a contribution of this study, a quantitative metrics was designed to identify the relevant topics contained in a set of tweets selected as input.

2. Previews Studies

One of the traditional algorithms used in the inference of topics is LDA [6]. Its proven efficiency has made this method the basis of many studies in the detection of topics. The main feature of this method is the easy interpretation of the results, which are sets of words likely to belong to discovered topics. So far, most of the work is focused on creating dynamic applications [7] that collect information online and present topics at their output. Another current of research is the formulation of methods and special pre-processing of data [8] to improve the results and execution of LDA.

All the works described above present the common use of documents longer than 140 characters, which is a restriction for this study given the use of Twitter platform. Although this topic is discussed in [9], there is not a defined framework that employs standard methods such as LDA in the detection of topics with the restriction of document size. So, [10] propose improvements to the LDA output, which returns relevant topics associated with topics that have been specified by the researcher. However, this research looks for a simple way to identify if a topic is relevant with respect to all those obtained at the LDA output, just based on the input documents (tweets) of the process. For this reason, the researchers followed known guidelines in word processing and finally introduced an index for helping discern the importance of a topic with respect to all those generated by LDA.

3. Method

3.1 Users´ selection

For this research, a dataset of 10,750 sports tweets and 7,758 politics tweets were used, which were obtained from a group of Twitter users who served as a seed group. A list of 50 users was formed for each of the topics. By expert judgement, 30 users were chosen from each group whose profiles meet requirements such as constant activity, content focused on the area to which they belong, and quality in writing.

Among the seed users chosen for the Politics issue, there are some politics scientists, government and private sector workers, important political characters such as ex-candidates or current public officials who hold leadership positions in institutions such as mayors´ offices, ministries and the presidency. The seed users chosen for Sports include sport commentators, administrators and presidents of football clubs, professional players and users related to sport programs or magazines. For each of the selected users, a set of tweets was extracted for the time period covered in the study, as well as the date and time of the tweet.

3.2 Pre-selection and dataset processing

The collection of tweets was carried out using Twitter REST-API together with Python, starting by defining two specific topics (sports and politics) for testing the method, concluding that the topics mentioned above are completely exclusive and cover a large amount of data produced by twitter users. As mentioned by [11], the "Sports" category encompasses most of the topics within the social network, followed by the categories: "Other", "Other news", "Music", "Tv & Movies", "Tech" and "Politics" as the most relevant. Although politics is not on the top of the list, it was chosen because of the national panorama and the events of political nature that are occurring in the country. Considering that, as the experience shows, the events that include popular expressions of people are always reflected in social networks, the study seeks to verify specific facts (topics) emerging with the analysis of the tweets.

Once the expert users in both topics were defined, the researchers proceeded to collect the maximum number of tweets per user that REST Twitter API allows to gather in a single query. The collection is done by making a query per user, obtaining 10,750 and 7,758 tweets of users related to Sports and Politics respectively.

Finally, the JSON is analyzed and the tweet is extracted, eliminating line breaks, punctuation marks and correcting sequences of characters that are relevant for the study. As an example: all chains with sequences of 'ja' or 'ha' are mapped to 'jaja', so 'jajaja', 'jajajaja' or 'hahaha' are translated as 'jaja'.

In addition, hashtags are separated by the capital letter (standard writing within Twitter), so the string #vamosMiSelection was translated as 'vamos mi selección', and in case of not finding capital letters, it will remain the same. This process is based on the study of Mitchell et al. [12], in which important words are identified for consideration by researchers and corrected or eliminated to obtain greater accuracy in the execution of the algorithms. In addition to the tweet, the username and timestamp are extracted from each tweet and everything is stored for later analysis in a database.

3.3 Text simplification

The task of simplifying the text was carried out with the use of two libraries: NLTK Steven Bird et al. [13] for the elimination of stopwords and the software provided by [14] which includes methods for detecting language forms (pronouns, nouns, verbs, etc.) within a sentence and also for the simplification of nouns to singulars and verbs to their most basic form, which is important since words such as 'playing' or 'played' are reduced to the verb 'play', or nouns such as 'fans' to 'fan', avoiding further processing and significantly improving the work of algorithms such as LDA.

3.4 Description of parameters for the LDA process

In Figure 1, M represents the number of documents of the collection and N the number of words in each document, α is the Dirichlet parameter of the distribution of topics per document and β is the Dirichlet parameter of the distribution of words by topic. In this study, the values of 0.1 and 0 are used in all executions for α and β respectively. The theory about LDA [15] establishes that, for small values of α, the content of the documents is restricted, which means that each document (tweet) is a mixture of a few topics or even just one, as well as small values of β indicate that each topic is a mixture of a few words, which is an optimal fit when working with documents of no more than 140 characters and with a high probability that each document (tweet) deals only with one topic.

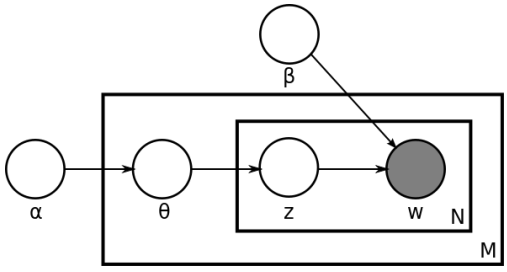

Figure. 1. Representation of the LDA model, where M represents the number of documents and N represents the number of words per document [16]

4. Results and discussions

4.1 Evaluation of document configuration

As explained before, two configurations were made to the LDA input documents, the first one for making each tweet a document and the second one for making each document a group of five consecutive tweets in the timeline of the same user. As a result of the algorithm execution, very similar results are shown, although the execution of LDA with the configuration of grouped tweets show more diffuse results for certain topics, including words that could be easily left out of the topic. For example,

for the day 21 of October of 2018, the most relevant topics for both configurations according to the index are:

- Configuration a document-a tweet: guaira South American Catholic cup match.
- Configuration a document-five tweets: South American cup first catholic guaira.

As noted, the first word in the second configuration does not represent a major meaning for the topic. Despite this, an advantage of grouping by consecutive tweets is the decrease in the impact of repetitive spam tweets, since when they are grouped in the same document, the impact of those words for a topic decreases. Finally, Table 1 shows the results of the LDA execution for 12 consecutive days with the configuration of a tweet-a document using the Sports topic. It is worth mentioning that the LDA algorithm was executed for the days in which the documents exceeded a limit value of 55 due to the fact that below this value the resulting LDA topics are not coherent.

Table 1. Topics generated by LDA for the configuration of a tweet-a document.

Topics per day	**Table of Contents**	**Holder**
Day: 2018-10-17		
guaira south american catholic cup match	55.42	T1
change being duty soccer go out	6.22	
time minute guaira second uniandes	33.47	T1
Day: 2018-10-17		
bsc barcelona almada be able to go	98.57	T2
video colombia today madrid match	14.33	
united manchester video champion bruja	14.12	T3
Día: 2018-10-21		

4.2 Classification by topics

During the execution of the algorithms for identifying topics as required by LDA, as part of the input parameters of the number of topics executed by LDA with this parameter established in 5, 7 and 10, the next step was to compare the results, deciding that 7 is an ideal number given the grouping by days. Establishing the topics to 10 provides many irrelevant topics as a result and, on the other hand, establishing it to 5 produces diffuse topics, as combinations are clearly observed between topics.

Next, the results of the analysis with the respective index of relevance for each topic was exposed, establishing 7 as the maximum number of topics for LDA. Table 1 shows the result of the LDA execution process, together with the relevance index for each topic. It is worth mentioning that the topics (three per day) are the three classified ones with the highest index of the seven that are obtained in the LDA output. Repeated headlines are displayed among some of the topics because they refer to the same event, and some topics are also shown in blue color for representing events that occurred on those days but El País newspaper page does not reflect them with a headline similar to the obtained topic.

5. Conclusions

This research proposes a simple method that can be included in any application that requires an exploration by topics. Once the seed users were chosen and the topics were extracted, it was observed that using a simple to calculate index it is possible to determine the most relevant topics by using LDA. This is confirmed by the fact that most of the resulting topics could have been associated with El País

newspaper headlines issued for the same days on which the method was executed. The ease in the calculation of this index and its proven efficiency is evident both for the resulting topics in the subjects of Sports and Politics. The discrimination of the not resulting topics is observed coherent since topics with words that together result in too ambiguous topics are eliminated.

Finally, two keys for improving this process are the careful selection of expert users in each of the subjects to be analyzed and the careful processing, such as the elimination of stopwords, simplification of language as a reduction of verbs to their basic form and singularization of nouns, processes that were applied in the present study but that can be even improved.

The improvement in the detection of spam at the initial stage of the process is expected to provide excellent results with greater precision in the terms that make up the topic. Additionally, future researchs can apply the index of relevance of each topic and measure its change with respect to time zone for identifying the when a topic reached greater relevance. So, given that the method accepts, as input, users of the Twitter social network, it can establish and test its efficiency in more diverse scenarios and not just aimed at revealing topics such as politics and sports.

References

[1] Amelec, V., & Carmen, V. (2015). Relationship Between Variables of Performance Social and Financial of Microfinance Institutions. *Advanced Science Letters*, *21*(6), 1931-1934.

[2] Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[3] Guyon, I., Elisseeff, A., An introduction to variable and feature selection, Journal of machine learning research, 3, 2003, pp. 1157-1182.

[4] Kohavi, R., John, G., Wrappers for feature subset selection, Artificial Intelligence Journal, Special issue on relevance, 1997, pp. 273-324.

[5] Abdul Masud, M., Zhexue Huang, J., Wei, C., Wang, J., Khan, I., Zhong, M.: Inice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres. Inf. Sci. (2018). https://doi.org/10.1016/j.ins.2018.07.034

[6] Nic Newman, William H Dutton, Grant Blank: Social media in the changing ecology of news: The fourth and fifth estates in britain. InternationalJournalofInternetScience,7(1):6–22, 2012.

[7] Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[8] Avery E Holton Hsiang Iris Chyi: News and the overloaded consumer:Factors influencing information overload among news consumers. Cyberpsychology, Behavior, and Social Networking, 15(11):619–624, 2012.

[9] Eytan Bakshy, Jake M Hofman, Winter A Mason, Duncan J Watts: Identifying influencers on twitter. In Fourth ACM International Conference on Web Seach and Data Mining (WSDM), 2011.

[10] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary: Twitter trending topic classification. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pages 251– 258. IEEE, 2011.

[11] Leite, R., Brazdil, P., Decisión tree-based attribute selection via subsampling, Workshop de minería de datos y aprendizaje, VIII Iberamia, Sevilla, Spain, Nov, 2002, pp. 77-83.

[12] Piramuthu, S., Evaluating feature selection methods for learning in data mining applications, Proc. 31st annual Hawaii Int. conf. on system sciences, 1998, pp. 294-301.

[13] Liangjie Hong, Brian D Davison: Empirical study of topic modeling in twitter. In Proceedings of the First Workshop on Social Media Analytics, pages 80–88. ACM, 2010.

[14] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 569–577. ACM, 2008.

[15] Kira, K., Rendell, L., The feature selection problem: traditional methods and a new algorithm, Tenth nat. conf. on AI, MIT Press, 1992, pp. 129-134.
[16] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. Indian Journal Of Science And Technology, 9(47). doi:10.17485/ijst/2016/v9i47/107371