Using machine learning and information visualisation for discovering latent topics in Twitter news

Vladimir Vargas-Calderón[1][0000−0001−5476−3300], Marlon Steibeck Dominguez[2], N. Parra-A.[1][0000−0002−1829−4399], Herbert Vinck-Posada[1], and Jorge E. Camargo[3]

[1] Grupo de Superconductividad y Nanotecnología, Departamento de Física, Universidad Nacional de Colombia, AA 055051, Bogotá, Colombia {vvargasc, nparraa, hvicnkp}@unal.edu.co
[2] Facultad de Ingeniería, Unipanamericana Fundación Universitaria, 111321, Bogotá, Colombia madominguez@unipanamericana.edu.co
[3] Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, AA 055051, Bogotá, Colombia jecamargom@unal.edu.co

Abstract. We propose a method to discover latent topics and visualise large collections of tweets for easy identification and interpretation of topics, and exemplify its use with tweets from a Colombian mass media giant in the period 2014–2019. The latent topic analysis is performed in two ways: with the training of a Latent Dirichlet Allocation model, and with the combination of the FastText unsupervised model to represent tweets as vectors and the implementation of K-means clustering to group tweets into topics. Using a classification task, we found that people respond differently according to the various news topics. The classification tasks consists on the following: given a reply to a news tweet, we train a supervised algorithm to predict the topic of the news tweet solely from the reply. Furthermore, we show how the Colombian peace treaty has had a profound impact on the Colombian society, as it is the topic in which most people engage to show their opinions.

Keywords: latent topic analysis · text mining · information visualisation · machine learning.

1 Introduction

Social interaction on social networks has been a central research topic during the last decade. It has applications in modernisation of government's Twitter-based networking strategies [7,22,23,11,21], e-commerce [13,16,27] and business performance [5,9]. Of particular interest is the study of how people respond to a stimulus, forming either strong collective responses or weak, isolated responses. For example, [4,19], studied users' influence in microblogging networks, where implementing the InfluenceRank Algorithm for identifying people with high social media influence can be of great commercial value. An interesting area within

the stimulus–response field in social networks is the response of people to the news. In the case of Twitter, it has been identified since its early days as a news-spreading social network [12] where mass media are central actors. Previous work in this area is typically focused on topics such as sports, politics, health, education, travel, business, and so on [14,28,1,6].

In this work, we inquire on the recognition of the semantic structure of people's response to news tweets by one of the largest Colombian news mass media, Radio Cadena Nacional (RCN), with (7,7 million followers by August 2019). If there is a distinction between how people respond to different kinds of news, then state of the art natural language processing and machine learning techniques should be able to detect this difference. Therefore, we propose a two-stage pipeline to investigate the claimed differences. During the first stage, topic analysis is performed over the tweets by RCN, showing the main latent topics that they tweet about. In the second stage, we take the comments by regular users to those tweets and ask ourselves if it is possible to predict which is the topic of the tweet they are responding to. Our research is two-folded. On the one hand, we investigate the characterisation of how people respond to news from different topics. On the other hand, we pay particular attention to one of the most critical social phenomena of the past decade worldwide: the response by Colombian population to the peace treaty between the Colombian government and the Colombian Revolutionary Armed Forces guerrilla (FARC) [25]. These two entities were in a war for over 50 years, a conflict responsible for hundreds of thousands of deaths and millions of people forced from their homes, and being one of the largest human tragedies of modern times [3].

This paper is divided as follows: in Section 2 we describe in detail the dataset that we used to carry out our study as well as the workflow of the proposed method. In Section 3, we present the results and analysis of the research. Finally, the main conclusions and future work are presented in Section 4.

2   Method and Materials

In this section, we describe in detail the two stages of our research: a topic modelling and a topic prediction stage. The work pipeline is shown in Fig 1.

We collected tweets from the Twitter account @NoticiasRCN from 2014 to the present using the Twitter API. During this period, the mass media giant published 258,848 tweets, having 1,447,440 comments from their public. Then, we proceeded to pre-process our tweet database. We removed punctuation, links, hashtags and mentions from all the tweets. We further lemmatised the text in lowercase.

With the pre-processed data, we trained topic modelling algorithms in order to discover topics in the tweet corpus. Particularly we used the Latent Dirichlet Allocation (LDA) model [8] and a combination of unsupervised FastText model [2,10] and K-Means clustering [15].

LDA assigns K probabilities to a tweet of belonging to K different classes. The classes are learnt in an unsupervised fashion from the co-occurrence of words
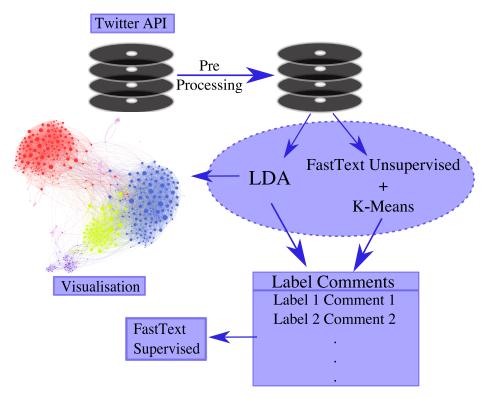
Fig. 1: Overview of the proposed method. We crawl tweets from @NoticiasRCN Twitter account. Then, pre-processing is made over this set of tweets. After that, two different unsupervised topic models are trained, from which visualisations of the tweets only from @NoticiasRCN can be built. We use the discovered latent topics to label comments to @NoticiasRCN tweets. Finally, we train a supervised FastText model to predict the topics of the comments.

within the tweets of the corpus. Each of the classes is called a latent topic. Therefore, the K probabilities assigned to a tweet by LDA can be interpreted as a vector of K components, where the k-th one shows the content percentage of the k-th topic in a tweet. The number of latent topics K is a parameter that one provides to the model. In principle, each latent topic corresponds to a topic a human would understand. However, if a low number of latent topics is provided, each latent topic may contain several real topics. On the other hand, if a high number of latent topics is provided, many latent topics may refer to the same real topic. To keep a good correspondence between latent topics and human-understandable topics or real topics, the number of latent topics should be carefully selected. Refs. [20,24] have shown that the best way to pick the number of latent topics is by measuring the C_V coherence, which has shown a large correlation with human judgements of the interpretability of the topics

extracted by LDA. For the @NoticiasRCN tweets, the optimum number of latent topics is 12, as it is shown in Fig 2.
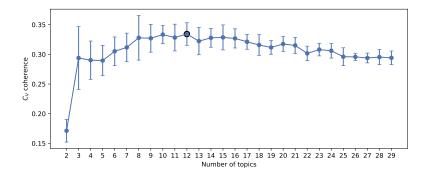

Fig. 2: C_V coherence as a function of number of topics for an LDA model trained with tweets from @NoticiasRCN. For each number of topics, a total of 20 LDA training runs were performed, and bars indicate a standard deviation of the corresponding 20 measured C_V coherences.

The other topic modelling algorithm consisted of training an unsupervised FastText model and performing K-means clustering on the unsupervised learnt vector representations for the tweets in the corpus. FastText is a memory-efficient and fast vector embedding algorithm based on the ideas of Word2Vec [18]. An important remark is that FastText uses sub-word information to enrich word vectors. This is particularly useful in a language such as Spanish, because it is largely inflected. What vector embedding algorithms do is to represent text as vectors of N components. Again, based on the co-occurrence of words or sequences of characters, FastText assigns N-dimensional vectors to each tweet. Here, the components do not have a human interpretation, as they embody abstract semantic features of the tweets. Once the embedded vectors for each tweet are learnt from FastText, the K-means clustering algorithm is used to group vectors in the embedded vector space. This combination of a vector embedding algorithm and a clustering algorithm has been used before with excellent results in unsupervised community detection models [26]. In order to compare both topic modelling algorithms (LDA and FastText + K-means), we used K-means to form 12 clusters, where each one represents a latent topic.

After that, we labelled the comments to the news tweets. Let $t_{\mathrm{N}}^{(i)}$ be a news tweet from @NoticiasRCN indexed by i, which runs from 1 to M, the total number of news tweets. Let $t_{\mathrm{C}}^{(j_i)}$ be a comment made by some Twitter user to the news tweet $t_{\mathrm{N}}^{(i)}$ indexed by j_i, which runs from 1_i to $N_i^{(i)}$, where $N^{(i)}$ is the total number of comments to the i-th news tweet $t_{\mathrm{N}}^{(i)}$. Then, we created a labelled comment dataset for each of the two considered topic modelling algorithms as

follows,

$$\left\{ \left(\ell_1, t_{\mathrm{C}}^{(1_1)} \right), \left(\ell_1, t_{\mathrm{C}}^{(2_1)} \right), \ldots, \left(\ell_1, t_{\mathrm{C}}^{(N_1^{(1)})} \right), \ldots, \left(\ell_M, t_{\mathrm{C}}^{(N_M^{(M)})} \right) \right\}, \qquad (1)$$

where ℓ_i is the latent topic (from either LDA or FastText + K-means) assigned to the i-th news tweet $t_{\mathrm{N}}^{(i)}$. Recall that ℓ_i is one of the 12 different latent topics.

From the vector representation of the tweets built with the unsupervised FastText model, we generated visualisations of the tweets in a 2D map. To do this, we applied a state of the art dimensionality reduction model called the Uniform Manifold Approximation and Projection (UMAP) [17]. UMAP learns topological relations between the FastText vectors of N components and finds representative projections of the data onto a two-dimensional vector space.

Finally, a supervised FastText model is trained to learn to predict the labels ℓ_i (provided by the two topic modelling algorithms) from the vector representations of the tweets $t_{\mathrm{C}}^{(j_i)}$, which are efficiently learnt by FastText.

3   Results and Discussion

After performing an LDA analysis on the news tweets, each news tweet was assigned a 12 component probability vector. We built the visualisation of such tweets by selecting the tweets that better represented each latent topic. To do this, we imposed a probability threshold on the LDA probability vectors. Tweets whose maximum LDA probability for some topic is above this threshold are called representative tweets for the corresponding latent topic. For a threshold of 0.8, Fig 3 shows an annotated visualisation of the most representative tweets. A couple of clusters are not annotated since their topic is not unique or simply not clear.

Concerning the FastText and K-means combination, a different selection mechanism for representative tweets was taken into account. K-means groups news tweets geometrically, building 12 vectors called the cluster centroids. We define the most representative tweets of each cluster as the set of tweets grouped in a cluster nearest to its corresponding centroid. Therefore, after defining a maximum distance threshold, we can visualise the clusters and their corresponding representative tweets. Note that this threshold does not have a probabilistic interpretation, and is in general different for different datasets, depending on the mean distance between the data points. In our case, the visualisation is shown in Fig. 4. Topics were easier to identify, and the visualisation shows clear clusters well-separated from the others. In terms of visualisation, the FastText + K-means technique is superior for our case study.

Now, we focus on the second stage of our research, where we predict the topic of a news tweet from a comment to that tweet. The precision and recall at $k = 1, \ldots, 10$ are shown in Fig. 5. From the results at $k = 1$ it is clear that for both LDA and FastText + K-means, the precision and recall are well-above the expected result of a random classifier. Again, FastText + K-means performs better. Of course, as k gets larger, recall increases and precision decreases.
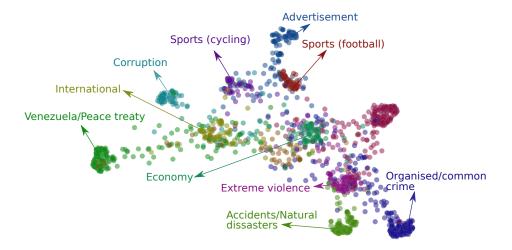
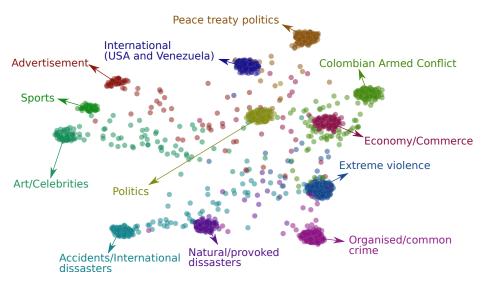Fig. 3: Visualisation of topics discovered with LDA.

Fig. 4: Visualisation of topics discovered with the combination of unsupervised FastText and K-means clustering.

Remarkably, having a 40% precision and recall at 1 means that there is a difference between how people respond to different topics. This result is quite good considering that topics were discovered with unsupervised learning, and that documents for classification are single tweets. This difference in response can be further examined with a histogram of the impact caused in the public by each latent topic. We do so only considering the FastText + K-means method,
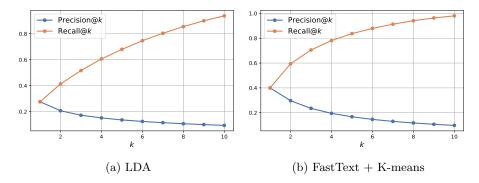
(a) LDA   (b) FastText + K-means

Fig. 5: Precision and Recall at $k = 1, \ldots, 10$ for the LDA and FastText + K-means topic modelling algorithms.

as it has consistently shown to yield better results. Fig. 6 shows that the topic identified as peace treaty politics is by far the more engaging one, having the most number of likes, retweets and replies per news tweet related to that topic. Therefore, our analysis confirms that the peace treaty has been the phenomenon with the largest impact on Colombian society.

4   Conclusions and Future Work

We presented two methods to automatically detect latent topics in news tweets and assess how significant were the differences between the response from the public to tweets from different topics. The first method consisted on discovering latent topics using the widely used LDA model. The second method consisted on generating embedded vectors with FastText and performing K-means clustering on those vectors. We visualised the most representative tweets for each latent topic using both methods and showed that the FastText + K-means method was superior both in the visualisation and in the interpretability of the topics.

Also, using the comments to the news tweets, we trained a supervised Fast-Text method to predict the topic of a news tweet from a comment of that tweet. Again, the better results were obtained with the FastText + K-means method, yielding 40% precision and recall at 1 in a 12-class classification problem.

The examination of the impact of each topic on the news Twitter account followers revealed that the topic identified as peace treaty politics was the most relevant topic, having the most number of likes, retweets and replies per news tweet, compared to the other topics.

We intend in the future to perform a sentiment analysis on the followers response in order to measure controversiality of topics as well as study which topics evoke positive/negative sentiment in the public.
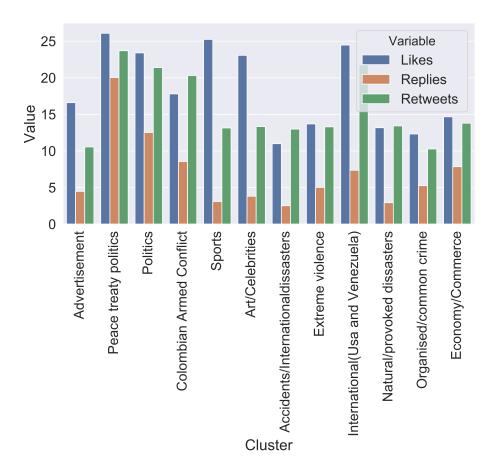
Fig. 6: Bar plot of number of likes, replies and retweets for the set of all news tweets belonging to each latent topic discovered by the FastText + K-means method.

References

1. Almutairi, N., Alhabash, S., Hellmueller, L., Willis, E.: The effects of twitter users' gender and weight on viral behavioral intentions toward obesity-related news. Journal of Health Communication **23**(3), 233–243 (2018). https://doi.org/10.1080/10810730.2018.1423648, pMID: 29388884
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
3. Cely, D.M.F.: Grupo de memoria histórica,¡ basta ya! colombia: Memorias de guerra y dignidad (bogotá: Imprenta nacional, 2013), 431 pp. 1. Historia y sociedad (26), 274–281 (2014)
4. Chen, W., Cheng, S., He, X., Jiang, F.: Influencerank: An efficient social influence measurement for millions of users in microblog. In: 2012 Second In-

ternational Conference on Cloud and Green Computing. pp. 563–570 (2012). https://doi.org/10.1109/CGC.2012.31

5. Culnan, M.J., McHugh, P.J., Zubillaga, J.I.: How large u.s. companies can use twitter and other social media to gain business value. MIS Quarterly Executive **9** (2010)

6. Garrett, R.K.: Social media's contribution to political misperceptions in u.s. presidential elections. PLOS ONE **14**(3), 1–16 (03 2019). https://doi.org/10.1371/journal.pone.0213500

7. Golbeck, J., Grimes, J.M., Rogers, A.: Twitter use by the u.s. congress. Journal of the American Society for Information Science and Technology **61**(8), 1612–1621 (2010). https://doi.org/10.1002/asi.21344

8. Hoffman, M.D., Blei, D.M., Bach, F.: Online learning for latent dirichlet allocation. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1. pp. 856–864. NIPS'10, Curran Associates Inc., USA (2010)

9. Ioanid, A., Scarlat, C.: Factors influencing social networks use for business: Twitter and youtube analysis. Procedia Engineering **181**, 977 – 983 (2017). https://doi.org/https://doi.org/10.1016/j.proeng.2017.02.496, 10th International Conference Interdisciplinarity in Engineering, INTER-ENG 2016, 6-7 October 2016, Tirgu Mures, Romania

10. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)

11. Kim, S.K., Park, M.J., Rho, J.J.: Effect of the government's use of social media on the reliability of the government: Focus on twitter. Public Management Review **17**(3), 328–355 (2015). https://doi.org/10.1080/14719037.2013.822530

12. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web. pp. 591–600. WWW '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1772690.1772751

13. ling Lai, L.S.: Social commerce – e-commerce in social media context (2010)

14. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 251–258. ICDMW '11, IEEE Computer Society, Washington, DC, USA (2011). https://doi.org/10.1109/ICDMW.2011.171

15. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)

16. Mata, F.J., Quesada, A.: Web 2.0, social networks and e-commerce as marketing tools. Journal of theoretical and applied electronic commerce research **9**, 56–69 (2014)

17. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints (Feb 2018)

18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

19. Nargundkar, A., Rao, Y.S.: Influencerank: A machine learning approach to measure influence of twitter users. In: 2016 International Conference on Recent Trends in Information Technology (ICRTIT). pp. 1–6 (2016). https://doi.org/10.1109/ICRTIT.2016.7569535

20. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. pp. 399–408. WSDM '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2684822.2685324

21. de Rosario, A.H., Sáez-Martín, A., del Carmen Caba-Pérez, M.: Using social media to enhance citizen engagement with local government: Twitter or facebook? New Media & Society **20**(1), 29–49 (2018). https://doi.org/10.1177/1461444816645652

22. Small, T.A.: e-government in the age of social media: An analysis of the canadian government's use of twitter. Policy & Internet **4**(3-4), 91–111 (2012). https://doi.org/10.1002/poi3.12

23. Sobaci, M.Z., Karkin, N.: The use of twitter by mayors in turkey: Tweets for better public services? Government Information Quarterly **30**(4), 417 – 425 (2013). https://doi.org/https://doi.org/10.1016/j.giq.2013.05.014

24. Syed, S., Spruit, M.: Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 165–174 (Oct 2017). https://doi.org/10.1109/DSAA.2017.61

25. Tellez, J.F.: Peace agreement design and public support for peace: Evidence from colombia. Journal of Peace Research **0**(0), 0022343319853603 (2019). https://doi.org/10.1177/0022343319853603

26. Vargas-Calderón, V., Camargo, J.E.: Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. Cities **92**, 187–196 (2019)

27. Zhao, W.X., Li, S., He, Y., Chang, E.Y., Wen, J., Li, X.: Connecting social media to e-commerce: Cold-start product recommendation using microblogging information. IEEE Transactions on Knowledge and Data Engineering **28**(5), 1147–1159 (2016). https://doi.org/10.1109/TKDE.2015.2508816

28. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) Advances in Information Retrieval. pp. 338–349. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)