Springer Link

Identifying Communities in Social Media with Deep Learning

International Conference on Social Computing and Social Media

SCSM 2018: Social Computing and Social Media. Technologies and Analytics pp 171-182 | Cite as

- Pedro Barros (1)
- Isadora Cardoso-Pereira (1)
- Keila Barbosa (1)
- Alejandro C. Frery (1)
- Héctor Allende-Cid (2)
- Ivan Martins (1)
- Heitor S. Ramos (1) Email author (heitor@laccan.ufal.br)

1. Instituto de Computação, Universidade Federal de Alagoas, , Maceió, Brazil
2. Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, , Valparaíso, Chile

Conference paper
First Online: 31 May 2018

- 4 Readers
- 444 Downloads

Part of the Lecture Notes in Computer Science book series (LNCS, volume 10914)

Abstract

This work aims at analyzing twitter data to identify communities of Brazilian Senators. To do so, we collected data from 76 Brazilian Senators and used autoencoder and bi-gram to the content of tweets to find similar subjects and hence cluster the senators into groups. Thereafter, we applied an unsupervised sentiment analysis to identify the communities of senators that share similar sentiments about a selected number of relevant topics. We find that is able to create meaningful clusters of tweets of similar contents. We found 13 topics all of them relevant to the current Brazilian political scenario. The unsupervised sentiment analysis shows that, as a result of the complex political system (with multiple parties), many senators were identified as independent (19) and only one (out of 11) community can be classified as a community of senators that support the current government. All other detected communities are not relevant.

Keywords

Community detection Deep Learning Text classification
Convolutional networks Autoencoder

This is a preview of subscription content, log in to check access.

Notes

Acknowledge

This work was partially funded by Fapeal, CNPq, and SEFAZ-AL. The work of Héctor Allende-Cid was supported by the project FONDECYT Initiation into Research 11150248.

References

Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1), 22–29 (1990)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Word%20association%20norms%2C%20mutual%20information%2C%20and%20lexicography&author=KW.%20Church&author=P.%20Hanks&journal=Comput.%20Linguist.&volume=16&issue=1&pages=22-29&publication_year=1990)

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
Google Scholar (https://scholar.google.com/scholar?q=Goodfellow%2C%20I.%2C%20Bengio%2C%20Y.%2C%20Courville%2C%20A.%3A%20Deep%20Learning.%20MIT%20Press%20%282016%29)

Hadgu, A.T., Garimella, K., Weber, I.: Political hashtag hijacking in the U.S. In: Proceedings of the 22nd International Conference on World Wide Web, WWW 2013 Companion, pp. 55–56. ACM, New York (2013). https://doi.org/10.1145/2487788.2487809 (https://doi.org/10.1145/2487788.2487809)

Jungherr, A.: Twitter use in election campaigns: a systematic literature review. J. Inf. Technol. Polit. **13**(1), 72–91 (2016)
CrossRef (https://doi.org/10.1080/19331681.2015.1132401)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Twitter%20use%20in%20election%20campaigns%3A%20a%20systematic%20literature%20review&author=A.%20Jungherr&journal=J.%20Inf.%20Technol.%20Polit.&volume=13&issue=1&pages=72-91&publication_year=2016)

Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
zbMATH (http://www.emis.de/MATH-item?1225.68219)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Visualizing%20data%20using%20t-SNE&author=LVD.%20Maaten&author=G.%20Hinton&journal=J.%20Mach.%20Learn.%20Res.&volume=9&pages=2579-2605&publication_year=2008)

Vaz de Melo, P.O.S.: How many political parties should brazil have? a data-driven method to assess and reduce fragmentation in multi-party political systems. PLOS ONE **10**(10), 1–24 (2015)
CrossRef (https://doi.org/10.1371/journal.pone.0140217)
Google Scholar (http://scholar.google.com/scholar_lookup?title=How%20many%20political%20parties%20should%20brazil%20have%3F%20a%20data-driven%20method%20to%20assess%20and%20reduce%20fragmentation%20in%20

multi-party%20political%20systems&author=POS.%20Vaz%20de%20Melo&journal=PLOS%20ONE&volume=10&issue=10&pages=1-24&publication_year=2015)

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media. Data Min. Knowl. Discov. **24**(3), 515–554 (2012)
CrossRef (https://doi.org/10.1007/s10618-011-0224-z)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Community%20detection%20in%20social%20media&author=S.%20Papadopoulos&author=Y.%20Kompatsiaris&author=A.%20Vakali&author=P.%20Spyridonos&journal=Data%20Min.%20Knowl.%20Discov.&volume=24&issue=3&pages=515-554&publication_year=2012)

Park, C.S.: Does twitter motivate involvement in politics? tweeting, opinion leadership, and political engagement. Comput. Hum. Behav. **29**(4), 1641–1648 (2013)
CrossRef (https://doi.org/10.1016/j.chb.2013.01.044)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Does%20twitter%20motivate%20involvement%20in%20politics%3F%20tweeting%2C%20opinion%20leadership%2C%20and%20political%20engagement&author=CS.%20Park&journal=Comput.%20Hum.%20Behav.&volume=29&issue=4&pages=1641-1648&publication_year=2013)

Rodrigues, J., Branco, A., Neale, S., Silva, J.: LX-DSemVectors: distributional semantics models for portuguese. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) PROPOR 2016. LNCS (LNAI), vol. 9727, pp. 259–270. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_27 (https://doi.org/10.1007/978-3-319-41552-9_27)
CrossRef (https://doi.org/10.1007/978-3-319-41552-9_27)
Google Scholar (http://scholar.google.com/scholar_lookup?title=LX-DSemVectors%3A%20distributional%20semantics%20models%20for%20portuguese&author=J.%20Rodrigues&author=A.%20Branco&author=S.%20Neale&author=J.%20Silva&pages=259-270&publication_year=2016)

Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487 (2016)
Google Scholar (https://scholar.google.com/scholar?q=Xie%2C%20J.%2C%20Girshick%2C%20R.%2C%20Farhadi%2C%20A.%3A%20Unsupervised%20deep%20embedding%20for%20clustering%20analysis.%20In%3A%20International%20Conference%20on%20Machine%20Learning%2C%20pp.%20478%E2%80%93487%20%282016%29)

Copyright information

© Springer International Publishing AG, part of Springer Nature 2018

About this paper

Cite this paper as:

Barros P. et al. (2018) Identifying Communities in Social Media with Deep Learning. In: Meiselwitz G. (eds) Social Computing and Social Media. Technologies and Analytics. SCSM 2018. Lecture Notes in Computer Science, vol 10914. Springer, Cham

- First Online 31 May 2018
- DOI https://doi.org/10.1007/978-3-319-91485-5_13

- Publisher Name Springer, Cham
- Print ISBN 978-3-319-91484-8
- Online ISBN 978-3-319-91485-5
- eBook Packages Computer Science

- Buy this book on publisher's site
- Reprints and Permissions

Personalised recommendations

© 2018 Springer Nature Switzerland AG. Part of Springer Nature.

Not logged in Not affiliated 131.161.25.63

Identifying Communities in Social Media with Deep Learning

Pedro Barros[1], Isadora Cardoso-Pereira[1], Keila Barbosa[1], Alejandro C. Frery[1], Héctor Allende-Cid[2], Ivan Martins[1], and Heitor S. Ramos[1]

[1] Instituto de Computação, Universidade Federal de Alagoas, Maceió – AL, Brazil
{pedro_h_nr,isadora.cardoso,keilabarbosa, acfrery, ivan.martins, heitor}@laccan.ufal.br
[2] Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso - Valparaíso, Chile
hector.allende@pucv.cl

Abstract. This work aims at analyzing twitter data to identify communities of Brazilian Senators. To do so, we collected data from 76 Brazilian Senators and used autoencoder and bi-gram to the content of tweets to find similar subjects and hence cluster the senators into groups. Thereafter, we applied an unsupervised sentiment analysis to identify the communities of senators that share similar sentiments about a selected number of relevant topics. We find that is able to create meaningful clusters of tweets of similar contents. We found 13 topics all of them relevant to the current Brazilian political scenario. The unsupervised sentiment analysis shows that, as a result of the complex political system (with multiple parties), many senators were identified as independent (19) and only one (out of 11) community can be classified as a community of senators that support the current government. All other detected communities are not relevant.

Keywords: Community detection, Deep Learning, Text classification, Convolutional networks, autoencoder

1 Introduction

The idea of community has changed with the growth of social media (such as forum, blogs, and micro-blogs), since it is possible to people, through the Internet, to connect and interact online based on shared interests and activities, even if they are not geographically close (Papadopoulos et al., 2012).

With 100 million daily active users, Twitter is one of the most popular social network nowadays. It enables the users to send short messages (called tweets) up to 280 characters and, has around 6000 tweets per second, which corresponds to over 500 million tweets per day[3]. Counting on this huge volume of data generated, Twitter can be seen as a valuable source of data that is useful for monitoring several social aspects, such as detecting and analyzing communities.

[3] https://www.omnicoreagency.com/twitter-statistics/

Jungherr (2016) reviews many studies that shows the power of Twitter to give meaningful insights about the American political scenario.

Brazil has a multi-party system, i.e., it admits the legal formation of several parties. This caused a highly fragmented party system, for instance, there are 81 senators currently in office, divided into 20 parties[4], which are: DEM (4 senators), PCdoB (1), PDT (3), PMDB (20), PODE (3), PP (3), PPS (1), PR (4), PRTB (1), PROS (1), PSB (4), PSC (1), PRB (1), PSD (4), PSDB (11), PT (9), PTB (2), PTC (1), REDE (1), and without parties (2). There are more parties in Brazil, but they do not have representative in Brazilian Senate.

Twitter data has been used to analyze political aspect in many different ways. For instance, Vaz de Melo (2015) showed that Brazil has and had many ideologically redundant parties, i.e., parties that are similar in the ideological space, being possible to reduce more than 20 to only 4 parties. Hadgu et al. (2013) analyzed tweets to show that some changes in political polarization of hashtags are caused by "hijackers" engaged in a particular type of hashtag war. Park (2013) have investigated the interrelationships between opinion leadership, Twitter use motivations, and political engagement.

Differently, our work aims to identify communities of Brazilian senators using Twitter posts. We analyze tweets of 76 senators (about 94 % of the total) from the beginning of their mandate rather than specific events, so we can have a better understanding of their political vision and if they share it. Thereunto, we collect the data using a Twitter API. After cleaning, we apply autoencoder and bi-gram to the content of tweets to find similar subjects and hence cluster the senators into groups. Thereafter, we applied an unsupervised sentiment analysis to identify the communities of senators that share similar sentiments about a selected number of relevant topics.

This article is organized as follows: Section 2 describes the methodology used to collect and analyze data; Section 3 presents the main results; and Section 4 concludes this work.

2 Methodology

Figure 1 depicts a schematic view of the methodology applied to this work. First and second steps, Data collection and preparation, are described in Section 2.1 and consist of collecting and preprocessing the twitter data to rule out obvious cluster and meaningless data that hinder the community detection process. Processing and Sentiment Analysis are described in Sections 2.2 and 2.3 and consist of the community detection technique used in this work. Finally, we described how we evaluated our approach in Section 2.4.

2.1 Data collection and preparation

We collected a real world dataset of tweets from Brazilian senators accounts. We obtained the list of senators official accounts from the Brazilian Senate offi-

[4] https://www25.senado.leg.br/web/senadores/em-exercicio/-/e/por-partido

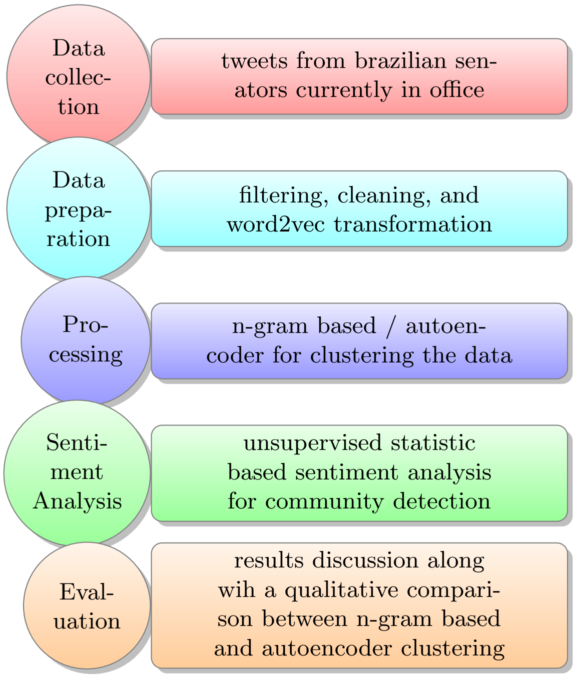

Data collection
tweets from brazilian senators currently in office
Data preparation
filtering, cleaning, and word2vec transformation
Processing
n-gram based / autoencoder for clustering the data
Sentiment Analysis
unsupervised statistic based sentiment analysis for community detection
Evaluation
results discussion along wih a qualitative comparison between n-gram based and autoencoder clustering

Fig. 1. Schematic view of the methodology used in this work

cial account. We found 76 accounts that represent a total of about 94 % of the actual number of senators currently in office. All extracted data are in Brazilian Portuguese and also performed all processing in the original language. We translated the tweets and subjects only when showing the results, for the sake of clarity and readability.

To gather the tweets, we use the API provided by Twitter[5]. The API returns each tweet in JSON format, with the content of the tweet, metadata (e.g., timestamp, replied or not, retweeted or not, etc.), and information about the user (username, followers, etc.). In this work, we only consider the username (Senator's account), the date, and the content of the tweet, in which we applied some cleaning techniques.

To clean the data, we removed stopwords (such as prepositions and pronouns), words with less than 3 letters, punctuation marks, and URLs. We also lower all the letters and removed graphic accentuation, in order to normalize writing. Moreover, we used a time filtering to collect data from January 1st 2014 to November 8th 2017, which corresponds to the date of the beginning of the current mandate of this legislature up to the day we decide to stop the data collection and analyze the data. It is worth mentioning that we are not able to collect all tweets for all accounts, because the tweeter API provides only a sample of the total amount of tweets. Furthermore, we manually removed some meaningless tweets, such as tweets with just greetings (e.g., "good morning", "hello friends", etc.) and automatic text (e.g., shared photo on different social networks).

We collected a total of $166,893$ tweets of 76 current Brazilian senators. Figure 2 shows the number of tweets per each senator twitter account during the period of data collection. As we can see most of the senators tweets more than 1000 times (59/76) and just few senators (9/76) tweets less than 5 times. This fact indicates that we collected a reasonable number of tweets. While red bars show the total number of collected tweets, blue bars show the final dataset after the cleaning and filtering processes. After all the filtering and the selection of relevant tweets (the ones related to a relevant topic), we have $33,550$ tweets in our final data set (the sum of blue bars).

We transformed all tweets into vectors of features using the Portuguese word embeddings set proposed by (Rodrigues et al., 2016), in which portuguese words are mapped into a vector of real numbers with 100 dimensions. We found in our sample of tweets that the maximum observed number of words is 29. Hence, we normalized the tweets in a way that they were mapped into a matrix of features of size 29×100. For tweets shorter than this dimension (with less than 29 words) we complete the gaps with zeros forming an array of $29 \cdot 10^2$ positions.

2.2 Processing

At this stage, the main goal is to find tweets of similar subject and cluster them into groups. In each group, we may find tweets that have positive or negative

[5] `http://www.tweepy.org/`

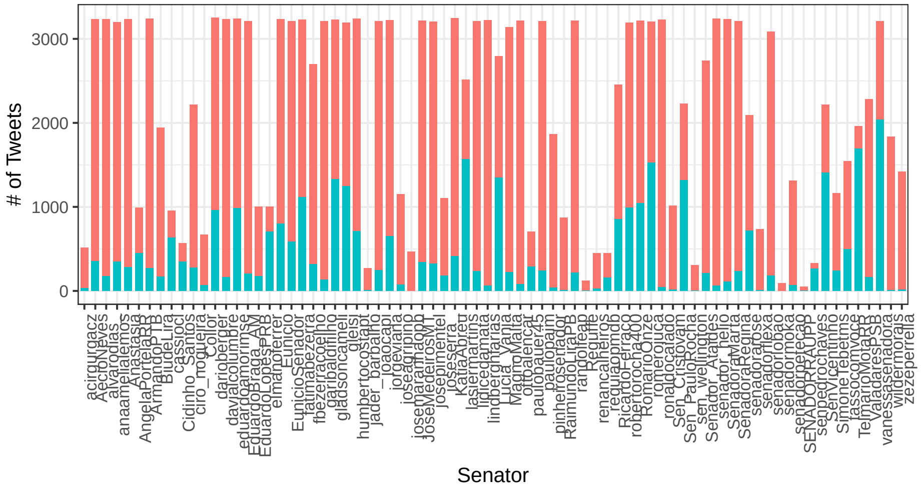


Fig. 2. Number of tweets per senator tweeter account during the period of data collection. Red bars show the total number of tweets, while blue bars show the number of tweets that were effectively used after cleaning and filtering the dataset.

sentiment about the subject, and we will further find communities of Senators that have positive or negative sentiments about these subjects.

We applied two different strategies to cluster the data. The first is a simple strategy and was used as a baseline. We count the top most frequent bi-grams of the dataset and applied two filters: (i) we ruled out all bi-grams with frequency smaller than 50, (ii) with the remain dataset, we manually removed all bi-grams that are irrelevant, for instance, greetings messages. We chose the number 50 because we observed that all bi-grams less frequent than this threshold were irrelevant and disconnected to real political scenario and sometimes they represent popular language terms that have no connection to relevant subjects.

After the filtering, we identified 13 subjects we judge relevant and highly correlated to the current political scenario in Brazil. For each bi-gram, we identified and grouped all senators that have tweets about this subject and we disregarded the senators that tweet less than 5 times about a subject.

N-gram is a well known and simple technique to apply to such scenario, but it is difficult to choose the correct granularity. For instance, we observed that we have to discard some tweets that were related to the so-called "Car Wash Operation" because this approach formed two different clusters: (i) "Wash Operation", and (ii) "Car Wash".

To try to overcome such situation, we adopted another approach to cluster tweets into groups of similar subjects. We use autoencoder (Goodfellow et al., 2016) as an unsupervised clustering technique. Goodfellow et al. (2016) defines an autoencoder as a neural network trained with the goal of copy its input to its output. The network comprises of two parts: (i) an encoder function, (ii) a decoder. Autoencoders have been applied to dimensionality reduction and information retrieval.

An autoencoder should be able to learn in a restricted way that makes it to copy only useful properties of the data. Recently, autoencoder has been used to perform unsupervised clustering as in (Xie et al., 2016). Authors used a variation of (Maaten and Hinton, 2008) and propose the Deep Embedded Clustering (DEC), which clusters a set of n points into a predetermined number of cluster (k). Instead of clustering directly in the data space, DEC transforms the data with a non-linear mapping $f_\theta \colon X \mapsto Z$. The new space Z has typically lower dimensionality than X, helping with the so-called curse-of-dimensionality. The function mapping f_θ is approximated with autoencoder, taking advantage of the theoretical function approximation property of neural networks.

DEC stands out from other clustering techniques due to the use of autoencoder to solve for feature space and cluster membership jointly. To do so, DEC uses a deep autoencoder network structure and finetunes it to minimize the reconstruction error. It then discards the decoder layers and use the encoder layers as its initial mapping between the data and the feature space. The second phase is the clustering itself that is done by computing an auxiliary target distribution and minimizing the Kullback-Leibler (KL) divergence to it. We chose DEC as our clustering method because we want to overcome the n-gram aforementioned problem.

2.3 Sentiment Analysis

In our approach to detect communities of senators from twitter data, the last step is the sentiment analysis. The main goal is to split, for each subject detected from the aforementioned clustering of tweets, senator into two groups: senators that have mostly positive posts and senators that have mostly negative posts. Hence, we can separate senators into two groups per subject. To do so, we use an unsupervised sentiment analysis proposed by Church and Hanks (1990), namely Pointwise Mutual Information (PMI), where words are statistically associated to words that are related to predetermined positive or negative sentiments.

PMI is an unsupervised technique that uses a list of predetermined words that are known in advance to be associated to positive or negative sentiment. It calculates the probability of an analyzed word be associated to a seed of positive or negative sentiment. This probability is calculated using the all tweets that the senator posted that contains the analyzed word. After that, it is calculated the statistic dependence of each word. PMI is thus defined as the mutual information between a given word c being positive (or negative) as follows

$$PMI(c, pos) = \log 2 \frac{\Pr(c, pos)}{\Pr(c) P(pos)}$$

and

$$PMI(c, neg) = \log 2 \frac{\Pr(c, neg)}{\Pr(c) P(neg)}$$

Informally, PMI is the probability of a given word c be positive with the probabilities of observing c and observing a positive *pos* word. Hence, for each word, it is calculated the difference between $PMI(c, pos)$ and $PMI(c, neg)$ as

$$\begin{aligned} PV_{PMI}(c) &= PMI(c, pos) - PMI(c, neg) \\ &= \log 2 \frac{\Pr(c, pos)/\Pr(pos)}{\Pr(c, neg)/\Pr(neg)} \\ &= \log 2 \frac{\Pr(c \mid pos)}{\Pr(c \mid neg)} \end{aligned}$$

Finally, we sum up the $PV_PMI(c)$ for all words of a tweet to calculate the sentiment about of that specific post.

It is worth mentioning that positive or negative sentiments are correlated to the opinion about a subject but it not necessarily means that a positive tweet agrees with a subject or, conversely, a negative tweet disagree with a subject. This happens because, sometimes, a text may use irony or some figure of speech that is difficult to be correctly recoginized by the techinque described herein.

2.4 Evaluation

We performed two analysis with the results of our proposal. In the first, we created a time series from the sentiment analysis of each senator per each subject we have studied. Hence, we registered the PMI value for the time span of the analysis. The main goal of this analysis is to observe some non-usual behavior such as a steep change of a senator sentiment in a specific subject, suggesting that this senator changed the opinion due to some relevant event. The second analysis is related to the community detection itself. For this analysis, we created a vector of PMIs for each subject detected by the clustering analysis, in our case, a 13-dimension vector, with the goal of analyzing what senators have similar sentiments about the selected subjects. To do so, we associated for each dimension (subject) a sentiment value of -1 (negative), 0 (neutral) or 1 (positive).

We used Principal Component Analysis (PCA) to reduce the dimensionality to 4, where we used a DBSCAN clustering method with parameters $\epsilon = 0.4$ and a minimum of 3 senators per cluster. We chose the DBSCAN clustering due to the fact that a senator can be independent, i.e., not associated to any other cluster. This is a common situation where a senator does not necessarily follow the party orientation and have an independent opinion about different subjects.

3 Results and Discussion

Figure 3 shows the results of PMI changing in time for some selected senators, representing the changes of the sentiment of a given senator for a specific subject.

The Car-wash Operation started at March 2014 and consists of a set of investigations against corruption at the state-controlled oil company Petrobras, in progress by the Federal Police of Brazil. Many parties of the Brazilian political system have politicians investigated by the Car-wash Operation.

Randolfe Rodriguez was cited by an accused businessman to be a participant in corruption in early 2016. We can see in Figure 3(a) that he expressed a very negative feeling about the operation that time, with PMI less than -20. After that, he showed less negative feelings about it, reaching almost a neutral sentiment by the middle of 2017. Figure 3(b) shows the PMI value for the tweets of Roberto Requião, senator for the same party (PMDB) as the current President of Brazil, Michel Temer. Senator Requião's tweets present a mild sentiment about the Car-wash Operation, varying from slightly positive to slightly negative and reaching a neutral sentiment by the end of 2017.

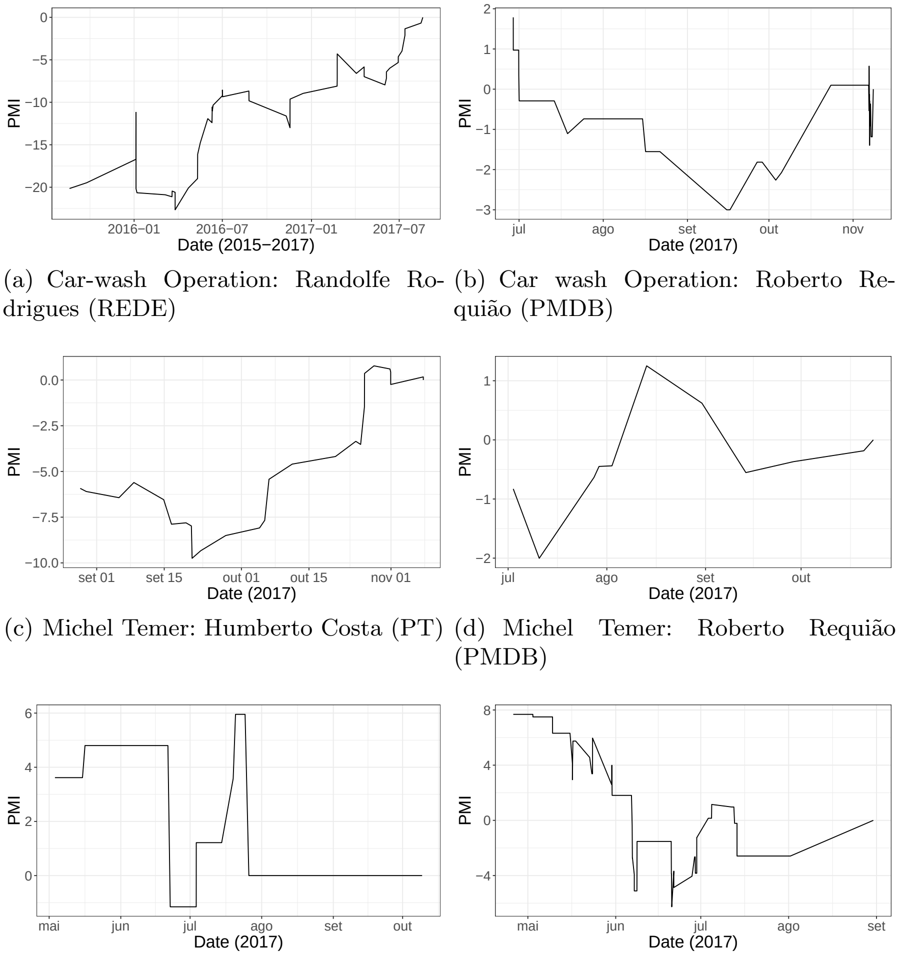
Michel Temer is the current Brazilian president, after the impeachment of Dilma Rousseff (August, 2016), which was received with mixed feelings by the Brazilian population. Humberto Costa is from the same party of the ex-president. We can see in Figure 3(c) that he shows strong negative feelings about Michel Temer, but this behavior evolved to an almost neutral sentiment by the middle of 2017. On the other hand, Roberto Requião is from the same party of the current President, hence showing more positive feelings about him.

PSC and PMDB are both in favor of Labor-rights Reform. Eduardo Amoring and Romero Jucá are two party members of PSC and PMDB, respectively. We can see in Figures3(e) that Eduardo Amorim have neutral or positive feelings about the subject. Romero Jucá, the responsible for voting the reform, shows some negative feelings in the same time that the Labor-rights Reform was accused of preventing the fight against slave labor, which forced him to change some reform points. This such behavior is shown in Figure 3(f).

The aforementioned results depicted how a sentiment analysis can be used in the understanding of the political momentum. As we can observe in the example of Romero Jucá about the Labor-rights reform, even tough he was in charge for voting the reform, and hence, he supports the reform, he posted some negative tweets showing his concern about the difficult he faced to conduct the process. Therefore, we cannot state that a politician that presents positive (or negative) sentiment about a topic necessarily supports or disapproves the matter.

In the following, we will present the results of the community detection method we have described and we will also cluster the politicians by the sentiment regarding a topic. Although some situations as the one described in Figure 3(f) is possible to happen, we can say that the sentiment about a topic correlates with (but does not necessarily leads to) the opinion that a politician have about a specific matter.

Figure 4 shows the communities we were able to detect by using the method described in Section 2. To detect these communities, we created a vector of values -1, 0 or 1 corresponding to the sentiment about a given topic (each topic corresponds to a position in this vector) for each senator. The topics we have considered are: Pension Reform, Aécio Neves (a senator also cited in the Car-wash Operation), Indigenous people, Michel Temer, President Dilma, Car-wash Operation, President Lula, Bolsa Família[6], Criminal responsibility age, Family



(a) Car-wash Operation: Randolfe Rodrigues (REDE)

(b) Car wash Operation: Roberto Requião (PMDB)

(c) Michel Temer: Humberto Costa (PT)

(d) Michel Temer: Roberto Requião (PMDB)

(e) Labor-rights Reform: Eduardo Amorim (PSC)

(f) Labor-rights Reform: Romero Jucá (PMDB)

Fig. 3. Variation of PMI for some selected senators

agriculture, Eduardo Cunha, Labor-rigths Reform, and Political Reform. All topics represent important aspects of the current Brazilian political momentum. In this plot, instead of showing individual politician names, we depicted how many politicians of each party showed up in each community. The font size is proportional to the number of senators of a given party that appears in that community. PMDB/NP and PDT/NP represents two senators that were elected by PMDB and PDT, but are not affiliated to these parties anymore.

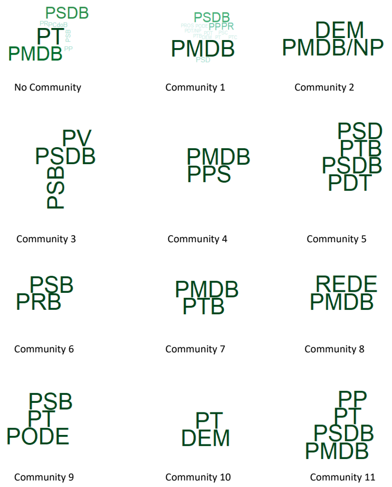
Table 1. Number of senators per community per political party

Party	NC	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	Total
DEM	0	1	1	0	0	0	0	0	0	0	1	0	3
PCdoB	1	0	0	0	0	0	0	0	0	0	0	0	1
PDT	0	1	0	0	0	1	0	0	0	0	0	0	1
PDT/NP	0	1	0	0	0	0	0	0	0	0	0	0	1
PMDB	5	9	0	0	1	0	0	1	1	0	0	1	18
PMDB/NP	0	0	1	0	0	0	0	0	0	0	0	0	1
PODE	0	1	0	0	0	0	0	0	0	1	0	0	2
PP	1	3	0	0	0	0	0	0	0	0	0	1	5
PPS	0	0	0	0	1	0	0	0	0	0	0	0	1
PR	1	3	0	0	0	0	0	0	0	0	0	0	4
PRB	0	0	0	0	0	0	1	0	0	0	0	0	1
PROS	0	1	0	0	0	0	0	0	0	0	0	0	1
PSB	1	0	0	1	0	0	1	0	0	1	0	0	4
PSC	0	1	0	0	0	0	0	0	0	0	0	0	1
PSD	0	2	0	0	0	1	0	0	0	0	0	0	3
PSDB	4	5	0	1	0	1	0	0	0	0	0	1	12
PT	6	1	0	0	0	0	0	0	0	1	1	1	10
PTB	0	1	0	0	0	1	0	1	0	0	0	0	2
PTC	0	1	0	0	0	0	0	0	0	0	0	0	1
PV	0	0	0	1	0	0	0	0	0	0	0	0	1
REDE	0	0	0	0	0	0	0	0	1	0	0	0	1
Total	19	31	2	3	2	4	2	2	2	3	2	4	76

The first group of parties (the top left group) depicts some senators that did not show any similarity with others. We can observe that PSDB, PT and PMDB present four, six and five senators, respectively, which have independent sentiment about the matters we have studied. These numbers correspond to about 33 %, 60 %, and 28 %, respectively, of the total representation of these parties in the Brazilian Senate. The community C1 has the PMDB as principal party and presents other parties such as PSDB, and PP, all of them support the current president Michel Temer. Therefore, this community is likely to be the government base in the Brazilian Senate. As we can see in Table 1, there are

[6] Assistance program for poor families

19 senators that cannot be associated to any community and 31 senators in the community that supports the government. These two groups amount about 66 % of all senators considered in this study. All other communities are comprised by a small number of senators, typically $2 - 4$ and represent groups that are most likely formed by chance. Observe in Table 1 that there is no community from C2 to C11 aggregating more than one senator of the same party. Showing that these communities are not too relevant from the point-of-view of the political analysis. Senators assigned to small communities like the ones from C2 to C11 are likely to have independent opinion, such as the ones assigned to NC. Hence, in this analysis, we conclude that there are only two relevant communities, C1, likely to be the community of senators that support the current government, and NC, formed by senators with independent opinion regarding the subjects studied herein.


No Community

Community 1

Community 2

Community 3

Community 4

Community 5

Community 6

Community 7

Community 8

Community 9

Community 10

Community 11

Fig. 4. Clusters of parties

4 Conclusion

This paper presents a study of a sample of 166, 893 tweets from Brazilian senators in office, in which a deep autoencoder was applied to cluster these tweets into groups of similar topics. We used autoencoder in contrast of a create clusters using the n-gram directly, because it is hard to find the granularity, n, which is more convenient to split the dataset into relevant topics. Hence, autoencoder is able to cluster the tweets without specify a specific granularity of important terms. We have identified 13 relevant topics about the current Brazilian political scenario. The final dataset presents $33,550$ tweets of 76 different senators. We further performed a sentiment analysis to split the senators that tweeted about the same topic into three groups having: (i) positive, (ii) negative, and (iii) neutral sentiment about each selected topic. Furthermore, we created a vector of 13 positions for each senator and applied DBSCAN to detect communities of senators that have similar overall sentiment about the selected topics. We observed that, as a result of the complex political system (with multiple parties), many senators were identified as independent (19) and only one (out of 11) community can be classified as a community of senators that support the current government. All other detected communities are not relevant. The techniques used in this analysis have some limitations. For instance, although, we observed that the sentiment is correlated to the politician opinion about a topic, we cannot conclude the sentiment is sufficient to state the politician opinion about a topic. Moreover, despite some limitations, this correlation is useful to create the communities and we can use this technique to extract a big picture of the political scenario, although we found we cannot make individual conclusions. It is worth mentioning that this was an observational study, hence, we cannot state any causal relation. Moreover, the conclusions about this sample cannot be generalized, although they are useful to improve the understanding of the current Brazilian complex political scenario.

Acknowledge

This work was partially funded by Fapeal, CNPq, and SEFAZ-AL. The work of Héctor Allende-Cid was supported by the project FONDECYT Initiation into Research 11150248.

Bibliography

Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. 16(1), 22–29 (Mar 1990)

Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)

Hadgu, A.T., Garimella, K., Weber, I.: Political hashtag hijacking in the u.s. In: Proceedings of the 22Nd International Conference on World Wide Web. pp. 55–56. WWW '13 Companion, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2487788.2487809

Jungherr, A.: Twitter use in election campaigns: A systematic literature review. Journal of Information Technology & Politics 13(1), 72–91 (2016)

Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(Nov), 2579–2605 (2008)

Vaz de Melo, P.O.S.: How many political parties should brazil have? a data-driven method to assess and reduce fragmentation in multi-party political systems. PLOS ONE 10(10), 1–24 (10 2015)

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media. Data Mining and Knowledge Discovery 24(3), 515–554 (May 2012)

Park, C.S.: Does twitter motivate involvement in politics? tweeting, opinion leadership, and political engagement. Computers in Human Behavior 29(4), 1641–1648 (2013)

Rodrigues, J., Branco, A., Neale, S., Silva, J.: Lx-dsemvectors: Distributional semantics models for portuguese. In: International Conference on Computational Processing of the Portuguese Language. pp. 259–270. Springer (2016)

Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning. pp. 478–487 (2016)

Med-e-Tel 2013

Electronic Proceedings

of

The International eHealth, Telemedicine and Health ICT Forum for Educational, Networking and Business

Editors
Malina Jordanova, Frank Lievens

April 10-12, 2013
Luxembourg, G. D. of Luxembourg

Publisher

International Society for Telemedicine & eHealth (ISfTeH)
Coordinating Office
c/o Frank Lievens
Waardbeekdreef 1
1850 Grimbergen
Belgium
Phone: +32 2 269 8456
Fax: +32 2 269 7953
E-mail: info@isfteh.org
www.isft.org

Med-e-Tel 2013
Electronic Proceedings: The International eHealth, Telemedicine and Health ICT Forum for Educational, Networking and Business
Editors: Malina Jordanova, Frank Lievens
ISSN 1818 - 9334

All rights reserved. No part of this proceedings may be reproduced, stored in a retrieval system or otherwise used without prior written permission from the publisher, ISfTeH.

© ISfTeH, 2013, Printed in G. D. of Luxembourg

With the publication of this Electronic Proceedings as well as with the sixth book from the series “Global Telemedicine and eHealth Updates: Knowledge Resources”, Med-e-Tel strengthens its position as a widely recognized International Educational, Networking and Business Forum for eHealth, Telemedicine and Health ICT.

Enjoy your reading!

Editors
Malina Jordanova, MD, PhD
Educational Program Coordinator, Med-e-Tel
Space Research & Technology Institute, Bulgarian Academy of Sciences
Bulgaria

Frank Lievens
Director, Med-e-Tel
Board Member, Secretary and Treasurer of International Society for
Telemedicine & eHealth (ISfTeH)
Belgium & Switzerland

http://www.academia.edu/3439958/Global_Telemedicine_and_EHealth_Updates_6th_ed._

Speed-up the Image Retrieval of Lung Nodules in the BigData Age

M. C. Oliveira[1], D. C. da Silva[1], K. B. C. dos Santos[1]
[1]Instituto de Computação(IC) e Laboratório de Telemedicina e Informática Médica do Hospital Universitário(HUPAA) da Universidade Federal de Alagoas (UFAL), Av. Lourival Melo Mota, s/n, Tabuleiro do Martins, Maceió, 57072-900, Brazil, oliveiramc@ic.ufal.br

Abstract: **The Content-Based Image Retrieval (CBIR) has received great attention in the medical community because it is capable of retrieving similar images that have known pathologies. However, the sheer volume of data produced in radiology centers has precluded the use of CBIR in the daily routine of hospitals. The volume of medical images produced in medical centers has increased fast. The annual data produced from exams in the big radiology centers is greater than 10 Terabytes. Therefore, we have reached to an unprecedented age of "BigData". We here present a bag-of task approach to speed up the images retrieval of lung nodules stored in a large medical images database. This solution combines texture attributes and registration algorithms that together were capable of retrieving images of benign lung nodules with greater-than-72% precision and greater-than-67% in malignant cases, yet running in a few minutes over the Grid, making it usable in the clinical routine.**

Introduction

The volume of data produced in medical centers has increasing fast. The annual production of the big radiology centers is about 10 Terabytes. This situation exists due to the ease that the data of the patients are obtained and stored, resulting mainly from the reduction of the cost of the equipment during the last years.

The Content-Based Image Retrieval (CBIR) has received great attention in the medical community because it is capable of retrieving similar images that have known pathologies. However, the sheer volume of images produced in radiology centers has precluded the use of CBIR in the daily routine of hospitals. Therefore, we have reached to an unprecedented age of "BigData" and it has been motivating research and companies to find new solutions.

The Grid Computing (GC) technology represent one of the most recent and promising tool in distributed computing. GC is the integration of many computers distributed geographically, making it possible to create a virtual computing platform, giving to users and institutions a virtually unlimited capacity to solve problems related to the storage and access of data, and also to process applications with high computational costs. Techniques focused on the medical image retrieval are a major beneficiary of the GC technology due to their characteristics and necessities: high processing and large storage. Besides, GC is a low cost solution for public hospitals and small clinics, because, it is able to use the idle recourses of computers [1]. This paper presents a Bag-of-Task GC approach to speed up the images retrieval of lung nodules stored in a big medical images database.

Application Description

The overall application is described in Figure 1 and a detailed description is given in the sections below. All the images inserted in the PACS have removed the patients` information in respect to the HIPAA [2]. We have used 20,000 images from the Lung Image Database Consortium (LIDC), which is a BigData of lung cancer [3]. In the LIDC each nodule is manually segmented and then classified in benign or malignant by physicians. To use as reference we selected 100 benign nodules and 100 malignant nodules

The application has two CBIR modules. The first module uses the second-order Texture Analysis (TA) (co-occurrence matrix) to filter the 1000 most similar images into the second module. The second module uses the Image Registration (IR) algorithms to find the similarity between an image defined by the user as a reference and the images filtered by the first module. The second module starts when the specialists select the registration module. The IR algorithm is executed in parallel on the grid machines using the reference image and the images classified by the first module. The application sorts in a list the most similar nodules according to the Cross Correlation values. Based on the DICOM protocol information, the application can also retrieve the information stored on the LIDC to aid the physician.

Because of the high computational cost related to the IR algorithms, the second module is processed on the OurGrid computational grid (www.ourgrid.org). This grid is free-to-join and cooperative, where sites share their idle computing resources and, when necessary, receive idle resources from other sites. OurGrid assumes that the parallel applications that run on it are a Bag-of-Tasks (BoT), i.e., those tasks are independent from each other. Currently, the OurGrid is composed of nearly 500 computers [4].

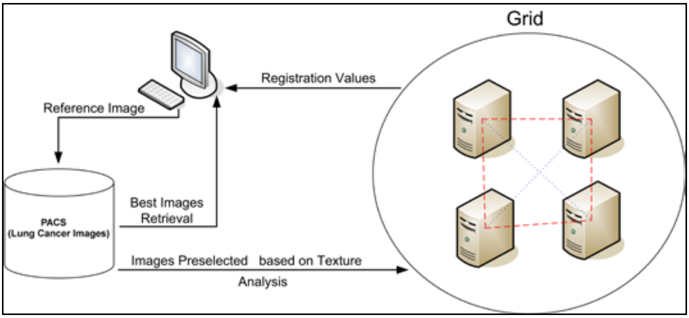


Fig. 1 CBIR lung cancer image retrieval schema using Grids

Results

The results of the CBIR tool developed in this work were assessed by the leave-one-out technic. The tests were repeated five times for each module and we have used different kinds of nodules for each test (ten benign nodules and ten malignant nodules). The nodules were selected in a random way. The algorithm precision was obtained by dividing the type number of nodules retrieved over the number of nodules retrieved.

The first module's time processing average was 1.9 minutes. This time was related to the calculation of the Manhattan Distance between the characteristic vector of each DICOM server's image and the characteristic vector of the reference image. The average precision of retrieving obtained by the first module was 0.73 (benign nodules) and 0.76 (malignant nodules), considering the ten most relevant nodules retrieved. To analyze the image retrieval capacity of the second module, a group of 1000 most similar nodules filtered by the first module has used in the second module. The average precision of retrieving obtained by the second module was 0.72 (benign nodules) and 0.67 (malignant nodules). In all the experiments, the results produced by the IR algorithms were very close to the traditional TA technique.

The Bag-of-Task approach was able to greatly reduce the high processing time of the IR algorithm. The GC was able to amortize the total time of the algorithm in 116.97 minutes, compared to the processing time obtained in one machine. The experiments used 50 Grid processors and 100 MB/s LAN. The application total time, considering the time to calculate the Manhattan Distance and the time to execute the IR, was 5.02 min.

Discussion and Conclusion

A computational algorithm was developed in this work, capable of using the high processing power of the Grid Computing technology to make feasible the CBIR using the Image Registration algorithms against a BigData of lung cancer images. Oliveira MC and colleagues [1] showed a preliminary work using IR as CBIR technique. However, the authors did not focus his attention in a specific disease. To our knowledge the results using IR and TA to retrieve lung nodules have never been published. Furthermore, our results evidenced that IR is precise and effective in retrieve similar lung nodules, besides that, the IR has showed very close results to the traditional TA technique.

The retrieval processing time was more than 100 minutes in only one computer using IR techniques. This time is impracticable for a CBIR to be applied in the clinical routine. In the grid, however, this time was reduced to less than 3 minutes in mean, making it affordable for clinical use. Therefore, a Bag-of-Task approach was fundamental to amortize the total processing time of the Image Registration algorithms against a BigData of lung cancer images. Besides this, we have shown a new methodology to evolve CBIR's state of art techniques through the Image Registration techniques.

Acknowledgment

This study was supported by Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL) and Hospital Universitário Prof. Alberto Antunes (HUPAA/UFAL).

References

[1] M. C. Oliveira Cirne, Walfredo, Marques, P. M. Azevedo, "Towards applying content-based image retrieval in the clinical routine," *Future Generation Computer Systems*, vol. 23, pp. 466–474, 2007.

[2] B. J. Liu Z. Zhou, H. K. Huang, "A HIPAA-Compliant Architecture for Securing Clinical Images," *Journal of Digital Imaging*, vol. 19, no. 2, pp. 172–180, 2006.

[3] H. Lin, Z. Chen, and W. Wang, "A pulmonary nodule view system for the Lung Image Database Consortium (LIDC).," *Academic radiology*, vol. 18, no. 9, pp. 1181–5, Sep. 2011.

[4] W. Cirne, Brasileiro F, Andrade N, Costa L, Andrade A, Novaes R, and M. M., "Labs of the World, Unite!!!," *Journal of Grid Computing*, vol. 4, no. 3, pp. 225–246, 2006.

Marcelo Costa Oliveira is member of the faculty at Computer Institute of the Federal University of Alagoas, in Brazil. Dr. Oliveira holds a PhD. in Medical Informatics from University of São Paulo (USP), M.Sc degree in Physics Applied to Medicine and B.Sc in Computer Science. Dr. Oliveira heads the Research and the Telehealth Project of the Hospital of the University of Alagoas. His current research interests include BigData Health, Cloud Computing and Medical Image Processing.

Darlan Chrystian da Silva is an undergraduate student at Federal University of Alagoas, in Brazil. He is member in the Research and the Telehealth Project of the Hospital of the University of Alagoas. His research interest BigData Health.

Keila Costa Barbosa's degree in Systems Analysis from the University of Northern Paraná, in Brazil. Degree in Systems Analysis and Development of Health State University of Alagoas. He is currently a MS in Computational Modeling Knowledge Federal University of Alagoas and Post graduate in Technologies for Web Applications at the University of Northern Paraná. His research interests include Health BigData, Cloud Computing and electronic medical records.