

Topic Modeling based on Louvain method in Online Social Networks

Guilherme Sakaji Kido
Universidade Estadual de
Londrina
Rodovia Celso Garcia Cid Pr
445 Km 380
Paraná, Londrina, Brazil
guilhermekido@gmail.com

Rodrigo Augusto Igawa
Universidade Estadual de
Londrina
Rodovia Celso Garcia Cid Pr
445 Km 380
Paraná, Londrina, Brazil
igawa.rodrigo@gmail.com

Sylvio Barbon Jr.
Universidade Estadual de
Londrina
Rodovia Celso Garcia Cid Pr
445 Km 380
Paraná, Londrina, Brazil
barbon@uel.br

ABSTRACT

Online Social Networks (OSNs) are the most used media nowadays, such as Twitter. The OSNs provide valuable information to marketing and competitiveness based on users posts and opinions stored inside huge volume of data from several themes, topics and subjects. In order to mining the topics discussed on an OSN we present a novel application of Louvain method for Topic Modeling based on communities detection in graphs by modularity. The proposed approach succeeded in finding topics in five different datasets composed of textual content from Twitter and Youtube. Another important contribution achieved was about the presence of texts posted by spammers. In this case, a particular behavior observed by graph architecture (density and degree) allows the classification of a topic as natural or artificial, this last created by the spammers on OSNs.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Text analysis; H.2.4 [Systems]: Textual databases; H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Measurement, Community, Topic Words

Keywords

Topic Model, Twitter, Youtube, ADVF, Louvain

1. INTRODUCTION

Nowadays, textual content is the most present in comparison to image, audio and video. Recent studies indicate that 80% of companies' information are text documents [1]. The production of digital data is increasing in volume because the accessibility of this media has become something

easy, fast and useful. People write articles on websites, forums, social networks, blogs and e-mails. These sources of information are a rich base of knowledge for organizations such as banks, universities, government, and marketing. The interests, concerns and criticism from users are stored in databases and can improve the products and services of organizations [3] [5] [6]; in policy, this data can adjust political placements in respect of sentiment analysis of your target audience [18].

Online Social Networks (OSNs) are services that have emerged as a new communication between individuals and organizations [14]. These services provide an essential platform for users to share thoughts, ideas, status, and experiences [21]. Due to the large number of texts, the OSNs have been extremely valuable to marketing companies and public organizations to find opinions about particular topics [12]. The standard methods used for Text Mining are usually applied to traditional texts in the Web, such as articles, news, and reports [20].

Micro-blog texts have more casual language and informal than traditional texts but are a source of similar relevant information in comparison to other textual sources. Due to the limit of the number of characters, users publish by a simplified way, using the colloquial language, abbreviations, slangs and generally links, emoticons, photos, videos, and others [8]. Colloquial and informal language usually creates specific words and terms that are considered noise. Noise is an undesirable textual feature and to eliminate or reduce it is necessary specific measures. One example is the *Adaptive Distribution of Vocabulary Frequencies* (ADVF) [10], capable of highlighting terms detected as textual noise.

The recent preoccupation about textual noise to obtain knowledge from OSNs is getting more attention due to spamming activities. One of the problems with knowledge discovery from OSNs when compared to traditional texts is the presence of spamming activities. These activities are practiced by fake accounts or compromised accounts. The first one, also called bot account, is an account for spreading malicious contents only [11]. A compromised account is a legitimate account which has been taken over by an attacker to publish fake or harmful content [12]. Directly or indirectly, both aims to spread content that need be avoided in knowledge discovery because do not contain real information. In an OSNs, a bot repeatedly publishes texts with the a subject, that can compromise the topic trends. The presence of artificial content made by a spammer could lead to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SBSI 2016, May 17th-20th, 2016, Florianópolis, Santa Catarina, Brazil
Copyright SBC 2016.

bias the result precision and application's goals.

Independent of textual-based application, an important task is to identify the textual content, which each text can be determined by one or more topics/keywords that describe the main subject. There are several methods to find those keywords. Topic Modeling is the main area of this activity. For Zeng et al. [22], topic is considered an aggregate of words and their frequency, which can be extracted from a document and is an important unit of the Topic Modeling Process.

This present work aims a new methodology for Topic Modeling based on a feasible solution to OSNs. Our solution is capable of handling problems of noise and spam, discovering the topics in the dataset. The kernel of proposed methodology was based on Louvain method and the concept of modularity that provides a quality measure of the communities in a graph [19]. In other words, we interpreted a community in a graph such as a group of terms around a topic and based on the modularity it is possible to detect the existence of different topics in the same dataset. To detect the noisy terms we applied ADVF [10], and to treat the causes of artificial topics, like spam, we suggest an architecture analysis of the graph, mainly the density value.

The dataset used in the experiments was Twitter¹, considered one of the largest existing micro-blogging services today; and the Youtube², one of the biggest video-sharing website. Their contents were extracted, filtered, processed and visualized in graphs in order to form word's networks. These graphs were analysed based on its structure and the topics found were classified in natural or artificial. The first one means topics that have some semantic context with the base's theme, and the other means topics created by spammer activity.

This paper is organized as follows. Next section provides our proposed approach for this paper and deals with the explanation of the ADVF and *Louvain* method. In Section 3 shows how the experiments were performed in this work. Section 4 presents our results and discussions about our method in OSNs datasets. Finally, Section 5 provides the implications and limitations found.

2. RELATED WORKS

Topic Modeling, as Latent Semantic Analysis (LSA) [13] and Probabilistic Latent Semantic Analysis (PLSA) [7], are popular in traditional text documents that need a vast amount of data, i.e., thousands of documents with thousands of words to generate coherent topics [4]. Many micro-blogs have limited number of characters (e.g., Twitter 140 character limit) per post, but present a high frequency of submissions that implies in a huge amount of data. This scenario would be the expected for Topic Modeling on OSNs. However, the presence of noise and malicious activities increases the difficulty of extraction topics on blogs [14] and OSNs.

About traditional techniques of Topic Modeling, Huang et al. [9] evaluated methodologies based on Vector Space Model and LSA. First, the tool developed by the authors performed the pre-processes data, eliminating stopwords and extracting scores. The TF-IDF, a weighting algorithm, was applied to each term resulted from pre-processing obtaining a value which measures the importance of a term concerning

¹www.twitter.com

²www.youtube.com

all dataset. This importance was based on frequency distribution of the terms and clustered by K-means. The dataset used in Huang et al. [9] was composed of Sina Weibo's texts, a Chinese micro-blog. Compared to the LSA, the work's conclusion determined that the methodology presented a better performance on indexing topics. However, the proposed approach does not handle noisy terms and malicious content.

In Tsai's work [20], the author analyze the Author-Topic method (AT) (a LDA extension) and compare which topics were similar to others, using the *Isometric feature mapping* (Isomap). The AT was applied on the Nielson BuzzMetrics's dataset, a blog about security threats and incident reports of cyber crime and computer virus. The author has succeeded in Topic Modeling, but the methodology of noise detection was based on manual labeling by users. This strategy can be non-trivial on a huge dataset, been effective just in a small dataset with few topics.

3. PROPOSED APPROACH

Our proposed methodology, as in the Figure 1, can be summarized as: 1) pre-processing; 2) ADVF analysis; 3) co-occurrence extraction; 4) Louvain calculus and, finally, topics identification. The datasets are formed by texts only, so no additional information besides the content was necessary to perform the proposed approach.

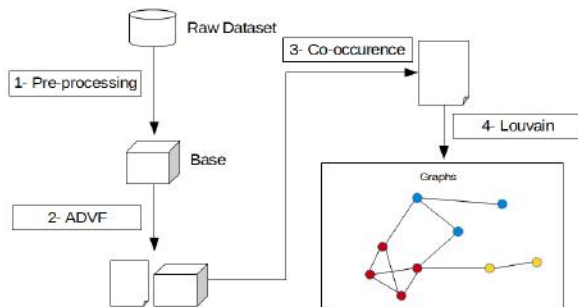


Figure 1: Proposed approach for Topic Modeling.

Since the datasets are acquired from OSNs, the first step is dedicated to pre-processing. In this step, is performed the traditional filtering and cleaning process maintaining the characteristics of text structure. Similar to conventional techniques of Text Mining for general textual content, this step consists in stopwords filtering and cleaning process (removal of links, special characters, and unnecessary spacing). Finally, a process of tokenization is applied. Others pre-processing strategies, i.e., stemming were not necessary since the ADVF will treat irrelevant terms as noise in the next step.

The ADVF method highlights the terms that might be considered as noise. These noisy words are terms that appear a lot or have a few times due to grammatical errors. After the tokenization process for each term from the token's list, it is checked the respective frequency among all tweets collected. The next step is to apply the ADVF method on the terms' frequencies. This method, explained in Section 3.1, creates a probabilistic frequency of terms, more sensitive to noise.

Then, the N terms with the smaller difference between real and probabilistic frequencies are selected. The N can

be interpreted as the sensitiveness of noise detection level.

Later, the co-occurrence of the selected terms is calculated, creating an adjacency list. The N selected terms compose the graph nodes, and its co-occurrence frequency is the edge weight. The adjacency list produces a graph.

The last step for the proposed approach consists of Louvain method application. Each graph will be divided into communities by Louvain method, always favoring the modularity optimization. The communities find by the graph structure analysis and metrics, are classified as natural or artificial topics.

3.1 Adaptive Distribution of Vocabulary Frequencies

The mathematical model ADVF proposed by Igawa et al [10] aims to evaluate the noise level of a dataset from social media corpus. Combined with other approaches, can improve pre-processing techniques toward noisy terms elimination. Frequently, the usage of techniques of pre-processing data on Text Mining has become critical to improving better results accuracy.

The ADVF model is based on the principle of Zipf's Law. The Zipf's Law [17] is a classical measure of literature that studies the frequency distribution of terms in a dataset. It was developed by George Kingsley Zipf, it is a potency law (Equation 1) that analyses the frequency distribution of terms concerning its ranking in descending order, i.e., the first term is most frequent of the entire database and the last, the least frequent. Let $f'(t)$ a desirable frequency of a term t and $r(t)$, the ranking term.

$$f'(t) \sim \frac{1}{r(t)} \quad (1)$$

It means that the second term will be repeated with a frequency of approximately half the first and third term, with a frequency of 1/3 and so on.

The Zipf's Law is a standard probability distribution of the terms which the straight line adapts to all the terms of distribution, but doesn't treat the noise evidence on the set.

Most of the highest frequencies correspond to terms that are prepositions, articles, and pronouns. For Text Mining, depending on the application purpose, these words are considered stopwords. To eliminate these stopwords, the ADVF considers the evidence of these noises in the frequency histogram. From two points in the Cartesian plane, t_1 (the most frequent term) and t_n (the least frequent term), you can find a straight line and its angular coefficient α (Equation 2 where $f(t)$ is the real frequency of the term t). The new line is not adapted so well to all terms, but the presence of noise will be evidenced. The ADVF line is given by Equation 3.

$$\alpha = \frac{\log(r(t_1)) - \log(r(t_n))}{\log(f(t_1)) - \log(f(t_n))} \quad (2)$$

$$ADVF(t) = \alpha(\log(r(t))) + \log(f(1)) \quad (3)$$

As ADVF is a linear distribution based only on the frequency of the terms, avoids extra processing and keep a low complexity $O(n)$.

3.2 Louvain Method

In the literature, the term "community" shows different meanings and connotations. In social science, community refers to a group of people who share the same kind of interest or activity. Once the networks are considered models for several real systems, the concept of community expands [16]. A new concept of community appears after the growing of social media, showing a diversity of on-lines entities with several relations and interactions among entities. The wide range of these networks on social media attracts more attention to areas such as computer science, psychology, economics, marketing and science of behavior [19]. One of the main tasks is to find communities whose members have more interaction with each other in the same community. The extracted communities can be used for analysis and visualization, marketing, training and development groups, clustering.

In graph application, a community is a group of nodes that the connectivity between them is dense. However, as Figure 2 shows, the connectivities between nodes of other communities are sparse. The ability to find and analyse this groups can provide more knowledge about the network's structure [15].

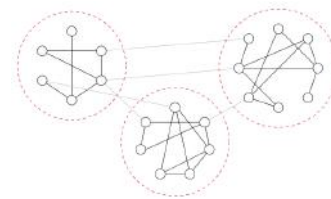


Figure 2: Graph divided into communities [15].

The concept of modularity [19] provides a measure of the quality of community within a network, quantifying a value given by the comparison of the fraction of edges within the community with edges between communities. The modularity Q receive a value between 0 and 1. When Q is closer to 1, the community connectivity is strong. In networks with weights, Q is defined according to Equation 4 [2]:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \quad (4)$$

where A_{ij} represents the weight between the nodes i and j , $k_i = \sum_j A_{ij}$ is the sum of weight between the edges that link the node i , c_i is a community that node i belongs, the function δ assign 1 if the communities are the same, otherwise 0 and $m = \sum_{i,j} A_{ij}$.

The Louvain method, developed by Blondel et al. [2], consists of two phases that are repeated iteratively. First, given a graph with N nodes, is assumed that each node is a community. For each node i and its neighbours j , it is rated the gain modularity among withdraw i of their community and putting in j communities. The node i assumes the new community where the modularity gain is maximum and positive, otherwise i still in the same community. Phase 1 is complete until no improvement can be achieved for all the nodes, i.e., the local maximum is reached when no moving can improve the modularity. The gain modularity ΔQ obtaining from the movement of i to the community C is demonstrated in Equation 5:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right], \quad (5)$$

where \sum_{in} is the sum of the weights of the edges in C , \sum_{tot} is the sum of the weights of the edges incident to nodes of C , k_i is the sum of weights of the edges incident to node i , $k_{i,in}$ is the sum of the weights of the edges of i to the nodes of C and m is the sum of weights of all edges in the graph. In practice, ΔQ evaluates the change of modularity removing i from community and then moving it to the neighbour community.

Phase 2 is the new graph construction, where communities (grouped nodes) of phase one become the new nodes. The weight edges between two new nodes are the sum of weight edges between the node of two communities. After the conclusion of Phase 2, Phase 1 can be performed again. The phases are iterated until no gain modularity is reachable. Figure 3 shows the operation of Louvain method.

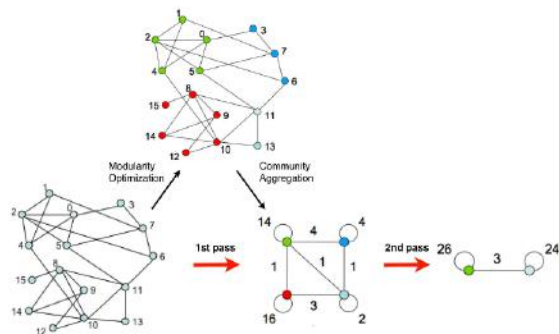


Figure 3: Louvain method application with 2 phases [2].

Although the exact computational complexity of Louvain method is not known, this method sometimes behaves as $\mathcal{O}(n \log(n))$, where most effort is in the first phase of the algorithm [10].

4. EXPERIMENTAL SETTINGS

The datasets used in our experiments were composed of texts from Twitter and Youtube social media, composing 5 datasets (Table 1) with different sizes and themes. The “TwitterGot”, “TwitterNatal” and “TwitterGame” are datasets formed by a keyword. By API services, it is possible to collect texts from social media using keywords, where all these texts contain the main term in its contents. The others two datasets, “TwitterTweets” and “Youtube”, are acquired without keywords, i.e., the sets are composed of texts about different themes and subjects. Another consideration is that all datasets have different sizes, between 2.100 and 20.100 posts. It was done to investigate the behavior of methodology in different datasets scenarios.

In the pre-processing, all alphanumeric characters were transformed to lowercase. By using regular expression, URLs and links were removed. These kinds of data do not represent an analysable term. Since all tweets have a maximum

Table 1: Online Social Network datasets characteristics used in the experiments

Dataset	Size (posts)	Keyword
TwitterGot	2.100	Yes
TwitterNatal	8.600	Yes
TwitterGame	20.100	Yes
TwitterTweets	4.200	No
Youtube	3.400	No

of 140 characters length, the usage of links’ shortcuts are popular, resulting in random links, without textual information. The messages from “Youtube” dataset presents a small number of words, similar to Twitter datasets. Most of the posts have colloquial language, with the presence of slangs, no-alphanumeric characters, and unnecessary spacing. The tokenization process was performed to transform each word from each post in a term.

It is possible to check that there are presence of multiple languages, predominantly the English language, after Spanish and Portuguese. So, we used stopwords addressing these three languages. Articles, pronouns, and prepositions were removed because they are considered noises for topic’s formation.

To calculate the difference between the real frequency and the ADVF’s frequency for each term, it was used the Euclidean Distance (DE). Terms presenting smaller DE are selected. The smaller is DE ’s terms, the greater is the probability of this term be a topic. The Equation 6 shows the calculus to obtain the Euclidean Distance:

$$DE(V_i, V_j) = \sum_{m=1}^n (V_{im} - V_{jm})^{1/2}, \quad (6)$$

where V_i is the real frequency, V_j is ADVF frequency and m is the term of N .

So, the co-occurrence was applied. For the co-occurrence verification, the adjacency’ list was formed by edges with weights greater than 1, due to a large number of edges with weight equal to 1. This cutting justifies the elimination of a dense graph, hard to be analyzed.

For the calculate of modularity and community determination, this study used the `igraph`³ library from the R platform. This library has functions able to import, view, explore, filter, manipulate and export any network. The Louvain method was implemented in this library. The result of this work is based on the formation of community and the complexity of each graph.

From the communities divided by Louvain method, each community was analysed by metrics about the graph structure:

- **Degree.** In a weighted graph, the node degree is the sum of the weight adjacency edges of a node. The weight edge in two nodes means the number of tweets that both terms appear simultaneously.
- **Density.** A dense graph is a graph in which the number of edges is close to the maximal number of edges. The opposite, a graph with only a few edges, is a sparse

³<http://igraph.org/r/>

graph. In undirected graph, the graph density D is defined in Equation 7 as:

$$D = \frac{2|E|}{|V|(|V| - 1)} \quad (7)$$

where E is the number of edges and V , the number of nodes in the graph.

Due to this kind of datasets were formed by texts from OSNs, the presence of spans was expressive. The spammer can post repeatedly texts with the same content, increasing the terms frequency. Although these terms do not represent the natural base theme, the great presence indicates that there are possible terms.

For this work, the found topics will be classified into two classes: natural and artificial topics. The first one represents topics that have a link with the base theme and the other, represents topics created with spans by bots.

5. RESULTS AND DISCUSSION

The Table 3 shows the results from the proposed approach. For each community created by Louvain method, it was selected one or more terms, with significant degrees, to become topics.

About the graphs from the “TwitterGot” and “TwitterNatal” datasets, it is possible to verify that the communities are quite distributed with the presence of centroid terms. Centroids are nodes for which the sum of distances to other nodes is minimal. In the results, the main centroid usually is the own base keyword, but there are presence of others unknown centroids discovered. For example, in “TwitterNatal”, the keyword “natal”, which means “Christmas”, is the main centroid of the graph, and it is considered as a topic. Using the proposed methodology, the application discovered other centroids like “ *festa* ” (“party”) and “ *dezembro* ” (“December”), considered as topics more specifics too.

Despite the “TwitterGame” have a keyword, the graph structure of this dataset is more complex due to the wide co-occurrence between its terms selected. The term “game” used as a keyword can have several semantics, depending on the contest. Due to this diversity, the proposed model still can find and determine topics, but they are more generics. The same situation can be verified in “TwitterTweets” and “Youtube”. The datasets do not have keywords, so the found topics were semantically diverse one each other, in other words, were not specifics topics from a unique theme.

It is not only humans that submit the posts to the OSNs. In OSNs, there are problems about spamming activities by bots. These type of activities can produce malicious contents to deceive users or to promote something. In Twitter, if a bot produces many tweets with the same content, for many applications on topic models, these terms can be considered as main topics. In “TwitterGot” dataset, there are tweets made by bots. In this case, the proposed model adopted natural or artificial topics specification. Natural are the topics that have links with the theme base and artificial are topics created by spamming activities. The Table 4 presents the communities of “TwitterGot”.

In the Community 1 from Table 4, the keyword “gameofthrones” is one of the centroids discovered, as described above. This graph has a lower density which means that the number of nodes is higher than the number of edges. The selected terms have one or few links in the graph, which means

that they are more specific. They can be subtopics from the natural topic “gameofthrones”. The Communities 3, 4 and 5 don’t have a centroid term because all nodes linked all nodes. In this case, the density is 1.00, because the number of possible edges is maximal. The selected terms are more generic and probably appears in same tweets. These topics are considered artificial.

In the Community 2, the term “trndnl” is another centroid term discovered. Due to this graph presented the same characteristics of the Community 1, it was classified as a natural topic. The “trndnl” means Trendinalia⁴ and it is an account service that analyses the classification of Trend Topics on Twitter. Trend Topics are terms delimited by “#” which is used to highlighted a topic. The Table 2 presents examples of tweets made by Trendinalia. These tweets have similar structure to each others. Due to the large amount of this kind of tweets in the dataset, these bots terms are considered artificial topics.

Table 2: Examples of noised *tweets* from Trendinalia.

Tweets
6. #GameOfThrones7. Leon Larregui8. Santos y Queretaro9. Chivas10. John Nash 2015/5/25 04:14 CDT #trndnl http://t.co/IN7801UqsL
6. Leopoldo Lopez7. Ceballos8. Pastor Maldonado9. Cersei10. Exxon 2015/5/25 04:44 VET #trndnl http://t.co/TZZWvFfY1p
1. 1. #charliecharliechallenge2. #MeCaesMalSi3. #GameOfThrones4. #camrenfeels5.#30MVamos Todos #trndnl http://t.co/TZZWvFfY1p

Therefore, the term “trndnl” can be classified in two different classes. Considering its density, the graph behaves as a natural topic like the term “gameofthrones”, but due to the standardized tweets and its huge amount, it is classified as an artificial topic. It is necessary another metric for graphs to correctly classify these cases.

To compare with traditional techniques from the other papers, we applied the TD-IDF, method used in Huang’s paper [9], in our datasets. Selecting the top 5 terms of TF-IDF values in “TwitterGot”, the result was: {9000x}, {astoria}, {babaye}, {bentley} and {brutaaal}. These terms have the same TF-IDF value. It can be explained by algorithm idea. The TD-IDF method is based on the words frequencies in each document and total collection, to calculate how important a word is to a document. Considering that a document is a tweet and the tweets length are short, a word that appears more than once in the same tweet is very unusual. So the TF-IDF in texts from OSNs, basically, is based on only words frequencies. Comparing these words with the selected terms and communities by our approach, it is possible to verify that the words with biggest TF-IDF values do not have links to the theme and topics found. Only high frequency is not enough to set a topic of OSNs datasets.

6. CONCLUSION

Our work presented a novel methodology able to extract text topics from online social medias. Different from traditional texts like articles and reports, OSN’s texts have particular characteristics in its structure. Abbreviation, slangs

⁴<https://twitter.com/trndnl?lang=pt>

and grammatical errors are very common in this kind of texts. It is necessary to adapt existing techniques or create new models.

Five datasets from Twitter and Youtube were used to evaluate the proposed methodology. After the pre-processing, the terms are evaluated by its frequency using Louvain method. The 50 nearest terms to the proposed line were selected. Then, the co-occurrence process was applied, creating an adjacency list. The terms were considered nodes and the co-occurrence frequency were the weight edges. Using the ADVF selection and cutting edges with frequencies less than 1 to prevent a complex graph. Applying the Louvain method, it is possible that the proposed solution performs a Topic Modeling

The datasets used to this works were composed of texts from different authors, and the presence of noise was inevitable. By analyzing the found communities, it was determined that topics can be artificial, due to the graph's density are higher or maximal. In other words, our approach can highlight topics artificially created reaching an accurate Topic Modeling.

However, there are cases that the terms' structure behave like one class, but the proposed model classifies as another class. These cases must be studied to find others metrics that discover its behavior.

Our method has better results than applying the TF-IDF method in OSN's datasets. Due to the TF-IDF values in OSN's datasets is based on only terms frequencies, the terms with biggest values does not have any links to the theme and topics found by our approach.

7. ACKNOWLEDGEMENTS

We are grateful to CNPQ that made this paper possible by sponsoring our work, process 479821/2013-5.

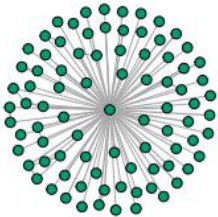
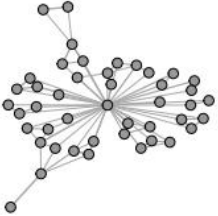
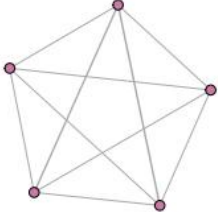
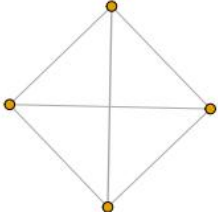
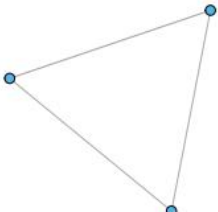
8. REFERENCES

- [1] A. Akilan. Text mining: Challenges and future directions. In *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, pages 1679–1684. IEEE, 2015.
- [2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, July 2008.
- [3] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, pages 43–52, New York, NY, USA, 2013. ACM.
- [4] Z. Chen and B. Liu. Mining topics in documents: Standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 1116–1125, New York, NY, USA, 2014. ACM.
- [5] K. Chitra and B. Subashini. Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, 3(8):219–226, 2013.
- [6] D. Choi, B. Ko, H. Kim, and P. Kim. Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, 38:16–21, 2014.
- [7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.
- [8] S. Huang, Y. Yang, H. Li, and G. Sun. Topic detection from microblog based on text clustering and topic model analysis. In *Services Computing Conference (APSCC), 2014 Asia-Pacific*, pages 88–92. IEEE, 2014.
- [9] S. Huang, Y. Yang, H. Li, and G. Sun. Topic detection from microblog based on text clustering and topic model analysis. In *Services Computing Conference (APSCC), 2014 Asia-Pacific*, pages 88–92, Dec 2014.
- [10] R. Igawa, G. Sakaji Kido, J. Seixas, and S. Barbon. Adaptive distribution of vocabulary frequencies: A novel estimation suitable for social media corpus. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 282–287, Oct 2014.
- [11] R. A. Igawa, S. Barbon Jr, K. C. S. Paulo, G. S. Kido, R. C. Guido, M. L. P. Júnior, and I. N. da Silva. Account classification in online social networks with lba and wavelets. *Information Sciences*, 332:72–83, 2016.
- [12] R. A. Igawa, A. M. G. de Almeida, B. B. Zarpelão, and S. Barbon Jr. Recognition of compromised accounts on twitter. 2015.
- [13] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.
- [14] H. Li, J. Yan, H. Weihong, and D. Zhaoyun. Mining user interest in microblogs with a user-topic model. *Communications, China*, 11(8):131–144, Aug 2014.
- [15] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [16] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554, 2012.
- [17] D. M. W. Powers. Applications and explanations of zipf's law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 1998.
- [18] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He. Interpreting the public sentiment variations on twitter. *Knowledge and Data Engineering, IEEE Transactions on*, 26(5):1158–1170, 2014.
- [19] L. Tang, X. Wang, and H. Liu. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33, 2012.
- [20] F. S. Tsai. A tag-topic model for blog mining. *Expert Systems with Applications*, 38(5):5330 – 5335, 2011.
- [21] M. Zappavigna. Ambient affiliation: A linguistic perspective on twitter. *New media & society*, 13(5):788–806, 2011.
- [22] J. Zeng, J. Duan, W. Cao, and C. Wu. Topics modeling based on selective zipf distribution. *Expert Systems with Applications*, 39(7):6541 – 6546, 2012.

Table 3: Example of topics extracted from different datasets.

Dataset	Keyword	Topics	N ^o Communities	Graph
TwitterGot	{gameofthrones}	{listen, jeffery, stone}, {teamyank3}, {thought, dude}, {dragons}, {go- texhibit}, {gift, s05s07}, {kings, landing}, {happy, take}, {winterfell, theon}, {else, series}, {people}, {dorne}, {trndnl}, {totti9, chapel10, anceldotti8, gal- liani, grevia}, {mamah- ablaespanol5, madres3, kanquimania4, exatempo- radabarbarella}, {santos4 , , junio5, llderteneceitamos- vivo2}, {tolerancia03}	18	
TwitterNatal	{natal}	{dezembro}, {publicar, foto, rn, tirol, acabei}, {quero, mais}, {seria, melhor}, {não}, {futuro}, {estar, sul}, {festa}, {tenho, amigo, secreto}	10	
TwitterTweets	No keyword	{people, dont, need}. {love}, {know, time, song}, {gift, card}, {going, back, last}, {think}, {posted, al- bum}, {lastest, insurance}, {follow, make}, {today}, {twitter}, {good}	12	
Youtube	No keyword	{time}, {love, song}, {good}, {video}, {know, great}, {make, better, ever}, {also, check}, {people}	8	
TwitterGame	{game}	{amazing}, {lose}, {trying, remember}, {white}, {fly- ers, thing, show}, {bring, sunday}, {bout, league}, {espn, finals}, {luck, girls, want}	10	

Table 4: Communities of TwitterGot

Id	Graph	Topic	Adjacency Terms	Class
1		gameofthrones Degree: 42.30 Density: 0.02	{wrong}, {world}, {winterishere}, {wedding}, {today}, {thank}, {steel}, {seven}, {right}, {whole}, {real}, {ready}, {whitewalker}, {need}, {minutes}, {things}, {make}, {thewalkingdead}, {long}, {iron}, {summer}, {help}, {story}, {karma}, {head}, {want}, {still}, {hate}, {guys}, {sparrow}, {walking}, {give}, {video}, {final}, {quede}, {fight}, {first}, {spamdecancionesdedespecho2}, {fire}, {moment}, {feeling}, {season5}, {demthrones}, {exhibition}, {crazy}, {ramsay}, {character}, {walker}, {black}, {despues}, {damn}, {arrived}, {left}, {throne}, {ghost}, {kill}, {gets}, {catching}, {casa}, {think}, {beer}, {esperar}, {intense}, {semana}, {sunday}, {serie}, {didn}, {every}, {said}, {stop}, {temporada}, {made}, {show}, {break}, {consultasparanormales5}, {fucking}, {khaleesi}, {books}, {mejor}, {better}, {back}, {foda}, {house}, {bolton}	Natural
2		trndnl Degree: 19.68 Density: 0.08	{top20}, {spoiler}, {took}, {hashtag}, {gotbr}, {joffrey}, {topic}, {sergio}, {uyvengarespete3}, {chile}, {queretaro2}, {posicion}, {cali2}, {bueno}, {trending}, {hours}, {welcomebackto1dzayn5}, {yajuniotyodavia2}, {sembrarsemillasdepaz7}, {wwechamber}, {mundialcomprado5}, {welcomebackto1dzayn2}, {teenchoice}, {nash10}, {gameofthrones8}, {feliz}, {nogueira9}, {felices}, {exxon}, {mujer}, {wonderwall7}, {chapel9}, {farez9}, {naena}, {cristina}, {bueno3}, {judas}, {azevedo}, {geraldodo}, {felizdiadelnino4}	Natural /Artificial
3		Without topic Degree: - Density: 1.00	{totti9}, {chapel10}, {ancelotti8}, {galliani}, {grevia}	Artificial
4		Without topic Degree: - Density: 1.00	{mamahablaespanol5}, {madres3}, {kanquimania4}, {exatemporadabarbarella}	Artificial
5		Without topic Degree: - Density: 1.00	{santos4}, {junio5}, {llidertenecesitamosvivo2}	Artificial