

Artificial and Natural Topic Detection in Online Social Networks

Sylvio Barbon Jr¹, Guilherme Sakaji Kido¹, Gabriel Marques Tavares¹

¹Computer Science Department – State Univerisity of Londrina (UEL)
86057-970 – Paraná – PR – Brazil

barbon@uel.br, guilhermekido@gmail.com, gabrielmrqstvrs@gmail.com

Abstract. *Online Social Networks (OSNs), such as Twitter, offer attractive means of social interactions and communications, but also raise privacy and security issues. The OSNs provide valuable information to marketing and competitiveness based on users posts and opinions stored inside a 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 community architecture (density and degree) allows the indication of a topic strength and the classification of it as natural or artificial. The later created by the spammers on OSNs.*

1. Introduction

Currently, people have instant access to massive amounts of data. Learning and discovering becomes easier when the data format is more logical and structured. Unfortunately, 90% of the available information are in unstructured forms [Verma et al. 2015]. Extracting knowledge in large amounts of unstructured data is difficult and problematic. For companies, being able to do so might improve marketing strategy and business planning to attract more consumers.

Recent studies indicate that 80% of companies' information are text documents [Akilan 2015]. 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 [Chen et al. 2013] [Chitra and Subashini 2013] [Choi et al. 2014]; in politics, this data can adjust political placements in respect of sentiment analysis of your target audience [Tan et al. 2014].

Online Social Networks (OSNs) are services that have been emerging as a new communication between individuals and organizations [Li et al. 2014]. These services provide an essential platform for users to share thoughts, ideas, status, and experiences [Zappavigna 2011]. In this sense, the OSNs offer attractive means of online social interactions and communications, but also raise privacy and security concerns [Zhang et al. 2010]. 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 [Igawa et al. 2015]. The standard methods used for Text Mining are usually applied to traditional texts in the Web, such as articles, news, and reports [Tsai 2011].

Micro-blog texts are composed of more casual and informal language than traditional texts but are a source of similar relevant information in comparison to other textual sources. Due to the number of characters limit, users publish in a simplified way, using the colloquial language, abbreviations, slangs and generally links, emoticons, photos, videos, and others [Huang et al. 2014a]. Colloquial and informal language usually creates specific words and terms that are considered noise. Noise is an undesirable textual feature and specific measures are necessary in order to eliminate or reduce it. One example is the *Adaptive Distribution of Vocabulary Frequencies* (ADVF) [Igawa et al. 2014], capable of highlighting terms detected as textual noise. Roth et al. [Roth et al. 2013] presented a survey of noise reduction methods for textual databases. In their work, three different denoising principles are highlighted: At-least-one, Topic-based models and, Pattern Correlations. The advantages of this kind of approach to Sentiment Analysis considering statistical models are exposed, such as Pattern Correlations in [Roth et al. 2013]. More details concerning statistical models employed in noise analysis are given in [Aggarwal 2015], where the text sampling from a Zipf distribution is discussed. This strategy is the kernel of [Igawa et al. 2014]; the technique applied in our propose.

The recent concern 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 used for spreading malicious contents only [Igawa et al. 2016][Barbon et al. 2016]. A compromised account is a legitimate account which has been taken over by an attacker to publish fake or harmful content [Igawa et al. 2015]. Directly or indirectly, both aims to spread content that need to be avoided in knowledge discovery because they do not contain real information. In OSNs, a bot repeatedly publishes texts with a subject that can compromise the topic trends. The presence of artificial content made by a spammer could lead to bias the result precision and application's goals.

The malicious behavior was discussed by several authors. Jin et al. in [Jin et al. 2013] exposed several security and privacy threats. The authors highlighted the Sybil attack which means the register of multiple accounts maliciously. In this sense, a Sybil attack could create a biased topic due to undue influence. Sybil attack was compared with other malicious activities (information leak, de-anonymizing, phishing, malware and spamming) in OSNs by Gao et al. [Gao et al. 2011]. Their results reveal an association within attack difficulty, defense effectiveness and threat level to the users.

Diverse works propose solutions to mitigate the malicious activities, textual-based ones being a significant part of them. For example, the spam classification engine extracts the text from the post and runs it through a support vector machine (SVM) classifier, which assigns a score to the text [Abu-Nimeh et al. 2011]. More recently, Vanetti et al. [Vanetti et al. 2013] proposed a content-based filtering (textual) of unwanted content by the use of Machine Learning. More precisely, the use of Artificial Neural Networks built with a radial function (RBFN) was applied.

Independently of textual-based application, an important task is to identify the textual content, which can be determined by one or more topics/keywords that describe the main subject. There are several methods to find those keywords and Topic Modeling is the main area of this activity. For Zeng et al. [Zeng et al. 2012], 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.

Topic Modeling, as Latent Semantic Analysis (LSA) [Landauer et al. 1998] and Probabilistic Latent Semantic Analysis (PLSA) [Hofmann 2001], is 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 [Chen and Liu 2014]. 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 extracting topics on blogs [Li et al. 2014] and OSNs.

About traditional techniques of Topic Modeling, Huang et al. [Huang et al. 2014b] 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 all dataset. This importance was based on frequency distribution of the terms and clustered by K-means. The dataset used in Huang et al. [Huang et al. 2014b] 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 [Tsai 2011], 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 Buzz-Metrics'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 and was effective just in a small dataset with few topics.

In this work we focus on the possible strategy of topics as community centroids in a graph of terms. In this way, finding communities in a textual data is concerned as a data clustering problem. Several techniques to investigate the community structure of networks have been proposed in literature during last years [De Meo et al. 2011, Bhowmick and Srinivasan 2013, Campigotto et al. 2014, Traag 2015]. In [De Meo et al. 2011], the authors proposed a feasible solution considering a large network with a low computational cost by a Generalized Louvain method. In the same way, Traag in [Traag 2015] shows some improvements to the original Louvain algorithm able to reach a logarithmic runtime complexity in a clear community structure. On the other hand, in [De Meo et al. 2011] is presented a solution able to carry out over synthetic and real-world (noisy), showing efficiency and robustness. The urge for techniques capable of handling large datasets based on accurate and fast solutions was highlighted by [Bhowmick and Srinivasan 2013]. These authors proposed a shared-

memory parallel algorithm and exposed the advantages of a scalable solution in Louvain algorithm.

This present work aims to a new approach 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 an OSNs scenario highlighting the natural or artificial ones. The kernel of proposed approach was based on Louvain method and the concept of modularity that provides a quality measure of the communities in a graph [Tang et al. 2012]. 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 [Igawa et al. 2014], 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 an extended version of a previous work [Kido et al. 2016], including the study of additional related work, an improved explanation of the proposed approach, and new results discussion. It is organized as follows: the next section provides our proposed approach for this paper and deals with the explanation of the ADVF and *Louvain* method. 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. Proposed Approach

Our proposed approach, 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.

Since the datasets are acquired from OSNs, the first step is dedicated to pre-processing. In this step, the traditional filtering and cleaning process is performed in order to maintain 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 rarely due to grammatical errors. After the tokenization process for each term from the token's list, it is checked the respective fre-

¹www.twitter.com

²www.youtube.com

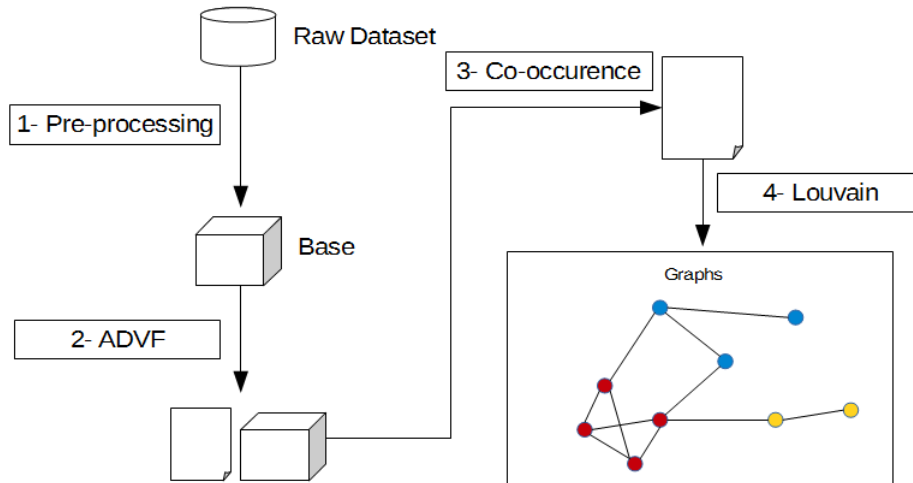


Figure 1. Proposed approach for Topic Modeling.

quency is checked among all tweets collected. The next step is to apply the ADFV method on the terms' frequencies. This method, explained in Section 2.1, 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 found by the graph structure analysis and metrics, are classified as natural or artificial topics.

2.1. Adaptive Distribution of Vocabulary Frequencies

Frequently, the usage of techniques of pre-processing data on Text Mining has become critical to improving better results accuracy. The mathematical model ADFV proposed by Igawa et al. [Igawa et al. 2014] aims to evaluate the noise level of a dataset from social media corpus. In this way, the traditional pre-processing techniques combined with ADFV can improve the dataset quality by avoiding noise.

The ADFV model is based on the principle of Zipf's Law. The Zipf's Law [Powers 1998] 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, it avoids extra processing and keep a low complexity $O(n)$.

2.2. 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 interests or activities. Once the networks are considered models for several real systems, the concept of community expands [Papadopoulos et al. 2012]. 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 [Tang et al. 2012]. 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 where the connectivity between them is dense. However, 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 [Newman and Girvan 2004].

The concept of modularity [Tang et al. 2012] provides a measure of the quality of a 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 [Blondel et al. 2008]:

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

where A_{ij} represents the weight between the nodes i and j , $k_i = \sum_j A_{ij}$ is the weights' sum between the edges that link the node i , c_i is a community that node i belongs to, 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. [Blondel et al. 2008], 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 , the gain of modularity is calculated among i withdrew 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 when 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 obtained 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 2 shows the operation of Louvain method.

Although the exact computational complexity of Louvain method could reach $O(n \log(k))$, where k is the average community size for clear datasets [Traag 2015], our implementation sometimes behaves as $O(n \log(n))$, where most effort is in the first phase of the algorithm [Igawa et al. 2014].

3. Experimental Settings

The datasets used in our experiments were composed of texts from Twitter and Youtube social medias, composing 5 datasets³ (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

³<http://www.uel.br/grupo-pesquisa/remid/wp-content/uploads/DatasetiSys2016.zip>

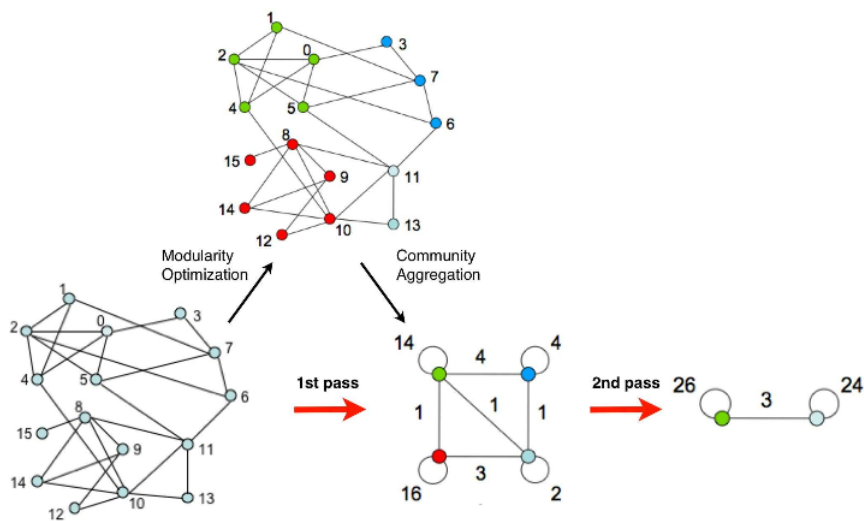


Figure 2. Louvain method application with 2 phases [Blondel et al. 2008].

all these texts contain the main term in its contents. The two others datasets, “TwitterTweets” and “Youtube”, are acquired without keywords, i.e., those sets are composed of texts about different themes and subjects. The “Youtube” dataset was published in [Thelwall et al. 2012], it was formed by English comments posted to videos on the YouTube web site. This represents comments on resources and any associated discussions. Another consideration is all datasets have different sizes, between 2.100 and 20.100 posts. This practice was performed to investigate the solution behavior in different datasets scenarios.

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 |

In pre-processing, all alphanumeric characters were transformed to lowercase. By the use of regular expressions, URLs and links were removed. These kinds of data do not represent an analysable term. Since all tweets have a maximum 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. In our experiments we did not treat each language separately, indeed the topics were found mixing the idioms.

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

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

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, which is hard to analyze.

For the calculation of modularity and community determination, this study used the `igraph`⁴ 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 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.

4. Results and Discussion

The results show important achievements, for each community created by Louvain method, it was selected one or more terms, with significant degrees, to become topics. This aspects are exposed in Subsection 4.1. Another important contribution is related to malicious behaviour of bots on OSN is available in Subsection 4.2.

⁴<http://igraph.org/r/>

4.1. Topic Modeling

An important tool used to observe the communities is the graphs. Graphs form the foundation representation of several types of data, ranging from Internet connectivity to social networks. In our case, we want to understand “what the graph looks like;” we want to know which vertex and edges are important and what are the significant features of the graph, as well as identifying the representative nodes, communities and links. In our results, we expose the communities with different colours in order to highlight the several communities found. About the graphs from the “TwitterNatal” (Figure 3) and “TwitterGot” (Figure 4) datasets, it is possible to verify that the communities are quite distributed with the presence of centroid terms.

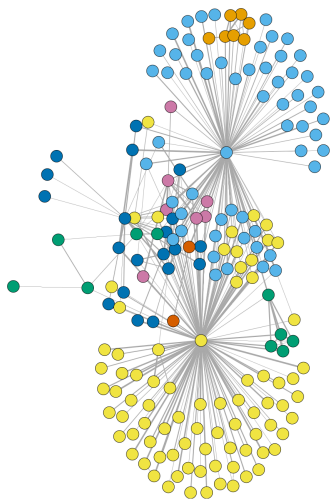


Figure 3. TwitterNatal communities

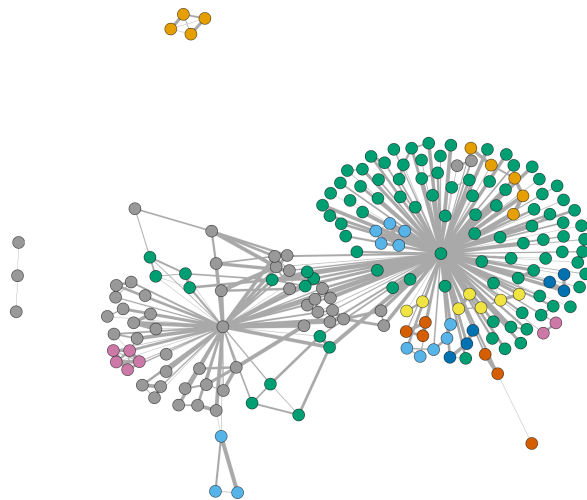


Figure 4. TwitterGot communities

About the graphs from the “TwitterNatal” (Figure 3) and “TwitterGot” (Figure 4) 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” (Figure 5), the keyword “natal”, which means “Christmas”, is the main centroid of the graph, and it is considered as a topic. Using the proposed approach, the application discovered other centroids like “*festa*” (“party”), considered as topics more specifics too, as in Figure 5 is shown.

It is important to observe that community created by centroid “*compra*” (blue graph in Figure 5) has a different graph topology, where we cannot identify a clear centroid. In this case, we can observe the terms that composed the graph had a similar co-occurrence. Thus it is not possible to specify a centroid, but treat this community with the presence of expressions to exposes the community sense, in other words, there are several terms for closely related concepts.

By observing the different communities identified from TwitterNatal dataset, we obtained 10 different senses and themes in a collection created by keyword “natal”. These communities are described in Table 2. For example, it was possible to obtain a disambiguation of terms with the same writing (but with different sense) as “natal” as Christmas’s meaning and “natal” the Brazilian city. Another important discovery was that some

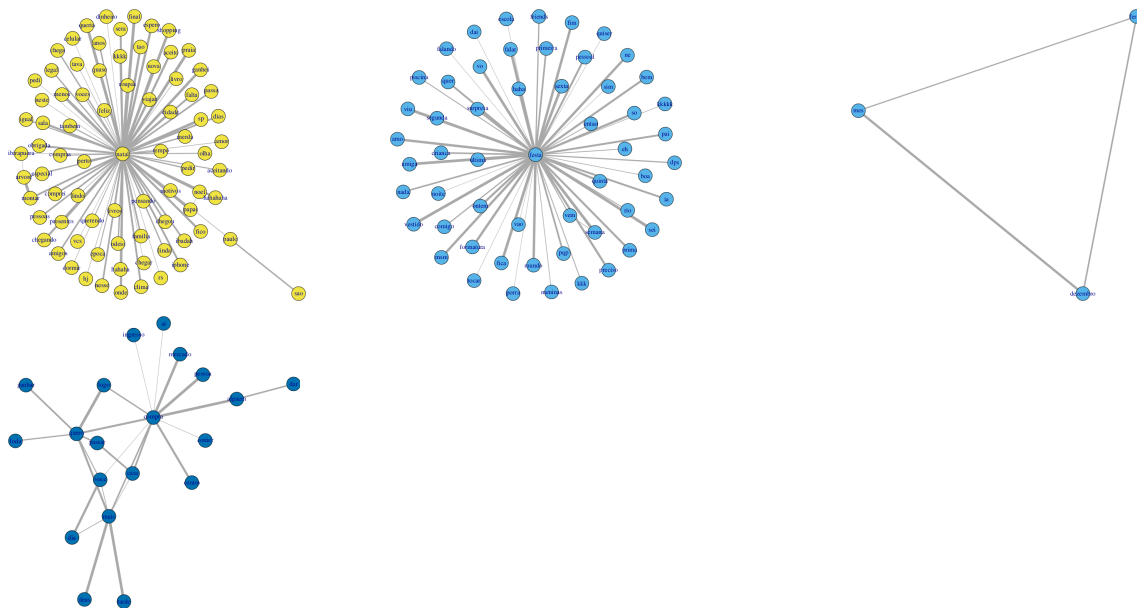


Figure 5. Communities of TwitterNatal dataset: centroid “natal” (yellow), centroid “festa” (cyan), without centroid “dezembro” (cyan with three terms) and without “compra” (blue).

topics found were hyponyms of the keyword used to build the dataset. Hyponym is a word of more specific meaning that a term applicable to it. Example, Hyponym of Natal city is a tourist attraction at the city of Natal.

Table 2. List of communities and terms sense interpretation of TwitterNatal dataset

| Term (centroid) | Sense | Size | Colour |
|---------------------------------------|------------------------------------|------|--------------|
| <i>natal</i> | Christmas' day | 84 | Green |
| <i>festa</i> | graduation parties | 56 | Gray |
| <i>compra, quero, mais and centro</i> | go shop / go to the mall | 20 | Blue |
| <i>estar</i> | be in a country region | 7 | Red |
| <i>nao</i> | “I don't like” | 7 | Pink |
| <i>foto, tirol, rn, publicar</i> | a tourist attraction at Natal city | 5 | Cyan |
| <i>amigo, secreto, oculto</i> | Kriss Kringle play | 5 | Light Orange |
| <i>ferias, mes and dezembro</i> | vacation on December | 3 | Pink |
| <i>futuro</i> | past and future wishes | 3 | Gray |
| <i>melhor and seria</i> | could be better | 2 | Light Blue |

A similar result was obtained in TwitterGot dataset. We identified 18 different communities, represented by centroids (or mor frequent terms): {listen, jeffery, stone}, {teamyank3}, {thought, dude}, {dragons}, {gotexhibit}, {gift, s05s07}, {kings, landing}, {happy, take}, {winterfell, theon}, {else, series}, {people}, {dorne}, {trndnl}, {totti9, chapel10, ancelotti8, galliani, grevia}, {mamahablaespanol5, madres3, kanquimania4, exatemporadabarbarella}, {santos4, junio5, lldertenesitamosvivo2} and {tolerancia03}. The keyword used to build this dataset is related to a TV Series. It was observed the most of the communities are assessed to series character. Thus, it was

possible to identify some different stories and scenarios from the tv show.

Another important observation is the presence of malicious activities. In the TwitterGot dataset spamming activities by a bot user named as *trndnl* were identified. A more precise description of this aspect is done in Sub-section 4.2.

Despite the “TwitterGame”, the graph structure of this dataset is more complex due to the wide co-occurrence between its terms selected, as Figure 6 illustrate. The term “game” used as a keyword can have several semantics, depending on the context. Due to this diversity, the proposed model still can find and determine topics, but they are more generics.

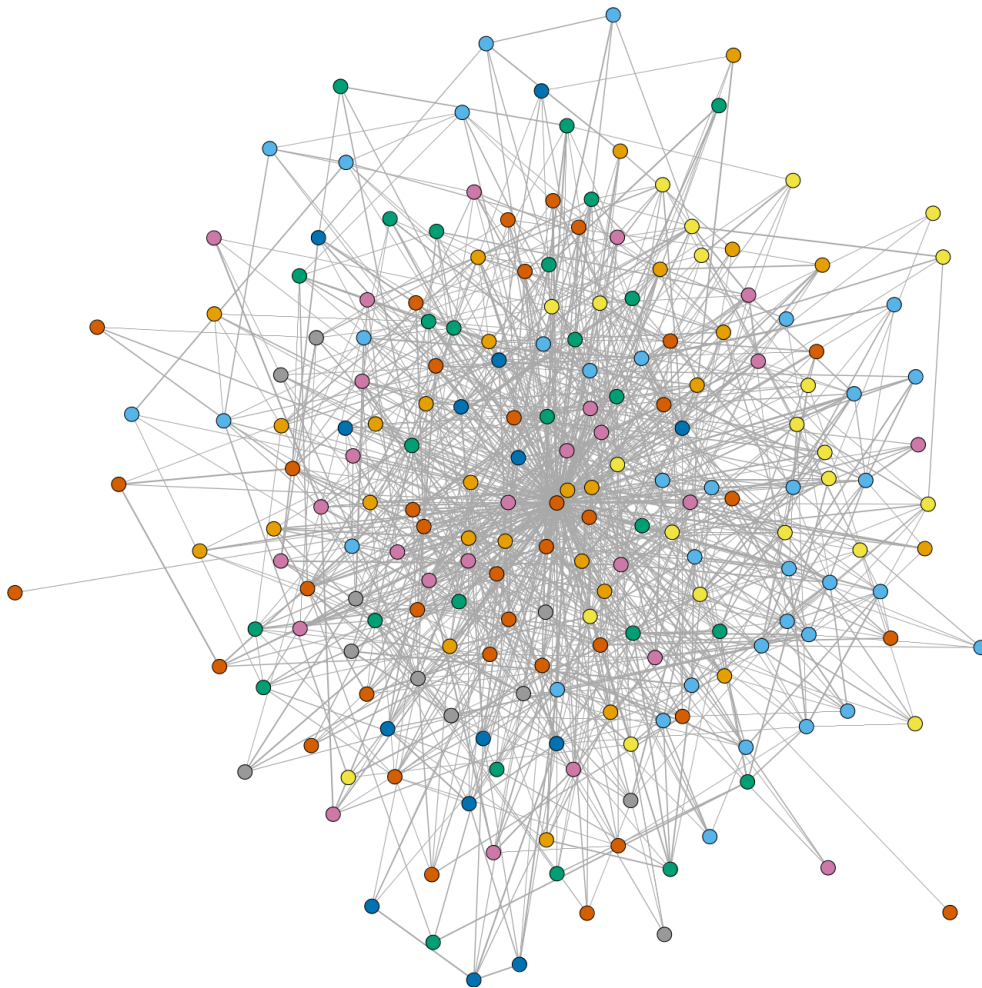


Figure 6. TwitterGame communities obtained

The communities obtained in TwitterGame were 10. Considering the wide sense of the keyword used to compose this dataset, “game”, we distinguish important topics. The communities extracted were: {amazing}, {lose}, {trying, remember}, {white}, {flyers, thing, show}, {bring, sunday}, {bout, league}, {espn, finals}, {luck, girls, want}. For example, the community with centroid “lose” presents terms as “cavs” (Cleveland Cavaliers from National Basketball Association - NBA), “george” (athlete of Indiana Pacers NBA team), “score” and “shot”, there are related to a NBA game (Indiana Pacers versus Cleveland Cavaliers) that happened in the time of dataset acquisition. Another

NBA game happened during the dataset acquisition. The community {trying, remeber} contains the terms “wade” (Dwyane Wade an athlete of Miami Heat) and other terms related to sports. Thus we could interpret as a community related to a NBA game.

In this way, the dataset TwitterGame was split into communities that represent different sports events during the dataset acquisition. Not just NBA games, the community {flyers, thing, show} presents the term “flyers” and “rangers”, both are hockey teams of National Hockey League (NHL): Philadelphia Flyers and New York Rangers.

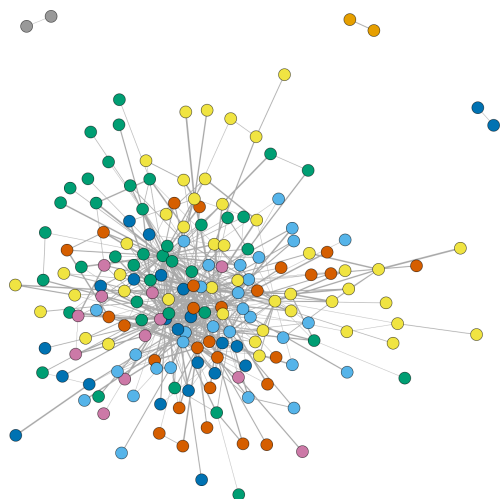


Figure 7. TwitterTweets communities

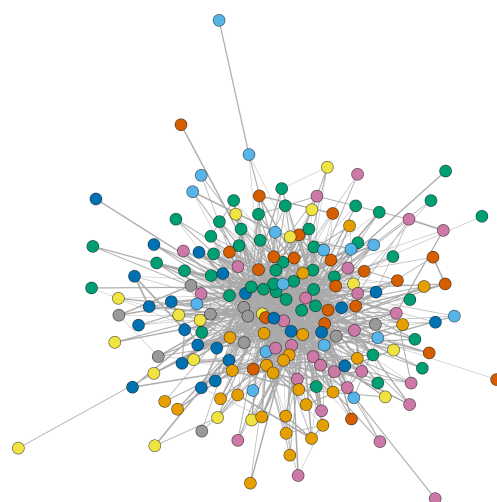


Figure 8. Youtube communities

A scenario of a term with several meanings or several terms used to build a dataset has a more complex graph structure, e.g. Figure 6, Figure 7 and Figure 8. On the other hand, a more specific term can be represented as a less complex graph, as Figure 3 and Figure 4 show. This

This phenomenon is not related to the number of communities (i.e. topics). The complex graphs as obtained by Youtube, TwitterTweet and TwitterGame were composed by 8, 12 and 10 communities, respectively. The less complex graphs of TwitterNatal and TwitterGot were formed by 10 and 18 communities. In this way, the graph complexity obtained by performing Louvain is related to equality of communities density. The community density is the symmetry between the number of terms and co-occurrence that compose it. A density equal 1.00 means the total co-occurrence between the terms. On the other hand, low density (near to 0.00) is related to less co-occurrence between all terms. The low density is a sign of diversity of sentences establishing a community and the evidence of a centroid presence.

Another important metric is the average degree. We can assess this metric as the strength of the topic found. Thus a high degree indicates a more popular topic. In the Table 3 it is possible to observe the average degree 42.30 of “gameofthrones” community. This dataset was built with a keyword related to Game of Thrones TV series. The other communities obtained with this dataset have lower degree in comparison to “gameofthrones”, designed through a centroid with the same name of TV series. Another example is the communities “natal” and “festa” in Figure 3 (yellow and cyan chart), they have the highest average degree of TwitterNatal dataset. Thus, are the most strength topics.

Considering the presence of a centroid as the implication of a distinct topic, more precisely, an occurrence of a single term that represents a collection of terms (community), we can affirm that a complex graph has more unique words that describe a topic, leading to semantical difference between them. A less complex graph presents high density and the indication of repetitive sentences.

The repetitive sentences in a community is a sign of artificial sentences presence. This artificial sentences could be the consequence of malicious action, as discussed in Subsection 4.2.

4.2. Malicious Detection

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 3 presents the communities of “TwitterGot”.

In the Community 1 from Table 3, 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 (terms) is higher than the number of edges (co-occurrence). 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 co-occurrence is maximal. The selected terms are more generic and probably appear in same tweets. Theses topics are considered artificial.

In the Community 2, the term “trndnl” is another centroid term discovered. Since this graph presented the same characteristics of the Community 1, it was classified as a natural topic. The “trndnl” means Trendinalia⁵ 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 highlight a topic. The Table 4 presents examples of tweets made by Trendinalia. These tweets have a similar structure to each other. Due to a large amount of this kind of tweets in the dataset, these bots’ terms are considered artificial topics.

⁵<https://twitter.com/trndnl?lang=pt>

Table 3. The natural and artificial communities of TwitterGot dataset

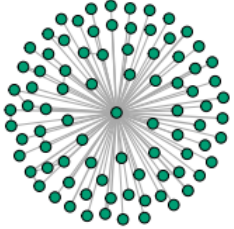
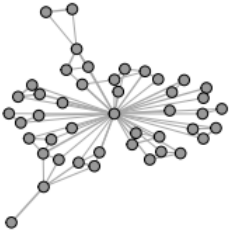
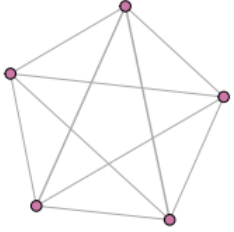
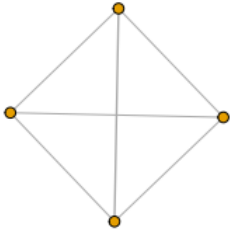
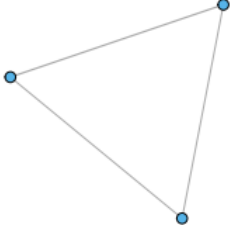
| Id | Graph | Topic | Class |
|----|---|---|------------------------|
| 1 |  | gameofthrones Degree: 42.30 Density: 0.02 | Natural |
| 2 |  | trndnl Degree: 19.68 Density: 0.08 | Natural /Artificial |
| 3 |  | Without topic Degree: - Density: 1.00 | Artificial |
| 4 |  | Without topic Degree: - Density: 1.00 | Artificial |
| 5 |  | Without topic Degree: - Density: 1.00 | Artificial |

Table 4. 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. |

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 [Huang et al. 2014b], 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 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. Thus, this type of traditional technique for Topic Modeling cannot be applied to Topic Modeling in Online Social Networks due to the lexicon size of a tweet. In other words, only high frequency is not enough to set a topic of OSNs datasets.

5. Conclusion

Louvain method was the base of our OSN topic extraction proposal. Different from the traditional non-structured text (news, whitepapers, and reports), the OSNs’ content has particular characteristics. Abbreviation, slangs and grammatical errors are very common, besides the use of few terms (small lexicon).

Five datasets from Twitter and Youtube were used to evaluate the proposal. Using the ADVF terms selection, terms co-occurrence, and Louvain method, it was possible to perform the Topic Modeling in OSNs.

The datasets used in the experiments were composed of texts from different authors, and the presence of noise was inevitable. However, there were discovered several topics. By analyzing the found communities, we can infer if a topic was artificial or natural, due to the graph’s density.

In other words, our approach can highlight the topics avoiding noise and obtaining an accurate Topic Modeling. When the dataset was obtained by the use of a keyword, the topics found were in the most cases hyponyms of the keyword. In addition, our proposed method can identify an artificial or natural topic in the dataset. Artificial topics need to be avoided once they were created by the malicious action of a bot.

However, there are cases where the community has characteristics of both artificial and natural topics. The use of alternative metrics to distinguish this aspect is the object of our future work.

6. Acknowledgements

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

References

- Abu-Nimeh, S., Chen, T. M., and Alzubi, O. (2011). Malicious and spam posts in online social networks. *Computer*, 44(9):23–28.
- Aggarwal, C. C. (2015). Outlier analysis. In *Data Mining*, pages 237–263. Springer.
- Akilan, A. (2015). Text mining: Challenges and future directions. In *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, pages 1679–1684. IEEE.
- Barbon, S., Igawa, R. A., and Bogaz Zarpelão, B. (2016). Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications*, pages 1–21.
- Bhowmick, S. and Srinivasan, S. (2013). A template for parallelizing the louvain method for modularity maximization. In *Dynamics On and Of Complex Networks, Volume 2*, pages 111–124. Springer.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+.
- Campigotto, R., Céspedes, P. C., and Guillaume, J.-L. (2014). A generalized and adaptive method for community detection. *arXiv preprint arXiv:1406.2518*.
- Chen, Y., Amiri, H., Li, Z., and Chua, T.-S. (2013). 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. ACM.
- Chen, Z. and Liu, B. (2014). 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. ACM.
- Chitra, K. and Subashini, B. (2013). Data mining techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, 3(8):219–226.
- Choi, D., Ko, B., Kim, H., and Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, 38:16–21.
- De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. (2011). Generalized louvain method for community detection in large networks. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 88–93. IEEE.
- Gao, H., Hu, J., Huang, T., Wang, J., and Chen, Y. (2011). Security issues in online social networks. *IEEE Internet Computing*, 15(4):56–63.
- Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196.
- Huang, S., Yang, Y., Li, H., and Sun, G. (2014a). Topic detection from microblog based on text clustering and topic model analysis. In *Services Computing Conference (APSCC), 2014 Asia-Pacific*, pages 88–92. IEEE.

- Huang, S., Yang, Y., Li, H., and Sun, G. (2014b). Topic detection from microblog based on text clustering and topic model analysis. In *Services Computing Conference (AP-SCC), 2014 Asia-Pacific*, pages 88–92.
- Igawa, R., Sakaji Kido, G., Seixas, J., and Barbon, S. (2014). Adaptive distribution of vocabulary frequencies: A novel estimation suitable for social media corpus. In *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*, pages 282–287.
- Igawa, R. A., Barbon Jr, S., Paulo, K. C. S., Kido, G. S., Guido, R. C., Júnior, M. L. P., and da Silva, I. N. (2016). Account classification in online social networks with lba and wavelets. *Information Sciences*, 332:72–83.
- Igawa, R. A., de Almeida, A. M. G., Zarpelão, B. B., and Barbon Jr, S. (2015). Recognition of compromised accounts on twitter. In *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective*, volume 1, pages 9–14.
- Jin, L., Chen, Y., Wang, T., Hui, P., and Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9):144–150.
- Kido, G. S., Igawa, R. A., and Barbon Jr, S. (2016). Topic modeling based on louvain method in online social networks. In *XII Brazilian Symposium on Information Systems - Information Systems in the Cloud Computing Era*, pages 353–360.
- Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.
- Li, H., Yan, J., Weihong, H., and Zhaoyun, D. (2014). Mining user interest in microblogs with a user-topic model. *Communications, China*, 11(8):131–144.
- Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.
- Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3):515–554.
- Powers, D. M. (1998). Applications and explanations of zipf’s law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning*, pages 151–160. Association for Computational Linguistics.
- Roth, B., Barth, T., Wiegand, M., and Klakow, D. (2013). A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78. ACM.
- Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J., Chen, C., and He, X. (2014). Interpreting the public sentiment variations on twitter. *Knowledge and Data Engineering, IEEE Transactions on*, 26(5):1158–1170.
- Tang, L., Wang, X., and Liu, H. (2012). Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33.
- Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

- Traag, V. A. (2015). Faster unfolding of communities: Speeding up the louvain algorithm. *Physical Review E*, 92(3):032801.
- Tsai, F. S. (2011). A tag-topic model for blog mining. *Expert Systems with Applications*, 38(5):5330 – 5335.
- Vanetti, M., Binaghi, E., Ferrari, E., Carminati, B., and Carullo, M. (2013). A system to filter unwanted messages from osn user walls. *IEEE Transactions on Knowledge and data Engineering*, 25(2):285–297.
- Verma, V., Ranjan, M., and Mishra, P. (2015). Text mining and information professionals: Role, issues and challenges. In *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on*, pages 133–137.
- Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on twitter. *New media & society*, 13(5):788–806.
- Zeng, J., Duan, J., Cao, W., and Wu, C. (2012). Topics modeling based on selective zipf distribution. *Expert Systems with Applications*, 39(7):6541 – 6546.
- Zhang, C., Sun, J., Zhu, X., and Fang, Y. (2010). Privacy and security for online social networks: challenges and opportunities. *IEEE Network*, 24(4):13–18.