Short Text Tagging Using Stochastic Block Models: Yelp Case Study

John Bowllan[1], Kailey Cozart[2], S. M. Mahdi Seyednezhad[3], Anthony Smith[4], and Ronaldo Menezes[5]

[1] Department of Mathematics, Middlebury College,
Middlebury, Vermont, USA.
jbowllan@middlebury.edu,
[2] School of Engineering and Technology,
University of Washington, Tacoma, Washington, USA.
kncozart@uw.edu
[3] Department of Computer Engineering and Sciences,
Florida Institute of Technology, Melbourne, Florida, USA.
sseyednezhad2013@my.fit.edu
[4] Department of Computer Engineering and Sciences,
Florida Institute of Technology, Melbourne, Florida, USA.
anthonysmith@fit.edu
[5] Computer Science Department,
University of Exeter, Exeter, UK.
r.menezes@exeter.ac.uk

Abstract. Short text grouping, and topic modeling are hard tasks in text processing applications. There is not enough information in the short text to build an acceptable statistical model. In this paper, we propose a short text tagging algorithm that groups the business categories, while discriminating dissimilar ones. Multiple applications stem from this methodology, namely drawing inferences of related businesses by placing them in the appropriate group of semantically-similar business categories. This method accomplishes dimensionality reduction while capturing important underlying semantic relationship between the categories. We compare the results with the original Yelp categories and a number of random topic assignments. The results shows that our model outperforms others.

Keywords: Network science, nested stochastic block model , Yelp, topic modeling.

1 Introduction

With the rapid growth of online services and users, a considerable number of people write online comments, posts, blogs, reviews, or tags, many being short texts. Short text mining is beneficial for many applications such as advertisements, item recommendations, and user behavior characteristics. However, short texts are hard to work and classify [7] due to lack of enough information for traditional statistical models. The problem becomes even harder where people use some tags instead of a whole sentence.

In order to address these issues, we pick the 2018 Yelp data set containing short text categories, which business owners use to define their businesses. Unfortunately, the regular topic modeling algorithms such as Latent Dirichlet Allocation (LDA) [4] does not work properly on short texts. LDA turns documents to as a mixture of topics; Then it extract each word from one of its topic. In a short document with only a few of words, LDA suffers from too few observations for the parameter estimation task. Besides that, it may severely be biased towards some specific words in a set of short texts. We ran the LDA on the Yelp business categories. The top words of the topics are *restaurants, services, spas, automotive, bars, food, life, shopping, medical, home.* As can be seen the LDA model is heavily biased towards food, shopping, home, and automotive while we have 22 main groups of business categories.

We intend to extract the “topics” of business categories by creating groups of thematically-related business categories using Nested Stochastic Block Model (NSBM) [17]. NSBM demonstrates considerable advantages above other community detection algorithms [13]. The motivation for utilizing NSBM on the Yelp business categories is three-fold: 1- Yelp categories are short and not similar to regular texts. They are a set of tags. 2- Although Yelp defines 22 initial categories, we cannot rely on these categories because business owners may choose sub-categories from more than one main category. However, we compare our findings with the predefined categories in the Yelp data set, an advantage we use to evaluate our method. 3- This unsupervised network-based model can capture the characteristics of the business categories at both entity and structural levels [20, 9]. The NSBM generates a hierarchical community structure [16] from a weighted network input [16, 13] and there is no requirement to specify the number of desired communities [17]. Moreover, it is shown that networks can extract both semantics and sentiments of the entities [3, 20].

The structure of this paper is as follows. In Section 2 we provide a brief review on related work. Then, in Section 3 we explain relevant characteristics of the data set followed by Section 4 where we delineate the model construction. Then, in Section 5 we evaluate our method using two criteria: semantic similarity and normalized mutual information. Finally, we conclude this paper, and suggest potential future work.

2 Related Work

In the past, different methods have been used to determine topic representations in regular texts. Lee et al. [15] used Non-negative Matrix Factorization (NMF) to find parts-based representations of data. NMF, as well as Vector Quantization and Principal Component Analysis, were used on a database of faces provided by Bell Laboratories. Additionally, the process of applying NMF to analyze text was illustrated in detail. While this paper illustrated and explained the process of using NMF for text analysis, the process of text analysis was only explained and was not fully tested by the researchers. Moreover, non-negative matrix factorization assumes that the number of topics in a data set is known.

In their paper, Arun et al. [2] focus on finding the proper number of topics in order to improve the features used in machine learning. By using Latent Dirichlet Allocation (LDA) for matrix factorization, Arun et al. illustrate the ability of LDA to find the ideal number of topics. Both text and image data sets were used. While this paper is useful for finding the optimal number of topics in text, new problems arise when short texts are used.

With the advent of social media, short texts are abundant. However, when topic modeling is used for short texts, certain difficulties arise because of the sparsity of the data. Currently, assembling several short texts into a larger document has been the proposed solution to the problem. In a paper by Quan et al. [18], topic modeling and text aggregation were used on a data set of NIPS conference papers, as well as a data set of Yahoo! Answers. Using Short and Sparse Text Topic Modeling and Self-Aggregation (STAM), the researchers created a topic model that performed better than both Latent Dirichlet Allocation (LDA) and Biterm models.

Also recognizing the difficulties of working with topic modeling and short text, Hong et al [14]. address the issue by proposing ways to better train models that will be used for short text identification. They discuss an Author-Topic Model version of LDA, as well as 3 schemes that can be used alongside LDA to create greater accuracy on a data set of messages collected by Twitters streaming API. While this paper did not introduce new methods for short text topic modeling, it discusses better ways in which researchers can approach LDA.

Besides the above traditional methods of text mining, network sciences have been found useful for text analysis [8, 1, 21]. Furthermore, in the case of topic modeling, Gerlach et al. [11] used Nested Stochastic Block model. They create a network of words and documents, then they extracted the communities using NSBM to define the topic of the communities. The difference from their research and this paper is that we experiment very short texts containing a limited number of words – in some cases we have only 5 words. Moreover, we create the network using the co-occurrence cliques extracted from the words in each short text.

3 Data

We obtained the 2018 Yelp data set, comprised of six data sets related to businesses, users, reviews, check-ins, tips, and photos [22]. For the purposes of our studies, we fixated our efforts on the business data set. In this data set, each business has a “business categories” field containing short text tags it indicates best represent the services it offers. Wide reaching and varying in specificity, some example categories include “Comfort Food”, “Seafood”, “Venues and Event Spaces”, “Internet Service”, and “Ophthalmologists”. According to Figure 1, businesses indicate a minimum of 0 categories and a maximum of 36 categories with the vast majority of businesses providing 2 or more categories. We exclude businesses indicating 0 business categories in this analysis, as the analysis relies on the presence of business categories.

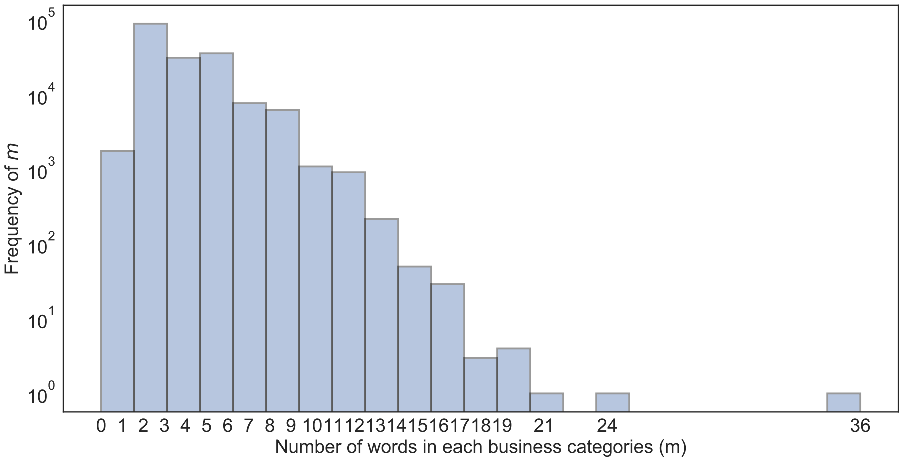


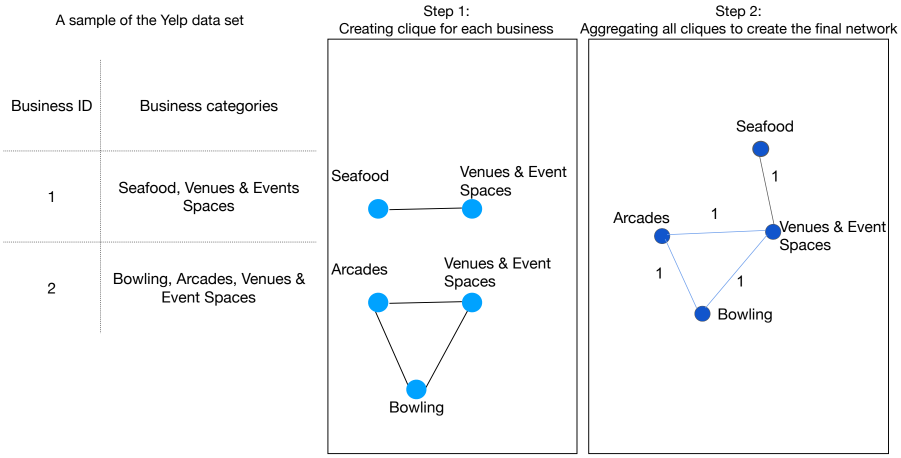
Fig. 1. Distribution of number of categories businesses use to self-identify. Some business owners may use up to 36 words to describe their business category.

4 Short Text Tagging Model

In this section, we describe the short text tagging model consisting of two main components: network representation of semantic proximity and iterative nested stochastic block model algorithm. We describe each of these components accompanied by supplemental visualizations in the following sub-sections.

4.1 Semantic Proximity using Networks

Networks provide insight into the dynamics and structure of elements, represented by nodes, and their connections, represented by edges. First, we create relationships between categories via a network-based approach. Let $m = 174,567$ denote the number of businesses and $l = 1293$ the total number of business categories. Define $C_i = \{c_1, c_2, \ldots, c_n | 1 \leq n < l\}$ where $1 \leq i \leq m$ as the set of business categories describing the i$^{\text{th}}$ business. All l categories represent the nodes in the network. We define an undirected edge between categories c_i and c_j as $e_{i,j} = (c_i, c_j)$ where $1 \leq i, j \leq l$ and $i \neq j$ if and only if $c_i, c_j \in C_k$ for any C_k. That is, we define an edge between two categories if both are contained in the same set of categories. The weight $w_{i,j}$ of an edge $e_{i,j}$ is the number of sets C_k for which $c_i, c_j \in C_k$. In practice, we create an edge list by generating all combinations of business categories for every business. The higher the weight, the stronger the connection is between business categories. Figure 2 shows the steps to create the network.



A sample of the Yelp data set

Business ID	Business categories
1	Seafood, Venues & Events Spaces
2	Bowling, Arcades, Venues & Event Spaces

Fig. 2. The steps to create the business category network. The nodes are the words in business categories defined by the business owners. The Edges are the co-occurrence of the categories in the same business category describe by the owners. The final network is weighted, and undirected.

4.2 Tagging via Iterative Nested Stochastic Block Model (NSBM)

Modularity is one metric to measure structure within networks. Highly modular networks, when divided into modules of nodes, exhibit dense inter-modular connections and sparse connections between nodes of different modules [10]. In our case, when optimizing for modularity, we intend to find business categories that are frequently mentioned together within the same module. We uncover the modular structure of the network of business categories via the nested stochastic block model, a generative model which hierarchically groups nodes into said blocks, or communities [17]. Figure 3 displays the hierarchical block structure while the Block Membership table depicts sample business category block memberships. It should be noted that in our simulations, we use 15 communities at the level 2 to assign the topics to the business categories. Figure 4 shows the number of categories at the level two in 100 different runs.

Once we determine the block membership of each node, or business category, for simplicity, we label each block by its highest degree node. Recall that the goal is to relabel each business by its predominant block membership, which

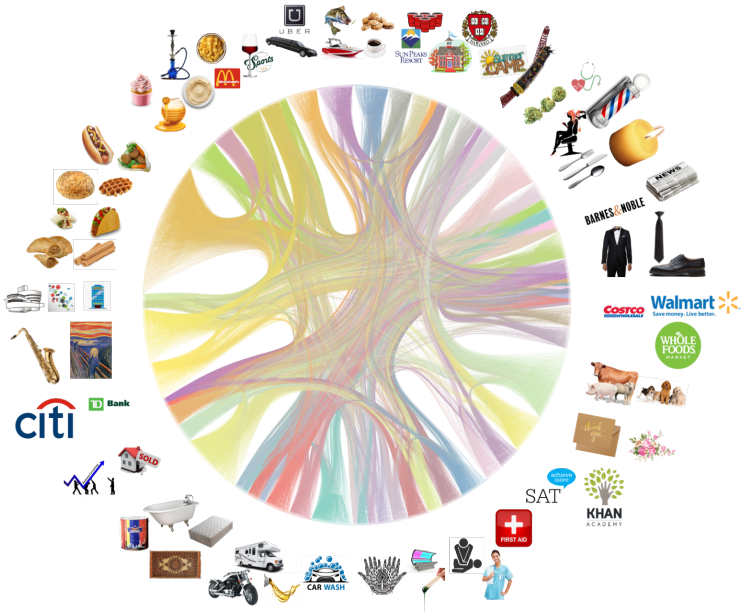


Fig. 3. NSBM diagram depicts the hierarchical structure of business categories. All l business categories line the perimeter of the circle, logos representing select categories. Blocks of semantically similar categories are distinguished by color. The interior shows the block-block relationships.

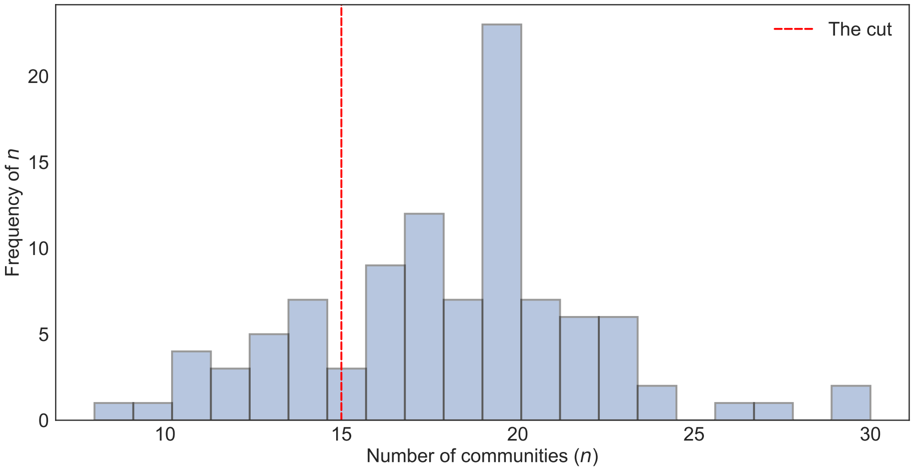


Fig. 4. This is the frequency of the number of communities for 100 runs of NSBM on the network of words. in 76% of the times, we observe more than 15 categories.

Block #	Member Business Categories
1	Bankruptcy Law, Tax Law, Trusts, Payroll Services, ...
2	Antiques, Thrift Stores, Used, Vintage and Consignment, ...
3	Jewelry Repair, Pawn Shops, Watches, Appraisal Services, ...
4	Mailbox Centers, Passport and Visa Services, ...
5	Nightlife, Restaurants, Party and Event Planning, Food, ...
6	Banks and Credit Unions, Business Financing, Financial Advising, Investing, ...
7	Computers, Electronics Repair, Mobile Phones, Computer Repair, ...
8	Laundromat, Laundry Services, Dry Cleaning, Personal Shopping, ...
9	Car Wash, Gas Station, Backshop, Auto Detailing, ...
10	Boat Dealers, Golf Cart Dealers, Motorcycle Dealers, RV Dealers, ...
11	Surfing, Sledding, Skiing, Skate Shops, ...
12	Bikes, Hunting and Fishing Supplies, Outdoor Gear, Sporting Goods, ...
13	Hotels and Travel, Resorts, Tours, Travel Services, ...
14	Cafes, Coffee and Tea, Beer, Wine and Spirits, ...
15	Seafood, Steakhouses, Buffets, Burgers, ...
16	Health Markets, Herbs and Spices, Juice Bars and Smoothies, Tea Rooms, ...
17	Cupcakes, Donuts, Shaved Ice, Custom Cakes, ...
18	Tobacco Ships, Vape Shops, Pop-up Shops, Souvenir Shops, ...
19	Landscaping, Plumbing, Hardware Stores, Contractors, ...
20	Colleges and Universities, Employment Agencies, Educational Services, ...
21	Arts and Crafts, Cards and Stationary, Florists, Flowers, Gifts, ...
22	Barbers, Hair Salons, Wigs, Cosmetics and Beauty Supply, ...
23	Books, Mags, Music and Video, Bookstores, ...
24	Day Spas, Massage, Nutritionists, Skin Care, ...
25	Urology, Ophthalmologists, Orthodontists, Dentists, ...
⋮	⋮

Table 1. In this table a number of block numbers are shown with the corresponding business categories grouped in them using NSBM.

the business categories determine. Consider the following iterative process for relabeling each business:

Consider a business with a set of categories C_i. First, we determine the block membership $\forall\ c_i \in C_i$ and note the frequency of each block present. We repeat this process for a prescribed amount of iterations, 15 in our case, keeping track of the frequencies of all blocks ever present for business i over all iterations. This allows for variations in block structure due to the probabilistic nature of this technique. After completing the iterative process, we label each business by the block with the highest frequency since the business categories predominately belong to said block.

This iterative approach synthesizes a large and diverse host of business categories into a significantly smaller set of blocks, which capture semantic relationships between the categories. Table 2 shows all of the assigned categories extracted using our method. As can be seen, “Bars” and “Restaurats”

5 Method Evaluation

In this part, we evaluate our method using two criteria:

- Semantic similarity [12] of the words inside a cluster of categories.
- Normalized Mutual Information (NMI) between the group of businesses categorized by Yelp, and our method [6].

Semantic similarity can give us valuable information regarding the similarity of words in a text [5, 12]. To obtain the semantic similarity of the method, we calculate the sum of the semantic similarities between the pairs of the words in the categories of the businesses clustered in the same community. We call it ”inter-semantic similarity” (ISS). Then we normalize the ISS of each community by dividing it by the number of unique words in the community. We also calculate ISS for 3 other cases to compare our method against them. Table 3 shows the summary of the methods and the inter-semantic similarity of them in the simulations. As can be seen, the categories extracted by the major Yelp original category (i.e. Yelp-business) has the best ISS. After normalizing other ISS’s by dividing them on it, our method shows the best results.

To have a better insight regarding the ISS of different methods, we show the distribution of ISS’s of the communities/clusters in Figure 5. It can be seen that even in the Yelp-business method we have a wide normal distribution. It means that the ISS can be very low in some cases. However, in our method we do not have a community with very low ISS.

The second criterion is the normalized mutual information (NMI) [19] between the businesses labeled with the Yelp-business method and the other methods. In this case, we also shuffle the final categories (i.e. labels) to have another random-based method. Table 4 shows the NMI between the Yelp-business categories and the categories assigned by other methods. In addition to analysis the the label of the businesses, we also perform this test for the words that are labeled in different ways. In both cases, our method shows an acceptable NMI with the Yelp-business categories.

```
Result: Words and businesses labeled by synthesized categories
$Net$ = Create undirected weighted network(Yelp business categories);
$maxIt$ = number_iterations;
labels = A dictionary of the words with their labels as the dictionary values;
/* In the label dictionary, each word is a key, each key gets a list
   of labels with the size of $maxIt$                                  */
final_word_label= dictionary of words;
final_business_label = dictionary of the business ID's;
for $i = 1$ to $maxIt$ do
    /* Extract the communities using NSBM                              */
    $COMs$ = NSBM.extract_communities($Net$);
    foreach word in $Net$ do
        /* Every word is a node in our network.                        */
        deg = Calculate the degree of the nodes;
    end
    foreach community in $COMs$ do
        community_label = The node with the highest degree in community;
        foreach word in community do
            labels.word[i] = community_label;
        end
    end
end
foreach word in $Net$ do
    final_word_label.word = The most frequent label in labels.word;
end
/* In the next loop we extract the category of each business.          */
foreach BusinessID in Yelp Business do
    business_label_candidates = Empty list;
    foreach word in business_categories do
        Add word to business_label_candidates;
    end
    final_business_label.BusinessID = The most frequent label in
      business_label_candidates;
end
```

Algorithm 1: Short-Text Tagging Pseudocode. It should be noted that in each iteration we assign a label to each node. Finally we assign the label appeared in more iterations to each word. We also extract the label of business by finding the most frequent label for each business category.

Category #	Assigned Category	Portion of data set
1	Preschools	0.0062
2	Sports Clubs	0.0121
3	Cosmetics & Beauty Supply	0.0167
4	Financial Services	0.0188
5	Oil Change Stations	0.0191
6	Pubs	0.0259
7	Used	0.0290
8	Home Cleaning	0.0305
9	Hair Removal	0.0349
10	Active Life	0.0431
11	Fashion	0.0648
12	Home & Garden	0.0957
13	Beauty & Spas	0.1304
14	Restaurants	0.2177
15	Bars	0.2392

Table 2. The final 15 extracted and assigned categories with the portion of data they cover.

Method	Extracted categories	Total ISS	Normalized
Our method	Using NSBM explained in Algorithm 1	2.65	0.62
Random category	Categories are randomly assigned to the businesses.	2.04	0.47
Yelp-business	The major categories in each business description	4.28	1.00
Yelp-raw	Just the words and theirs categories on the Yelp website.	0.53	0.12

Table 3. Inter-semantic similarity of different methods. The normalized column is the total ISS of a given method divided by the total ISS's of the Yelp-business

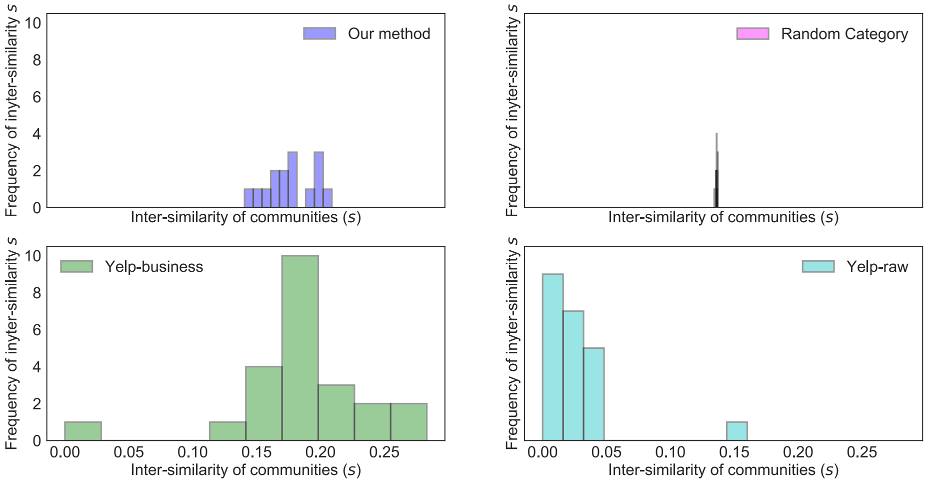


Fig. 5. The distribution of inter-semantic similarity (ISS) of the communities of the words based on the main categories extracted by different methods. Yelp-raw and Random category have the worst ISS in most of the communities.

Method	NMI - Business level	NMI - Word level
Our method	0.4773	0.3816
Our method with shuffled labels	0.0002	0.1374
Random category	0.0002	0.0782

Table 4. The NMI between three methods and the Yelp-business categories. Our method shows an acceptable amount of mutual information.

6 Conclusion and Future Work

In conclusion, we created the co-occurrence network of words in the categories of businesses for the Yelp data set. Then we ran NSBM to find the communities of the words. We labeled the communities at level 2 by the node with the higher weighted degree. Then we ran this procedure for 15 time, and assigned 15 different categories (i.e. labels) as the topic of the categories. While other methods are not applicable for this purpose, we compare our method with the original Yelp categories and some other random-based methods. The results suggested that our method assigned an acceptable set of categories to the business.

This algorithm is widely applicable to myriad feature engineering tasks. For example, consider a recommender system, where the input consists of user, item, and contextual features and the model outputs a predicted rating of the item. High-dimensional user or item text-based features, such as the "business categories" feature, can provide relevant perspective on user-item interactions, but can also increase the variance in the input space. Applying the algorithm to said features can capture the underlying relationships between collections of text while reducing the feature's dimension. Another application is the tag grouping on an online website, or social media. This short-text tagging method comprises a new approach to dimensionality reduction for high-dimensional short text features.

7 Acknowledgement

This section is currently omitted due to peer review restrictions.

References

1. Amancio, D.R., Nunes, M.d.G.V., Oliveira Jr, O., Pardo, T.A.S., Antiqueira, L., Costa, L.d.F.: Using metrics from complex networks to evaluate machine translation. Physica A: Statistical Mechanics and its Applications 390(1), 131–142 (2011)
2. Arun, R., Suresh, V., Madhavan, C.V., Murthy, M.N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 391–402. Springer (2010)
3. Biemann, C., Roos, S., Weihe, K.: Quantifying semantics using complex network analysis. In: Proceedings of COLING 2012. pp. 263–278 (2012)

4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
5. Bollegala, D., Matsuo, Y., Ishizuka, M.: A web search engine-based approach to measure semantic similarity between words. IEEE Transactions on Knowledge and Data Engineering 23(7), 977–990 (2010)
6. Byrd, R.J., Ravin, Y.: Identifying and extracting relations in text. na (1999)
7. Chen, M., Jin, X., Shen, D.: Short text classification improved by learning multi-granularity topics. In: Twenty-Second International Joint Conference on Artificial Intelligence (2011)
8. Drieger, P.: Semantic network analysis as a method for visual text analytics. Procedia-social and behavioral sciences 79, 4–17 (2013)
9. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences 106(36), 15274–15278 (2009)
10. Fortunato, S.: Community detection in graphs. Physics Reports pp. 75–174 (2010)
11. Gerlach, M., Peixoto, T.P., Altmann, E.G.: A network approach to topic models. Science Advances 4(7) (2018), https://advances.sciencemag.org/content/4/7/eaaq1360
12. Gomaa, W.H., Fahmy, A.A.: A survey of text similarity approaches. International Journal of Computer Applications 68(13), 13–18 (2013)
13. Hartman, R., Seyednezhad, S.M., Pinheiro, D., Faustino, J., Menezes, R.: Entropy in network community as an indicator of language structure in emoji usage: A twitter study across various thematic datasets. In: International Conference on Complex Networks and their Applications. pp. 328–337. Springer (2018)
14. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. pp. 80–88. acm (2010)
15. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788 (1999)
16. Peixoto, T.P.: Efficient monte carlo and greedy heuristic for the inference of stochastic block models. Physical Review E 89(1), 012804 (2014)
17. Peixoto, T.P.: Hierarchical block structures and high-resolution model selection in large networks. Physical Review X 4(1), 011047 (2014)
18. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
19. Seifzadeh, S., Farahat, A.K., Kamel, M.S., Karray, F.: Short-text clustering using statistical semantics. In: Proceedings of the 24th International Conference on World Wide Web. pp. 805–810. ACM (2015)
20. Seyednezhad, S.M.M., Fede, H., Herrera, I., Menezes, R.: Emoji-word network analysis: Sentiments and semantics. In: The Thirty-First International Flairs Conference (2018)
21. Silva, F.N., Amancio, D.R., Bardosova, M., Costa, L.d.F., Oliveira Jr, O.N.: Using network science and text analytics to produce surveys in a scientific topic. Journal of Informetrics 10(2), 487–502 (2016)
22. Yelp: Yelp open dataset (2018), https://www.yelp.com/dataset