Short text Tagging using Nested Stochastic Block Model: A 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. From online reviews and product descriptions to tweets and chats, many modern applications revolve around understanding both semantic structure and topics of short texts. Due to significant reliance on word co-occurrence, traditional topic modeling algorithms such as LDA perform poorly on sparse short texts. In this paper, we propose an unsupervised short text tagging algorithm that generates latent topics, or clusters of semantically similar words, from a corpus of short texts, and labels these short texts by stable predominant topics. The algorithm defines a weighted undirected network, namely the one mode projection of the bipartite network between words and users. Nodes represent all unique words from the corpus of short texts, edges mutual presence of pairs of words in a short text, and weights the number of short texts in which pairs of words appear. We generate the latent topics using nested stochastic block models (NSBM), dividing the network of words into communities of similar words. The algorithm is versatile—it automatically detects the appropriate number of topics. Many applications stem from the proposed algorithm, such as using the short text topic representations as the basis of a short text similarity metric. We validate the results using inter-semantic similarity and normalized mutual information, which show the method is competitive with industry short text topic modeling algorithms.

Keywords: Network science, nested stochastic block model, topic modeling, machine learning, short text tagging.

1 Introduction

With the rapid growth of online services, users often contribute to the immense corpus of short texts in the form of blogs, posts, tweets, reviews, and short tags. Understanding the topics and semantic significance of short texts is crucial in many applications ranging from item recommendation to monitoring hate speech. Traditional topic modeling algorithms such as Latent Dirichlet Allocation (LDA) fail to reveal latent topics within sparse texts, exhibit an inability to correctly choose the number of topics, and demonstrate bias towards specific words in short texts as there are too few observations for parameter estimation [4,7].

We propose a network-based unsupervised algorithm that generates topics by extracting communities of similar words, from a corpus of short texts, and assigns one topic to each short text, taking community instability into account. Given a set of short texts, we first define a weighted undirected network of all words within the corpus, where edges between words represent co-occurence within a short text. Another interpretation of the network is the one mode projection of the bipartite network between words and users who wrote the short texts. After, we extract the topics by uncovering its modular structure using the nested stochastic block model (NSBM) [17], which demonstrates considerable advantages over other community detection algorithms [13]. Each short text is then represented by either the predominant topic or a combination of topics in the form of a community distribution.

We use the Yelp 2018 dataset [23] containing business names and corresponding sets of descriptors, each forming a short text, to evaluate the algorithm. From this dataset, the algorithm clusters business descriptors into communities of thematically-related descriptors and assigns each business a unique topic. With this, we are able to discover the theme(s) of businesses via topic distributions and use this engineered feature for other predictive modeling purposes.

The motivations for the network interpretation of the corpus of words and discovering topics with the NSBM, especially using the Yelp dataset, are as follows:

1. Yelp businesses are described by personalized sets of descriptors, nouns and adjectives, defined by the registered business, which constitute well-defined pre-processed short texts.
2. This unsupervised network-based model can capture the characteristics of the business descriptors at both entity and structural levels [20,9] because networks can extract both semantics and sentiments of the entities [3,20].
3. Although Yelp defines 22 initial categories (or topics) [22], each a set of descriptors, we should not fully rely on these categories since business owners may choose descriptors from more than one main category. Our method for finding topics is purely user-driven.
4. Since the NSBM is unsupervised, there is no requirement for specifying the number of communities, or topics [17].

The structure of this paper is as follows. First, in Section 2 we briefly survey related work followed by an explanation of relevant characteristics of the Yelp business dataset used in this analysis in Section 3. We then outline the model schema together with experimental evaluation measuring cluster cohesion, semantic similarity and normalized mutual information in Sections 4 and 5. Finally, we suggest future work and further applications of the proposed algorithm.

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 dataset is known.

Arun et al. [2] focused 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 datasets were used. While this paper is useful for finding the optimal number of topics in text, new problems arise when dealing with short texts.

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 work by Quan et al. [18], topic modeling and text aggregation were used on a dataset of NIPS conference papers, as well as a dataset 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] addressed the issue by proposing ways to better train models that will be used for short text identification. They discussed an Author-Topic Model version of LDA, as well as 3 schemes that can be used alongside LDA to create greater accuracy on a dataset of messages collected by the Twitter streaming API. While this paper did not introduce new methods for short text topic modeling, it discusses better ways in which researchers can approach the problem with LDA.

Besides the above traditional methods of text mining, network sciences have been found useful for text analysis [1,8,21]. Furthermore, in the case of topic modeling, Gerlach et al. [11] used the NSBM. They created a network of words

and documents, and then extracted the communities using NSBM to define the topic of the communities. In our research, as we work with short texts containing a very limited number of words, we do follow an NSBM approach, but construct the network using word co-occurrence cliques in short text and label each short text by a stable topic, taking into account community instability.

3 Dataset Description

The 2018 Yelp dataset contains six datasets related to businesses, users, reviews, check-ins, tips, and photos [23]. We used the business dataset where each business has a *categories* field, a set of descriptors in the form of short text tags that represents the services offered by the business. The dataset contains 174,567 businesses, 1,293 unique descriptors. Some descriptors include *comfort food*, *seafood*, *venues and event spaces*, *Internet service*, and *ophthalmologists*. According to Figure 1, businesses indicate a minimum of 0 descriptors and a maximum of 36 descriptors with the vast majority of businesses providing 2 or more descriptors. We exclude businesses providing 0 descriptors, as the analysis relies on short text presence. Thus, each business is assigned to one short text, namely its set of business descriptors.

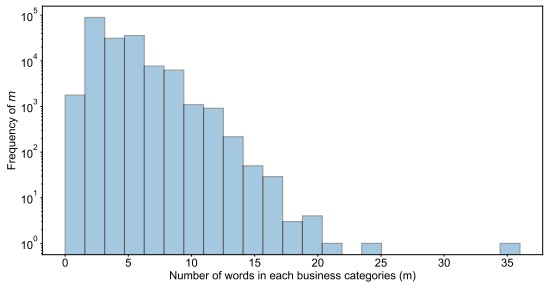

Fig. 1. Distribution of number of descriptors. Some business owners may use up to 36 words to describe their business.

4 Short Text Tagging Model

For the Yelp case study, the proposed short text tagging model defines a mapping from the set of descriptors to a smaller set of latent topics, which allows for creating a topic representation of each business. Using these topic representations, we may establish some business similarity metric. To achieve this, we represent descriptor relationships in the form of a weighted undirected network and extract its modular structure in an iterative process using NSBM. In this section, we outline the model's components, accompanied by supplemental visualizations.

4.1 Modeling Descriptor Similarity

Networks provide insight into the dynamics and structure of elements, represented by nodes, and their connections, represented by edges. In our study, the network is the one mode projection of the bipartite network between descriptors and businesses. In other words, the set of all unique descriptors represent the nodes. We define an undirected edge between two nodes if the corresponding descriptors are both contained in same set of descriptors for a particular business. The edge weight is defined as the number of businesses where the two corresponding descriptors satisfy the aforementioned edge condition. In practice, we create a weighted edge list by generating all combinations of descriptors for every business, group by the edge, and aggregate by count. Figure 2 shows a visual representation of the network creation. Higher edge weight corresponds to businesses more frequently using a pair of descriptors together to characterize their services. Thus, if a subset of descriptors are frequently used together to describe more businesses, there exists a user-defined thematic relationship the descriptors in the form of a latent topic.

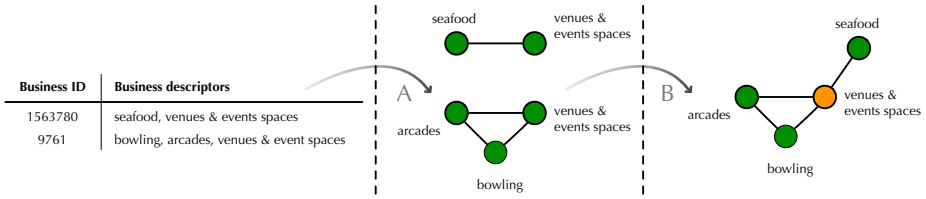

Fig. 2. (A) From a list of businesses and their respective sets of descriptors, we can generate a series of cliques representing pairs of descriptors used to characterize a business. (B) We then create a network where the cliques are combined based on common nodes. Note that in this figure we highlight the descriptor *venues & events spaces* to show that the cliques were combined and that node was common to both cliques. Weights can also be present if we have pairs of descriptors that are common to multiple businesses. The final network is weighted, and undirected.

4.2 Tagging via Iterative Nested Stochastic Block Model

Modularity is one metric to measure structure within networks. Highly-modular networks, when divided into communities of nodes, exhibit dense inter-modular connections and sparse connections between nodes of different communities [10]. The NSBM is a generative model extracting hierarchical modular structure in networks by grouping nodes into communities from a weighted network input [17,16,13]. In our case, when optimizing for modularity, we represent abstract topics in the form of communities using the NSBM. Figure 3 displays the hierarchical block structure while the Table 1 shows sample community (topics).

```
Result: Descriptors and businesses labeled by synthesized categories
Net = Build an undirected weighted network (Yelp business descriptors);
maxIt = number_iterations;
labels = A dictionary of the descriptors with their labels as the dictionary
 values;
/* In the label dictionary each descriptor is a key, each key gets a
   list of labels with the size of maxIt                              */
final_word_label= dictionary of descriptors;
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 descriptor in Net do
        /* Every descriptor is a node in our network.                 */
        deg = Calculate the degree of the node;
    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 descriptor 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 pseudo-code. It should be noted that in each iteration, we assign a label to each descriptor. This label is the hub of the community in which the descriptor is contained. Then, we assign to each descriptor the most frequent label. After, we extract the label of businesses by finding the most frequent label for each business descriptor.

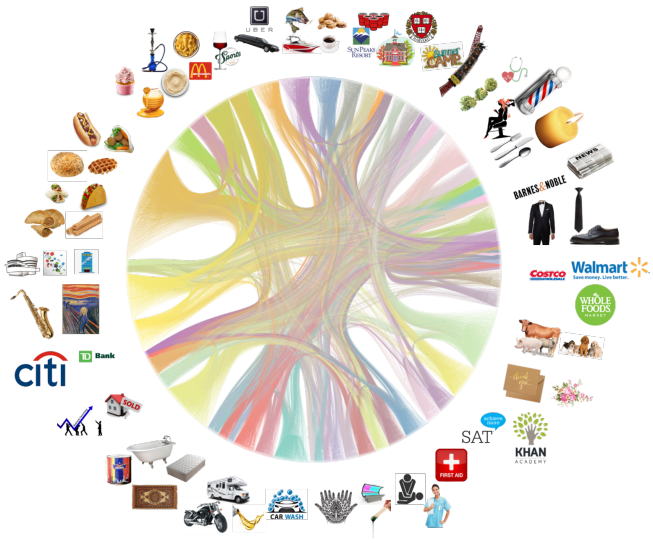

Fig. 3. NSBM diagram depicts the hierarchical structure of descriptors, which line the perimeter of the circle, logos representing select descriptors. Communities of semantically similar descriptors are distinguished by color. The interior shows the community-community relationships. It should be noted that the icons around the network is for visualization purposes. Each icon is a famous business related to the category of the community.

We determine the topic representation of the short texts by an iterative process using NSBM. For n iterations, we fit a NSBM to the weighted undirected network input to create a mapping from the nodes to the extracted communities. For each short text, we map each descriptor to its corresponding community and track the counts of unique communities over all n iterations. This allows for variations in community structures since the NSBM is a generative probabilistic model. After n iterations, each short text exhibits a topic representation, that is, a combination of communities at varying counts, which we normalize to create a topic distribution. This approach synthesizes a diverse host of business descriptors into a significantly smaller set of communities capturing descriptor semantic relationships and user-defined business themes. In a nutshell, to label descriptors in each iteration, we use the hub of the community as the label of the descriptors in that community. After all iterations, we assign the descriptors to their most frequent label, then we assign the businesses to the most frequent label of their descriptors.

5 Method Evaluation

We evaluate our proposed method using two separate criteria:

1. Semantic similarity [12] of descriptors within communities.

Table 1. The business descriptors are grouped in blocks after the application of NSBM. Here we see a sample of some of blocks containing semantically similar descriptors.

Block #	List of categories in the block
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, ...
⋮	⋮

Table 2. Test table.

index	name	categories	**final_label**
1	Dental by Design	[Dentists, General Dentistry, ...]	$\mathbf{c_{i_1}}$
2	Vans	[Shopping, Men's Clothing, Shoe Stores, Fashion, ...]	$\mathbf{c_{i_2}}$
3	Royal Fades Barbershop	[Barbers, Hair Salons, ...]	$\mathbf{c_{i_3}}$
4	Quik Chik	[Restaurants, Chicken Wings]	$\mathbf{c_{i_4}}$
⋮	⋮	⋮	⋮

2. Normalized Mutual Information (NMI) between Yelp's categorization of descriptors and those determined by our method [6,24].

Semantic similarity provides valuable information regarding the similarity of words in a text [5,12]. To obtain the semantic similarity of descriptors within communities, we first calculate the sum of the semantic similarities between all pairs of descriptors contained in a community, which we call "inter-semantic similarity" (ISS). Then we adjust the total ISS computed by each method by dividing by the total ISS computed by Yelp's categorization technique, which we deem the "ground truth". We must note a detail regarding this analysis: Recall that the NSBM hierarchically partitions nodes into communities, which are themselves grouped into higher level communities and so on until we reach the top of the tree. In this evaluation, we used "level 2" communities to assign descriptors to topics instead of "level 1" communities. More specifically, a "level 2" community contains communities, each containing descriptors. Although we lose ISS since higher level communities contain more words, the number of "level 2" communities closely resembles the number of groups (22) Yelp used for their categorization method [22], integral to our evaluation. Figure 4 shows the number of "level" communities over 100 iterations of the NSBM. For future study, we intend to expand our evaluation metrics using lower level communities.

We use 3 other method to assign categories to words: Random category, Yelp-business, Yelp-raw For the Yelp-business method, mentioned in this table, we assign the main category by voting over the descriptors the owners wrote. As a result, for each business we assign the most frequent category. For example, we may have *bar, food, spa* for a business. If on the Yelp website *bar* and *food* are from the *Restaurant*, and *spa* is from *Pleasure*, then the main categories based

Table 3. As per our community-naming convention, in addition to integer representations, we label a community by the descriptor in the community with the highest degree. The % in dataset column shows the predominance of that descriptor in the dataset.

Community	Assigned Descriptor	% in dataset
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

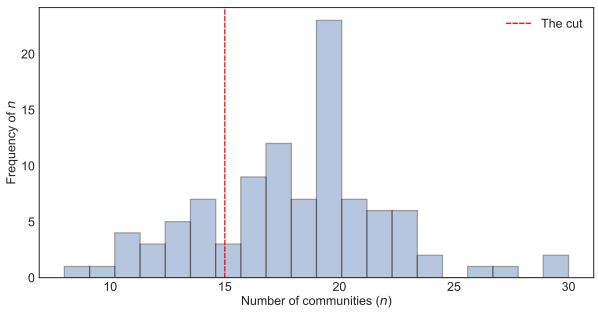

Fig. 4. This is the frequency of the number of communities for 100 runs of NSBM on the network of descriptors. 76% of the time, we observe more than 15 communities which is close to the number of categories of descriptors Yelp uses.

on Yelp is *Restaurant, Restaurant, Pleasure*. Then, we assign *Restaurant* to this business as the main category. On the other hand, the yelp-raw has nothing to do with the data set of businesses. It is solely the words set and assigned by Yelp found on the Yelp category web page[6]. This is the one that defines *food* is a subcategory of *Restaurant*.

We compare our method's ISS calculation against those computed by 3 other methods mentioned in Table 4. Although it shows that Yelp's categorization of descriptors yields the best ISS, if we establish Yelp's categorization of descriptors as the "ground truth", or the true total ISS, then when we divide the total ISS of the other three methods by this ground truth ISS, our method outperforms the other methods. Figure 5 shows the ISS distributions for the different methods. Note that Yelp's categorization method has a wide normal distribution, meaning the ISS is very low in some cases. However, in our method, we do not have a community with very low ISS.

Table 4. Inter-semantic similarity of select methods. The normalized column is the total ISS of a given method divided by the total ISS of the Yelp-business.

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

The second evaluation criterion is normalized mutual information (NMI) [19] between businesses labeled by synthesized categories for different methods. For this evaluation, we labeled each business with its highest represented community in its topic representation distribution. Table 3 shows all of the assigned categories extracted using our method. We also shuffle the final business labels to have another random-based method. Table 5 shows the NMI between Yelp-defined categories and the labels assigned by other methods. We also perform this test for descriptors that are labeled in different ways. In both cases, our method shows an acceptable NMI with Yelp-defined categories.

6 Conclusion and Future Work

We created the co-occurrence network of descriptors from the Yelp business dataset. With this network as input, through an iterative process, we fit nested hierarchical modular structure on the network using NSBM, resulting in a topic (or community) distribution for each business. We compare our method against the 22 category aggregations Yelp defines and some other random-based methods. The results suggested competitive inter-semantic similarity using this algorithm. However, for broader cases, the words do not have to be semantically

[6] https://blog.yelp.com/2018/01/yelp_category_list

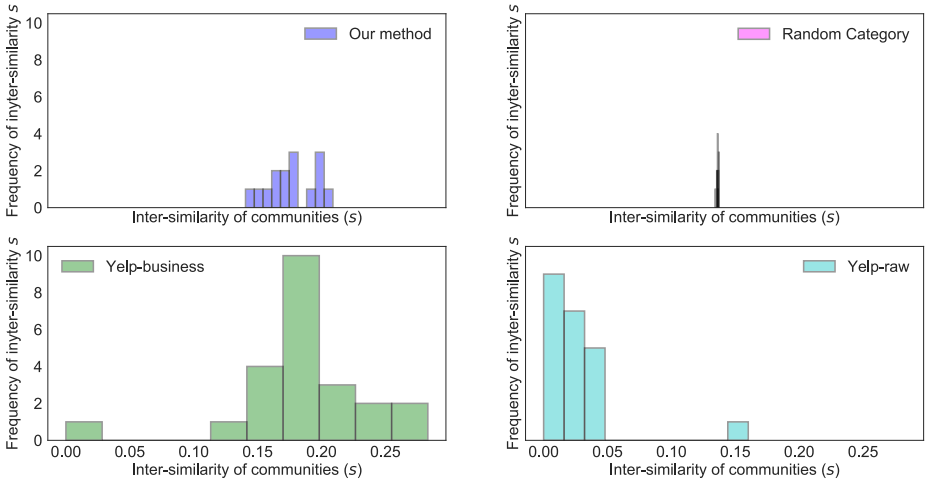

Fig. 5. The distribution of inter-semantic similarity (ISS) of the communities of the words based on the main categories extracted by different methods. Yelp-raw and Random category have the worst ISS in most of the communities.

Table 5. The NMI between three methods and the Yelp-business categories. At the business level, we compare the categories assigned to businesses, and at the word level we compare the categories assigned to the words. Our method shows an acceptable amount of mutual information.

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

similar. This algorithm is completely user-driven, uncovering the topics users define, not business entities that create them. For further study, we intend to explore other community detection algorithms for short text tagging and other evaluation metrics to analyze the algorithm's performance.

This algorithm is also widely applicable to a myriad of 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 "business categories", can provide relevant perspective about user-item interactions, but due to the raw structure, can also increase the variance in the input space. Applying the algorithm to such features can capture the underlying relationships between collections of text while reducing the feature's dimension. Other applications stem from this algorithm, namely computing similarity between short text reviews by computing the similarity between their respective topic distributions generated by this algorithm.

7 Acknowledgement

The authors would like to thank the NSF for funding the AMALTHEA REU and Florida Institute of Technology for hosting the program. The authors would also like to acknowledge support from the NSF grant No. 1560345.

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: The complete yelp category list (2018), https://blog.yelp.com/2018/01/yelp_category_list
23. Yelp: Yelp open dataset (2018), https://www.yelp.com/dataset
24. Zhang, P.: Evaluating accuracy of community detection using the relative normalized mutual information. Journal of Statistical Mechanics: Theory and Experiment 2015(11), P11006 (2015)