Topic Modeling Using Community Detection on a Word Association Graph

Mahfuzur Rahman Chowdhury
Brac University
Dhaka, Bangladesh
mahfuzur.rahman@bracu.ac.bd

Intesur Ahmed
Brac University
Dhaka, Bangladesh
intesur.ahmed@bracu.ac.bd

Farig Sadeque
Brac University
Dhaka, Bangladesh
farig.sadeque@bracu.ac.bd

Muhammad Nur Yanhaona
Brac University
Dhaka, Bangladesh
nur.yanhaona@bracu.ac.bd

Abstract

Topic modeling of a text corpus is one of the most well-studied areas of information retrieval and knowledge discovery. Despite several decades of research in the area that begets an array of modeling tools, some common problems still obstruct automated topic modeling from matching users' expectations. In particular, existing topic modeling solutions suffer when the distribution of words among the underlying topics is uneven or the topics are overlapped. Furthermore, many solutions ask the user to provide a topic count estimate as input, which limits their usefulness in modeling a corpus where such information is unavailable. We propose a new topic modeling approach that overcomes these shortcomings by formulating the topic modeling problem as a community detection problem in a word association graph/network that we generate from the text corpus. Experimental evaluation using multiple data sets of three different types of text corpora shows that our approach is superior to prominent topic modeling alternatives in most cases. This paper describes our approach and discusses the experimental findings.

1 Introduction

The goal of topic modeling is to find the underlying semantic structure in a corpus that succinctly describes the documents and the text forming the corpus without compromising the corpus's statistical characteristics. It is one of the oldest and most researched problems in the field of information retrieval and has numerous direct and downstream applications such as document grouping, classification, retrieval, and summarization. For a long time, the most prominent solutions to topic modeling are Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its variants. LDA works under the principle of 'interchangeability' of documents and describes a document as a random mixture of a fixed number of latent 'topics' drawn from a Dirichlet distribution where each topic is a distribution of words. LDA is a bag of words model as it disregards the position of words in a document.

LDA's assumption of a latent Dirichlet distribution for the topics and interchangeability of documents under the bag of words model is both for the mathematical tractability of the problem, as its authors admit (Blei et al., 2003), as opposed to its conformance with any empirical rule describing real texts (e.g., the Zipf's law (Zipf, 1935)). However, LDA provides the first principled approach to group both the documents and the words of a corpus and remains widely applicable. Consequently, most subsequent works on topic modeling focused on improving the LDA model which led to many LDA variants. For example, Biterm (Yan et al., 2013) adapts LDA for short texts, HDP (Teh et al., 2004) eliminates LDA's requirement of an input topic count, and SeededLDA (Jagarlamudi et al., 2012) incorporates human input in LDA training.

Most recent neural network solutions for topic modeling also focus on tackling specific shortcomings of LDA as opposed to proposing a better alternative mathematical foundation. For example, ProdLDA (Srivastava and Sutton, 2017a) addresses the concern of difficulty of Gibb's sampling and variational inference for LDA training by transferring the model parameters to the neural space, ETM (Dieng et al., 2020) addresses LDA's limitations in dealing with sparse and large vocabularies by using word embeddings, and CTM (Bianchi et al., 2021) generalizes LDA for cross-lingual topic modeling. All these solutions improve LDA in different respects but some fundamental limitations of LDA persist and hurt its effectiveness in many scenarios.

A frequently cited problem with LDA and its variants is the large discrepancy between their

Proceedings of Recent Advances in Natural Language Processing, pages 908–917
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_098

output and human judgment (Jagarlamudi et al., 2012). Various attempts are being made to overcome this problem by making LDA-based topic modeling semi-supervised, e.g., in ITM (Hu et al., 2014), GuidedLDA (Jagarlamudi et al., 2012), and SSHLDA (Mao et al., 2012). All these solutions apply some human-provided constraints on the LDA model training and only attain partial success as the document collections may not fit a latent Dirichlet distribution in the first place (Gerlach et al., 2018). Then given LDA tries to fit a probabilistic generative model against a corpus using maximum likelihood estimation or some other statistical measure, it tends to overlook small topics (Gerlach et al., 2018) and struggles when topics are overlapped (Jagarlamudi et al., 2012).

In this paper, we present an alternative to LDA-based topic modeling to address the above problems using a document-structure-sensitive topic modeling through community detection (Barabási, 2013) in a word co-occurrence graph. We call our solution **ComTM**[1]. In ComTM, we use the structure of the documents in the text corpus to generate the co-occurrence graph with words that capture the core information flow of the documents, as opposed to all words. Then we apply a novel overlapping community detection algorithm to extract the topics. ComTM is applicable when the type/source of the documents in the collection is uniform and known. This assumption is practical as topic modeling is frequently applied to a corpus of a specific type of document such as only news articles, scientific publications, or Wikipedia articles. Meanwhile, the assumption is necessary to apply a single strategy to capture the information flow of the documents. When the documents are of manifold type, it is unrealistic to assume the user knows their information flow structures.

ComTM is not the first attempt to apply community detection to the topic modeling problem. Community detection is a widely studied branch of network science that discovers meaningful clusters/communities in a graph by analyzing its wiring diagram (Barabási, 2013). It shows significant success in describing graphs originating from biological, physics, and human networks and begets several popular algorithms. The particular appeal of community detection is that it does not require any cluster/community count as input which was a major obstacle for the application of traditional graph partitioning algorithms (Buluç et al., 2016) in real-world networks. Another important characteristic of community detection algorithms is that they are non-parametric and can detect communities when their composition in the network is an uneven mixture of small and large communities.

These attractive features led researchers to apply community detection for topic modeling of text corpora. For example, community detection has been applied to guide LDA topic modeling using network-structured metadata such as citation information (Bouveyron et al., 2018a) (Hyland et al., 2021). Some recent works completely replace LDA with community-detection-based topic modeling using a different statistical criterion for community fitness calculation (Gerlach et al., 2018), called minimum description length (MDL) (Peixoto, 2013), that originates from information theory. However, to the best of our knowledge, no existing solutions apply overlapping community detection for which existing algorithms have exponential or inordinately high-degree polynomial running times, otherwise producing poor results. Secondly, none of them incorporates any notion of differential importance of words of the documents when constructing the word co-occurrence graphs. ComTM's ingenuity lies in these two aspects.

We compared ComTM with several LDA variants and a prominent community-detection-based topic model, called hSBM (Gerlach et al., 2018), on several data sets. The data sets are constructed or collected from online news articles from several outlets, scientific papers, and Wikipedia articles. We evaluated the topic models' output using both human annotators and cluster coherence measurement. The result shows that ComTM consistently outperforms other solutions for most data sets and for overlapping data sets in particular. This paper discusses our experimental findings along with the design methodology and algorithms for ComTM.

To summarize, the contributions of this paper are as follows:

1. Present the first document-structure sensitive topic modeling solution that uses community detection for topic identification.

2. Propose a novel algorithm for overlapping community detection on a network where nodes represent texts.

3. Discuss experimental findings from comparing the new topic modeling solution with

[1] https://github.com/ThreeSwordAI/ComTM

prominent alternatives.

4. Share the source code and instruction manual for the new topic modeling solution.

2 Design Methodology

In designing ComTM, we apply a notion/framework of document 'interchangeability' quite different from LDA. In our framework, the places a word occurs in a document are significant. Words in a document attain importance by virtue of their inclusion in larger semantic structures, such as sentences or paragraphs, that carry the core information of the document. In other words, we interpret words occurring in more important sentences/paragraphs as more important than other words. Under this interpretation, a pair of documents are more or less interchangeable depending on the similarity of their text contents in parts that carry the central information the documents try to convey. Consequently, the collection ComTM can process must contain documents that are structurally similar and whose structure reflects the relative importance of different semantic blocks within the documents.

Evidently, ComTM is not a generic topic model yet as its applicability relies on a preprocessing step that can explain the structure of an arbitrary document in terms of the relative importance of different parts. The current scope of ComTM covers three document classes: 'hard' news articles, scientific papers, and Wikipedia articles.

Hard news articles are those that report on recent incidents of local and global importance (Lehman-Wilzig and Seletzky, 2010) (thus, opinion pieces, interviews, and long-read articles are not hard news). They constitute the majority of daily news publications worldwide (Liebler and Smith, 1997; Tuchman, 1972; Patterson, 2000). Historically hard news articles follow an *inverted pyramid model* to capture the short attention span of typical news readers. In this model, the leading paragraph summarizes the key points of the event that subsequent paragraphs elaborate on in decreasing order of importance. Numerous studies from media sociology validate Adherence to this structure in English news articles (Po¨ttker, 2003; Smith, 1978).

Next we use scientific articles as they include an abstract section that abridges the contribution of the paper, which allow us to treat the abstract as a container of some of the most important words in this document category. Finally, we target Wikipedia articles as generic multi-section descriptive prose category and apply and concatenate extractive summaries of those sections to form the most semantically significant document part. We trust on extractive summaries for Wikipedia articles because the SOTA tools for summarization are frequently trained on Wikipedia data and extractive summaries are generated following a theory of information contribution of sentences to the meaning of their containing document.

We then construct a word co-occurrence (aka, association) graph using the words of the central information-bearing part of the documents. In this graph, the nodes are words and there is an edge between two nodes if the corresponding words appear in the same document – not necessarily in its central part – anywhere in the corpus. The graph is weighted, where the weight of an edge reflects in how many documents the corresponding pair of words co-occurred.

Then we apply a non-overlapping community detection on the word-occurrence graph. The goal here is to partition the graph into clusters. However, unlike the other community-detection-based topic models that consider the discovered communities to be topics and the most frequent words in the communities as the top topic words (Bouveyron et al., 2018b) (Gerlach et al., 2018), we apply eigenvector centrality measure to filter the most important words from identified communities. Eigenvector centrality encapsulates other notions of graph centrality such as between-ness, degree, and closeness centrality (Bonacich, 2007) and considers edge weights, which community detection ignores. There is a philosophical reason for choosing this alternative significance measure also that the eigenvector centrality measure reflects better:

> There is no reason to assume that the item which recurs most frequently is the most important ... the place occupied by the different elements is more important than the number of times the recur.
>
> *Oliver Burgelin (McQuail, 1972)*

In essence, we apply a standard community detection algorithm on the word occurrence graph to get the topic count and identify the central words of the individual topics. Subsequently, we construct a larger word co-occurrence graph by considering all words in the corpus and applying our own algorithm to associate other words with the central topic words based on a graph proximity calculation. At that time a single word can be associated with

multiple topics. This two-step process can be described as a new overlapping community detection algorithm for a weighted word co-occurrence graph that returns topics as word distributions.

3 Algorithm & Implementation

Although there is evidence that community detection algorithms can handle word co-occurrence graphs formed from unfiltered text corpora, we removed stop word and lemmatized in ComTM on the ground of pragmatism. In addition, we only considered nouns and verbs in the initial graph that ComTM uses for the topic count and central topic word identification. In that regard, ComTM uses the NLTK (Bird et al., 2009) package for parts of speech tagging. In addition, we apply WordNet (Fellbaum, 1998) super-subordinate relation among the filtered words and replace them with their immediate hypernyms. This has been done to capture the clustering tendency among the words in the co-occurrence graph at a higher conceptual level and also to make the graph more compact when the corpus is large.

For hard news articles and scientific papers, the word set each document contributes to the co-occurrence graph comes from the leading paragraph and the abstract section respectively. For Wikipedia articles, ComTM uses the Bert based extractive summarizer (Miller, 2019; Sabharwal et al., 2021) for each section then concatenates the summary of the sections. We found that the summaries incorporate most keywords of the documents. Still, we added the output of the KeyBERT keywords extractor (Grootendorst, 2020) in the word set of the combined summary in ComTM's initial word co-occurrence graph process. Finally, ComTM drops words from the sets that occurred in only a single document before the graph construction as they cannot influence the community structure of the corpus but increase the memory and processing footprint of the community detection algorithm.

3.1 Topic Count & Central Topic Words Determination

ComTM applies the Louvain community detection algorithm (Blondel et al., 2008) from the NetworkX package (Hagberg et al., 2008) in the word co-occurrence graph. Currently, Louvain is the fastest non-overlapping community detection algorithm for unweighted graphs with running time $\mathcal{O}(L)$ for an input graph having L edges. All community detection algorithms only consider the wiring structure of the input graph, consequently edge weights are ignored.

Louvain algorithm partitions a graph into communities based on the notion of 'modularity,' which says the participants in a community should be more interconnected to each other than nodes from other communities. Mathematically, the objective of the algorithm is to maximize the following equation:

$$M = \sum_{c=1}^{n_c} \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 \quad (1)$$

Here n_c is a community, L_c is the number of links/edges inside the community, and k_c is the number of links from n_c to other communities. Modularity maximization has the limitation that communities smaller than $\sqrt{2L}$ get merged into larger communities. So it is often advised to run the algorithm recursively in partitioned sub-graphs representing large communities (Barabási, 2013). However, ComTM only runs the Louvain algorithm once.

After communities are identified, ComTM recreates weighted, induced (West, 2000), sub-graphs for the communities before eigenvector centrality computation then keeps the topmost ten words from each community as the central topic words. Finally, ComTM shares topic words among the communities if a word of a community has a weighted degree centrality score higher than the last member of another community if being added to that community's induced sub-graph. The equation for the weighted degree centrality score for a word w in a community C with vertex set V and edge set E is as follows:

$$s_w = \sum_{u \in V, \exists (u,w) \in E} \iota_u \times weight(u, w) \quad (2)$$

Here ι_u is the eigenvector centrality score of word u in the subgraph representing community C.

3.2 Topic Identification

Once community count and the central words of each community are found, ComTM creates a new weighted word co-occurrence graph with all words in the corpus (except stop words). Now there is an edge between two graph nodes if they occur anywhere in the same document and the weight

of the edge is the number of documents they co-occur. Then weights are normalized and ComTM runs the following custom overlapping community detection algorithm with the graph and central topic words as input.

Algorithm 1: Topic Assignment Algorithm

Input: g - a weighted graph
τ_c - a multiset of central words
ϵ - a threshold parameter
Output: τ - topic assignments of all words

$M \leftarrow \emptyset$
$L \leftarrow |\tau_c|$
foreach $v \in g \ \ \& \ \ v \notin \tau_c$ **do**
  $\vec{T_v} \leftarrow 0_L$
  $M \leftarrow M \bigcup \{v : \vec{T_v}\}$
foreach $i \in [0, L)$ **do**
  $s_i \leftarrow \tau_c[i]$
  $r \leftarrow 0$
  while $\exists v \in g \ \ \& \ \ v \notin \{\tau_c, s_i\}$ **do**
    $r \leftarrow r + 1$
    foreach $(u, v) \in g \ \ \& \ \ u \in s_i \ \ \&$ $v \notin \{s_i, \tau_c\}$ **do**
      $T_v \leftarrow M[v]$
      $T_v[i] = T_v[i] + \frac{weight(u,v)}{2^r}$
    foreach $v \in g \ \ \& \ \ v \notin s_i \ \ \&$ $\exists u, (u, v) \in g \ \ \& \ \ u \in s_i$ **do**
      $s_i \leftarrow s_i \bigcup \{v\}$
$\tau \leftarrow \tau_c$
foreach $v \in M.keys$ **do**
  $T_v \leftarrow M[v]$
  $N_v \leftarrow normalizeVector(T_v)$
  $i \leftarrow maxIndex(N_v)$
  $\tau[i] \leftarrow \tau[i] \bigcup \{v\}$
  $s \leftarrow N_v[i]$
  foreach $j \in [0, L)$ **do**
    if $N_v[i] - N_v[j] \leq \epsilon$ **then**
      $\tau[j] \leftarrow \tau[j] \bigcup \{v\}$

Algorithm 1 is basically a gradient descent algorithm that assigns a per-topic significance weight to each word w in the corpus based on w's proximity to the central topic words. The significance weight drops exponentially with the w's distance from the set of central topic words. Then w gets assigned to the topic that it is closest to. Then based on a cutoff threshold parameter ϵ it is also shared with other topics. ComTM uses the output of this final algorithm as the topics for the corpus.

4 Experiments

Since the qualitative value of found topics under human judgment is ComTM's main target, statistical measures such as perplexity score or maximal likelihood commonly used for evaluating topic models (Blei et al., 2003; Gruber et al., 2007) are of little use. Earlier research shows that these measures do not typically correlate with human judgment (Chang et al., 2009). Therefore, you employed five human annotators to judge the topic outputs of ComTM and reference baseline implementations. We estimated the IAA(Inter-Annotator Agreement) of the annotators using Fleiss' kappa (Fleiss, 1971) to assess the quality of the annotations. The score– 0.4275 indicates that the human judgments were highly similar among the five annotators.

We compared ComTM with LDA (Blei et al., 2003), CTM (Bianchi et al., 2021), ETM (Dieng et al., 2020), HDP (Teh et al., 2004), ProdLDA (Srivastava and Sutton, 2017b), hSBM (Amini et al., 2023) and Seeded-LDA (Jagarlamudi et al., 2012). For reference implementations of these existing topic models, we use Gensim (Rehurek and Sojka, 2011) and Octis (Terragni et al., 2021) libraries. We used the graph-tool library (Peixoto, 2014) for visualizing the comparison results.

However, we applied two techniques to avoid making our evaluation completely subjective. First, we compared cluster coherence scores (Mimno et al., 2011) of different topic model outputs in each data set. Some empirical studies show cluster coherence scores for frequent words correspond well with human judgment. Second, we use curated datasets with known categories of documents (e.g., sports, business, and politics can be different categories of a hard news dataset) for various experiments to assess the topics' relevance to those categories.

4.1 Datasets

We used a total of eight datasets for the three classes of documents ComTM currently supports. For each class, the datasets are of different compositions. Descriptions of these datasets are given in Table 1.

There are four datasets for hard news articles. Among these, we created two and collected the remaining two from publicly available sources. We categorized the datasets into overlapping or non-overlapping based on their characteristics type. A dataset correlated to geopolitics, (e.g., the Ukraine-

Dataset	Type	Total Data	Categories	Distribution	Topic Types
1 (Custom)	News Articles	1480	3	Balanced	Overlapping
2 (Custom)	News Articles	2525	5	Imbalanced	Overlapping
3 (Gültekin, 2020)	News Articles	2225	5	Balanced	Non-overlapping
4 (Gültekin, 2020)	News Articles	775	5	Imbalanced	Non-overlapping
5 (Bonhart, 2020)	Scientific Abstracts	2229	8	Balanced	Non-overlapping
6 (Densil, 2020)	Scientific Abstracts	2500	6	Balanced	Overlapping
7 (Foundation)	Wiki-Data	204	4	Balanced	Non-overlapping
8 (Foundation)	Wiki-Data	153	3	Balanced	Overlapping

Table 1: Datasets Characteristics

Russia War, the Sri Lanka crisis, and the China-Taiwan conflict) is an example of an overlapping dataset. Meanwhile, a dataset containing business, sports, entertainment, tech, and political news is a non-overlapping dataset. Our news article datasets have either an even distribution of different categories of news or are intentionally uneven. In the former case, we call the dataset balanced, and in the latter case imbalanced.

We took the dataset of PubMed Abstracts from (Bonhart, 2020) in our experiments with abstract data. It covers 8 topics (Deep Learning, Human Connectome, Covid-19, Virtual reality, and Brain-Machine Interfaces (Electroactive Polymers, PEDOT electrodes, and Neuroprosthetics)). We also experimented with the abstracts of the Research Articles dataset (Densil, 2020), which comprises six areas (Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, and Quantitative Finance), in addition to PubMed. These datasets only contain abstracts – not the whole papers – so we had to restrict ComTM's topic identification phase (Section 3.2) to abstracts only.

Finally, to experiment with Wikipedia articles, we filtered six distinct category-based Wikipedia articles with both overlapping (capital cities, mythological places, and countries) and non-overlapping (sports, movies, universities, and countries) categories from the Wikipedia dump (Foundation) available at Hugging Face.

4.2 Evaluation

We evaluate ComTM against other topic models in three stages.

4.2.1 Stage 1: Comparision with SOTA

In the first stage, we applied LDA, CTM, ETM, HDP, ProdLDA, and hSBM on all datasets to compare ComTM's performance with the existing state-of-the-art. Before applying the models we lemmatized every word and removed stop words and the words with term frequency-inverse document frequency (TF-IDF) scores above 0.8 from the dataset. As LDA, CTM, ETM, and ProdLDA need the topic count as input, we applied topic coherence (Mimno et al., 2011) scores to calculate the count. To find the optimal number of topics with coherence; for each experiment, we ran the three models with the topic count of 1 to 40 and calculated the coherence value for each case. We kept the best result that gave the highest coherence score. HDP and hSBM do not require any topic count input. Consequently, we ran them on each dataset only once with their default configuration.

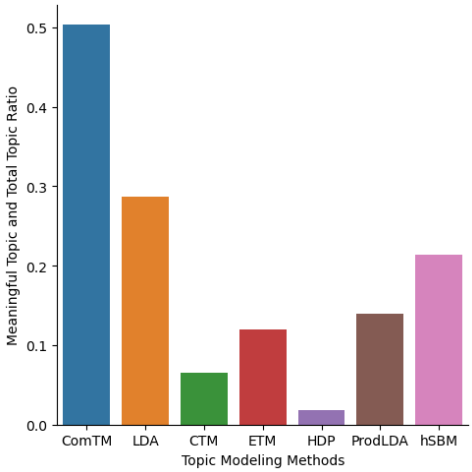

Figure 1: Meaningful Topic Count vs Total Topic Count Ratio.

The topic coherence score distribution of all topic models in the Stage 1 experiments we can identify that the coherence scores of ETM are significantly higher than any other algorithms and the scores of LDA are also much better compared to other models.

However, both ETM and LDA scored best in coherence scores for an unusually large number of topics. Therefore, the qualitative significance of their output is under question. So, we then asked the human annotators to evaluate the topic outputs of all models without telling them which topic

Dataset	Topic Number							Best Performing Algorithm	
	ComTM	LDA	CTM	ETM	HDP	ProdLDA	hSBM	Unknown Topics	Known Topics
1	10	26	37	26	150	10	35	ComTM	ComTM
2	10	28	37	26	150	35	23	ComTM	ComTM
3	9	2	31	34	150	10	23	ComTM	ComTM
4	11	1	33	5	150	10	20	ComTM	ComTM
5	5	1	33	22	150	9	35	ComTM	ComTM
6	3	3	28	26	150	10	9	ComTM	ComTM
7	5	26	8	26	150	14	2	ComTM	ComTM
8	8	22	13	26	150	12	5	ComTM	ComTM

Table 2: Topic Counts with Best Performing Topic Models Under Human Judgement

model produced what output. The evaluation has two parts. In the first part, the annotators rated every topic based on their eloquence and ranked them based on their meaningfulness and diversity. In the second part, annotators were informed about the categories of each dataset. Then they had to match the categories with the topics and rate the models based on their matching and coverage. Table 2 shows the detailed results of annotator evaluation.

Table 2 shows that in both parts of the evaluation, ComTM performs universally the best in all experiments. We then focused on determining why ComTM ranked best in the experiments.

Table 3 shows how many topics among all topics identified by the models in various experiments are judged meaningful (that is, relatable to any category included in a dataset) by the annotators.

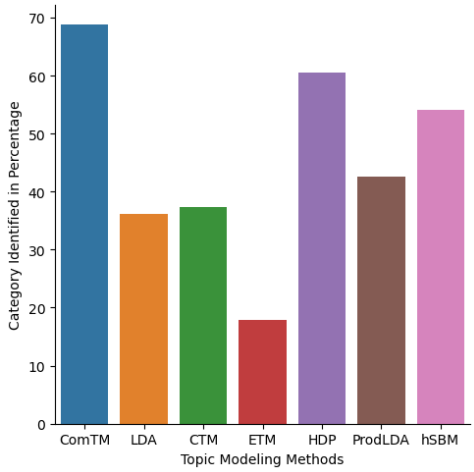

Figure 2: Average Document Category Coverage of Different Topic Models in Stage 1.

We can see that though the annotators declare ComTM better than others, sometimes hSBM provided more meaningful topics than ComTM. However, when we measure the average fraction of meaningful topics among spurious topics and duplicates then ComTM scores much higher than hSBM or any other models as shown in Figure 1.

As the final evaluation of Stage 1 experiments, we computed how many categories got covered by the topics generated by each model. As we know the category decomposition of the documents in our datasets, there must be at least one topic related to each document category for a topic model, dataset pair. According to the annotators, as shown in Figure 2, LDA and ETM miss many categories altogether despite having high coherence scores.

To summarize the Stage 1 evaluations, the higher number of meaningful topics, lower percentage of spurious topics, and better dataset coverage provide ComTM a competitive edge against competitor topic models.

Dataset	Topic Identified						
	ComTM	LDA	CTM	ETM	HDP	ProdLDA	hSBM
1	4	2	2	0	3	1	6
2	4	2	3	1	3	2	4
3	6	1	3	0	3	2	6
4	6	1	1	1	2	3	5
5	5	1	2	0	3	3	5
6	2	1	1	1	1	2	1
7	4	3	2	2	1	2	0
8	4	2	1	1	1	2	2

Table 3: Dataset-wise Total Topics Relatable to Document Categories.

4.2.2 Stage 2: Community Count Normalized Comparision with SOTA

Our human evaluations suggest that higher coherence scores are not good indicators of the actual topic count. Therefore, to give topic models that require a count input a boost, we used the number of topics identified by ComTM as the input in LDA, CTM, ETM, and ProdLDA. We then asked the annotators to rate the new outputs and also compared the coherence scores of the models under this new setting. The summary results and the category coverage percentage are shown in Table 4 and Figure 3. Comparing Figure 2 and Figure 3 we can observe a sharp decline in category coverage for LDA, CTM, ETM and ProdLDA.

Dataset	Topic Input	Topic Identified					Best Performer
		ComTM	LDA	CTM	ETM	ProdLDA	
1	10	4	2	0	1	1	ComTM
2	10	4	3	0	2	2	ComTM
3	9	6	4	1	3	3	ComTM
4	11	6	4	1	1	2	ComTM
5	5	5	3	2	1	3	ComTM
6	3	2	1	0	0	1	ComTM
7	5	4	2	1	0	2	ComTM
8	8	4	2	0	0	2	ComTM

Table 4: Best Performing Topic Model with Community Count as the Topic Count Input.

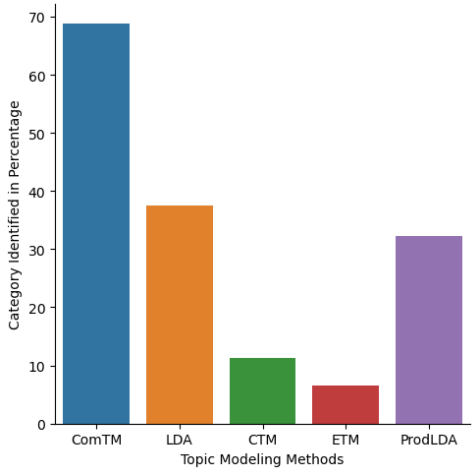

Figure 3: Average Document Category Coverage of Different Topic Models in Stage 2.

As the annotators already know the document category composition of individual datasets from Stage 1 experiments, here they rate each topic model output only once. We again observed that coherence scores remain high for ETM and LDA even in this setting. However, in terms of meaningfulness, category coverage, and avoidance of spurious topics; ComTM remains the best performer. This result also suggests that one cannot just use a community detection algorithm as a topic count input to significantly improve the performance of LDA-like topic models.

4.2.3 Stage 3: Comparision with Seed Boosted LDA-like Model

In the final stage of our experiments, we gave LDA-like topic models a further boost by providing the central words identified by ComTM in its topic counter and central topic words determination phase (Section 3.1) as initial seeds for LDA training. We used SeededLDA for this experiment as it accepts seeds to guide LDA training. The purpose of this stage is to investigate whether community detection output can guide existing topic models so much so that a full community detection-based topic model may be unnecessary.

We used the topic count and the top five words per topic from ComTM as the seeds for Seeded-LDA. As ComTM uses TF-IDF in the leading paragraph/abstract/summary and Seeded-LDA uses it in whole documents, sometimes ComTM produces seeds that are not in the word list of Seeded-LDA. We remove that word from the seed list to tackle this problem. Another problem can occur if the community size is smaller than the seed size. In that case, we removed the community from the community list.

Datasets	Number of Topics	Number of Seeds	Best Performer
1	10	5	ComTM
2	10	5	ComTM
3	9	5	ComTM
4	11	5	ComTM
5	5	5	ComTM
6	3	5	Seeded-LDA
7	5	5	Seeded-LDA
8	8	5	ComTM

Table 5: Performance Comparison with Seeded-LDA.

Table 5 shows the result of annotator evaluation for the top ten words per topic for both ComTM and Seeded-LDA. ComTM performs better six out of eight times than Seeded-LDA. Still, there are two experiments where the annotators rated Seeded-LDA better. Those two experiments show the prospect for a future hybrid topic model as we guided Seeded-LDA with ComTM.

5 Conclusion

In this research, we proposed ComTM, a topic model based on community detection. ComTM is document structure sensitive and does not require a preset topic count as input. Our experiments on multiple datasets of hard news articles, scientific abstracts and wiki data show that ComTM is generally superior to the dominant topic modeling alternatives in this particular domain. Given that ComTM utilizes the structure of documents to identify core words in each document as a pre-processing step in its modeling, alternative mechanisms to core words/terms identification augmented with ComTM has prospect for improvement in topic modeling in other domains as well. We encourage other researchers to investigate this prospect.

Acknowledgement

This work was funded by Brac University Research Seed Grant Initiative 2022.

References

Arash A. Amini, Marina S. Paez, and Lizhen Lin. 2023. Hierarchical stochastic block model for community detection in multiplex networks.

Albert-László Barabási. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *J*, 2008(10):P10008.

Phillip Bonacich. 2007. Some unique properties of eigenvector centrality. *Social Networks*, 29(4):555–564.

Bonhart. 2020. Pubmed abstracts.

C. Bouveyron, P. Latouche, and R. Zreik. 2018a. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 28(1):11–31.

C. Bouveyron, P. Latouche, and R. Zreik. 2018b. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 28(1):11–31.

Aydın Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. 2016. *Recent Advances in Graph Partitioning*, pages 117–158. Springer International Publishing, Cham.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Blesson Densil. 2020. Topic modeling for research articles.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Wikimedia Foundation. Wikimedia downloads.

Martin Gerlach, Tiago P Peixoto, and Eduardo G Altmann. 2018. A network approach to topic models. *Science advances*, 4(7):eaaq1360.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 163–170, San Juan, Puerto Rico. PMLR.

Habib Gültekin. 2020. Bbc news archive.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Charles C. Hyland, Yuanming Tao, Lamiae Azizi, Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. 2021. Multilayer networks for text analysis with multiple data types. *EPJ Data Science*, 10(1):33.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213.

Sam N. Lehman-Wilzig and Michal Seletzky. 2010. Hard news, soft news, 'general' news: The necessity and utility of an intermediate classification. *Journalism*, 11(1):37–56.

Carol M Liebler and Susan J Smith. 1997. Tracking gender differences: A comparative analysis of network correspondents and their sources. *Journal of Broadcasting & Electronic Media*, 41(1):58–68.

Xian-Ling Mao, Zhao-Yan Ming, Tat-Seng Chua, Si Li, Hongfei Yan, and Xiaoming Li. 2012. SSHLDA: A semi-supervised hierarchical topic model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 800–809, Jeju Island, Korea. Association for Computational Linguistics.

D. McQuail. 1972. *Sociology of Mass Communications: Selected Readings*. Penguin Books. Penguin.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Thomas E Patterson. 2000. Doing well and doing good. *Available at SSRN 257395*.

Tiago P. Peixoto. 2013. Parsimonious module inference in large networks. *Phys. Rev. Lett.*, 110:148701.

Tiago P. Peixoto. 2014. The graph-tool python library. *figshare*.

Horst Po¨ttker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modeling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Navin Sabharwal, Amit Agrawal, Navin Sabharwal, and Amit Agrawal. 2021. Bert model applications: Other tasks. *Hands-on Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing*, pages 139–171.

Edward J. Smith. 1978. Screw model has advantages over inverted pyramid. *The Journalism Educator*, 33(4):17–19.

Akash Srivastava and Charles Sutton. 2017a. Autoencoding variational inference for topic models.

Akash Srivastava and Charles Sutton. 2017b. Autoencoding variational inference for topic models.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Gaye Tuchman. 1972. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of sociology*, 77(4):660–679.

Douglas B. West. 2000. *Introduction to Graph Theory*, 2 edition. Prentice Hall.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

George Kingsley Zipf. 1935. *The psycho-biology of language: an introduction to dynamic philology*. The psycho-biology of language: an introduction to dynamic philology. Houghton Mifflin, Oxford, England.