AN INTEGRATED TEXT GRAPH REPRESENTATION LEARNING FOR SEMANTIC AND SYNTACTIC ENHANCED SHORT TEXT STREAM CLUSTERING

Tham Vo
Thu Dau Mot University, Binh Duong, Vietnam

Abstract: Text stream clustering is considered as a primitive task in natural language processing (NLP) which contains unique challenges related to the sparsity/noise, infinite length and cluster evolution of the input documents. In recent years, many mixture topic model, such as: Dirichlet Process Mixture Model (DPMM) based algorithms (e.g., MStream, OSDM, etc.) have demonstrated remarkable improvements in the accuracy performance of short text stream clustering task. However, these contemporary DPMM-based models still suffered limitations related to the capability of sufficiently capturing the sequential and long-range syntactic dependent relationships between words in texts in which can assist to leverage the quality of extracted clusters from given streams. To deal with these challenges, in this paper, we proposed a novel integrated graph convolutional network (GCN) with DPMM for handling text stream clustering task, called as GOWGCNStream. Our proposed GOWGCNStream model is an integration of GCN with BERT for capturing the joint syntactic structural and contextual representations of texts which are then used to facilitate the DPMM framework for dealing with short text stream clustering task. Extensive experiments in benchmark datasets (Tweet-Set and Google-News) demonstrated the effectiveness of our proposed GOWGCNStream model in comparing with recent state-of-the-art baselines.

Keywords: BERT; DPMM; GCN; graph-of-words; text stream clustering.

I. INTRODUCTION

In recent years, with the tremendous raises of Internet and social media platforms (e.g., Facebook, Twitter, TikTok, etc.), text stream clustering task has been widely attended by many researchers due to its potential applications in various domains, like as social event detection/tracking, social trend analysis, news/content recommendation, etc. In fact, there are a huge amount of short textual data in form of chats, comments, micro-blogs, etc. have been incrementally generated from large-scale social networks every day. These massive textual data are considered as valuable knowledge resources for multiple useful applications. Different from the traditional text stream clustering task which mainly focus on analyzing and extracting cluster information from static text corpora, the text stream clustering task has its own characteristic unique challenges regarding with the sparsity in textual data source, length-varied textual input and cluster/topic evaluation (aka topic/concept challenge). In the past few years, many researchers have attempted with different approaches to deal with major challenges of text clustering task, such as efforts in dynamic topic modelling (e.g., DTM [1], TOT [2], DMM [3], etc.) and vector-space similarity based (e.g., CluStream [4], OSKM [5], DenStream [6], Sumblr [7], etc.) approaches, however these previous techniques still have been unable to deal with topic/cluster drift related challenge in which the number of topics/clusters in a given text stream are rapidly changed over the time. Recently, with the tremendous progresses in the utilization of Bayesian non-parametric model and DPMM, there are number of notable works in this area have demonstrated significant enhancements in both accuracy and time-efficiency performances for the text stream clustering task. Among mixture model based approaches, there are some popular works, such as: GSDMM [8], DPMFP [9] and DCT [10] which are designed upon the well-known topic modelling paradigm in which clusters are learnt and extracted from given streams as the document-levelled latent topic generative process. Even though these attempts have presented significant results for text stream clustering task, they still have been considered as unable to handle the evolution of topics/clusters in the context of dynamism since the number of topics/clusters is static and must be pre-defined at the initial setup stage. Thus, the topic/concept drift challenge in text stream analysis and mining is still a big open issue at that time. In fact, for practical implementations of clustering algorithms in real-world text stream sources like as social networks, there is a huge number of short textual documents such as comments, posts, micro-blogs, etc. which are generated continuously and might belong to various topics, thus the proposed text clustering models must be capable to quickly characterize

Contact author: Tham Vo
Email: Thamvth@tdmu.edu.vn
Manuscript received: 30/6/2021, revised: 24/9/2021, accepted: 22/10/2021.

distinctive textual features from incoming texts in order to efficiently identify these topic/cluster evolutions. Moreover, unlike traditional static text clustering task, the text stream clustering also has challenges related to the document's length and data sparsity. In fact, most of generated contents from common text streams like as social networks are varied in length and most of them are considered as short textual documents in which each document only contains a single sentence or few words. Thus, proposed text clustering models might easily encounter problems regarding with the sparsity/noise in the textual feature engineering and representation learning processes.

A. Recent progresses, existing challenges & motivations

Recently, there are some well-known works (e.g., MStream [11], DP-BMM [12], etc.) which have demonstrated significant improvements in dealing with the topic/concept drift challenge in text stream clustering task. Especially in the MStream [11], Jianhua Yin et al. proposed a novel update/remove mechanisms to control the number of extracted clusters from a given text stream in which the outdated clusters are efficiently removed to eliminate the topic/cluster explosion during the mining process. In general, both MStream [11] and DP-BMM [12] mainly rely on the bag-of-word distribution analysis over text corpus within the DPMM framework to efficiently learn and extract the cluster information. In fact, these models still relied on the independently evaluation of word's distributions over latent topics/clusters within different document's batches of a given text stream, thus they might suffer limitations related to the capability of capturing the rich sequential and contextual relationships between words in the given text corpus. Moreover, the utilization of separated word distribution evaluation over short texts also easily lead to the poor performances in textual representation process in which can lead to the downgrades in the quality of extracted cluster information. To overcome limitations related to the word's co-occurrence distribution in texts, recently, there are some notable efforts (e.g., Si-DPMM [13], NPMM [14], etc.) related to the integration of pre-trained word embedding approach [15] [16] [17] with the Bayesian non-parametric and DPMM-based frameworks to facilitate and improve the performance of short text clustering task. By integrating with rich contextual word representations, these proposed models have successfully in learning and capturing salient cluster information from short text corpora as well as dealing with topic/concept drift challenge in text stream clustering task. Recently, there is another effort of Kumar, J. et al. in the proposed OSDM [18] is a DPMM-based model which utilize the joint evaluation of independent word and relational co-occurring relationship distributions in the text stream to extract high-quality clusters. Moreover, the OSDM [18] also applied the cluster update/remove mechanisms which are proposed in previous work [11] to efficiently manage number of extracted clusters in the given stream. However, recent works still encountered limitations regarding with the ability of capturing the rich syntactic and long-range dependent relationships between words in texts, thus can reach the better understanding and performance in the joint contextual and structural latent feature representations in which can support to significantly leverage the performance of short text stream clustering task.

In recent studies related to text analysis and mining domain, there are several efforts [19] [20] [21] have presented the usefulness of utilizing graph-based neural network (GNN) architectures, (e.g.,: GraphSage [22], graph convolutional network (GCN) [23] and graph attention network (GAT) [24], etc.) to learn and capture the syntactical long-range dependencies between words in texts to assist multiple textual data-driven learning problems. By taking the advantages of textual graph-based structural representation learning of GNN-based architectures over different constructed text graphs, these models have reached remarkable improvements in primitive tasks of NLP such as short text clustering and classification. Moreover, for the rich contextual text representation learning, with the tremendous raise of multiple advanced deep learning based architecture in NLP area, such as: sequence-to-sequence (seq2seq) [25] aka auto-encoding (AE), attention mechanism [26] [27] and transformer [28] [29] [30], the textual representation learning task have been shifted to a higher level and perform remarkable improvements in both accuracy and time-efficiency aspects. Among recent advanced textual representation model, BERT [30] is considered as the most powerful pre-trained language model which enables to learn rich-contextual information for words and sentences from the given text corpus. These recent achievements have played as the crucial motivations for most of enhancements in multiple downstream tasks of NLP, including short text stream clustering.

B. Our contributions

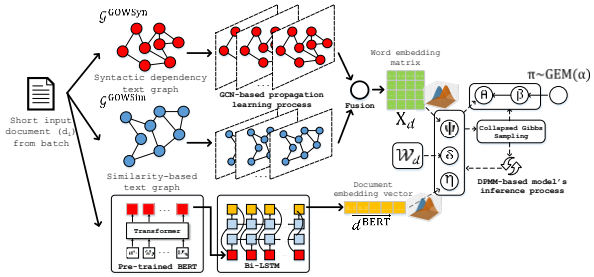

Figure 1. The illustration of the overall architecture of our proposed GOWGCNStream

To tackle above listed challenges, in this paper we proposed a novel semantic and syntactic enhanced short text stream clustering model, called GOWGCNStream (as illustrated in Figure 1). Our proposed GOWGCNStream model is an integration of syntactical and rich-contextual text graph representation learning approach with DPMM framework to efficiently deal with existing challenges in text stream clustering regarding with the sparsity/low-quality in textual representation and topic/concept drift. The utilization of text graph construction in our approach is majorly inherited from the graph-of-words (GOW) approach in which support to transform textual documents into graph-based structures. First of all, we propose a combination of using pre-trained BERT model to capture the rich-contextual representations of words and sentences in each input document. Then, these achieved BERT-based textual embedding vectors of words and sentences are used

to assist the structural representation learning of a given document by using the GCN-based architecture over different constructed text graphs. Finally, we utilize the obtained rich contextual and structural representations to facilitate the cluster inference process under the DPMM framework in which can support to effectively leverage the performance of short text stream clustering task.

Model's elaboration for short text handling. As most of DPMM-based approaches are considered as BOW-based approach in which the occurrence frequency for occurred words in documents is mainly focused. Thus, they might perform poor performance upon short/very short text corpora for dealing with clustering problem. Therefore, in this study, we utilized the integrated pre-trained BERT language model with text graph representation learning through GCN architecture to deal with the sparse feature extraction challenges for clustering problem within short text corpora.

In general, our contributions in this paper can be summarized as three-folds, which are:

- First of all, to efficiently learn the representations of words and sentences in each input document from stream, we utilize the pre-trained BERT model to extract rich contextual embedding vectors of words and sentences. Then, these word and sentence embedding vectors are used to constructed different text graphs, including similarity-based GOW and syntactical GOW. To extract syntactical dependent relationships between words in each document, we mainly reused the Stanford CoreNLP library [31]. Then from the structured text graphs, we apply the multi-layered GCN-based architecture to learn the structural representations of the input document via the graph-based propagation learning process.
- Next, the representations of documents and their contained words are utilized to facilitate the text clustering process under the DPMM framework with the participation of different textual representations beside the traditional evaluation on the word and document distributions over the model's inference process for exploiting cluster information which is majorly inherited in previous works [13] [14]. In fact, our proposed GOWGCNStream model is implemented upon DPMM framework, thus it can efficiently deal with the challenges related to the topic/cluster evolution in text stream.
- Finally, we presented extensive experiments in benchmark datasets to demonstrate the effectiveness of our proposed GOWGCNStream model in comparing with recent state-of-the-art baselines. Comparative experimental outputs show the usefulness of our ideas in this paper as well as shed some lights for further improvements in this direction.

In general, the left contents of our paper are organized into four sections. In the second section, we briefly present literature reviews about recent achievements in short text stream clustering and discuss about pros/cons of each model. In the third section, we provide detailed descriptions on the methodology and implementation of our proposed GOWGCNStream model. Next, we conduct extensive experiments and discuss about experimental results in the fourth section. Finally, we conclude our achievements in this paper and provide some potential directions for the future works.

II. RELATED WORKS

Among primitive tasks in NLP domain, text stream clustering is considered as an important task in which recently has been attracted by many researchers due to its wide applications in multiple disciplines. Different from the traditional static text clustering task, the text stream clustering task has its own challenges which are mostly related to the sparsity/noise of length-varied input documents and topic/cluster evolution in stream problem. In general, recent proposed models for text stream clustering task can be categorized into three main trends, which are: traditional topic modelling, vector space similarity and mixture model based approaches. The traditional topic model is considered as the earliest approach for text clustering in both static and dynamic contexts which are majorly inherited from the well-known latent topic inference mechanism of Latent Dirichlet Allocation (LDA) [32], proposed by Blei, D. et al. At that time, there are several topic modelling based techniques, such as: DTM (dynamic topic models) [1], TOT (topics over time) [2] and DMM (dynamic mixture models) [3] have been proposed to deal with the dynamism of clustering task in the given text stream corpora. Most of these topic modelling based techniques relied on the latent topic inference process to extract the topic-word and document-topic distributions over the given text corpora to handle the clustering task. However, these topic modelling based techniques still encountered several limitations related to the low-quality in textual feature representation which leads to the downgrades in the extracted topic/cluster information due to the noise and sparsity in the input textual data. Moreover, these recent approaches are also considered as unable to deal with the topic/cluster evolution in text streams due to the number of topic/cluster must be initially and statically identified.

On the other trend, there are several efforts of researchers who tried to applying the rich textual vector space representation and similarity-based techniques to handle the text stream clustering task. Such as notable contemporary proposed algorithms, like as: OSKM [5], DenStream [6] and Sumblr [7]. These vector space based models have utilized different document representation learning technique to learn the and measure the distance between document's as embedding vectors to efficiently identified their potential topics/clusters. By utilizing rich semantic textual representation approach, the vector space-based techniques can efficiently cope with the sparsity problem in textual representation learning process for improving the performance of clustering task. However, they have still been considered as unable to deal with the topic/concept drift challenge in text streams, because the number of topics/clusters also must be pre-defined at the initial stage like as the traditional topic modelling approach. Thus, the topic/concept drift challenge in text streams still had been considered a biggest open issue for years.

In recent time, with the emergence of mixture-model and Dirichlet Process (DP) based frameworks, many text

stream clustering models have been applied these novel paradigms to facilitate the mode's inference process for dealing with topic/concept drift challenge. In general, mixture-model like as DPMM [33] [34] utilized the well-known Bayesian non-parametric paradigm to effectively extract latent topics/clusters from texts through the inference learning process. In other words, DPMM has been coming as the main stream for most of recent proposed models in text stream clustering tasks due to the numerous advantages in comparing with previous approaches. Recent notable works like as: GSDMM [8], DPMFP [9] and DCT [10] are developed under the DP-based topic/cluster inference mechanism have demonstrated remarkable performances in the text stream clustering task within the context of dynamism in which the number of topics/clusters aren't required at the initial step. However, with the rapid increase in the number of topics/clusters from text streams, these previous DP-based model still suffered challenges regarding with the explosion in the quantity of extracted clusters from streams. To deal with this issue, recently Jianhua Yin et al. proposed a novel approach of DPMM-based model, called as: MStream [11] which is equipped with the cluster update/remove mechanism to control new/outdated clusters from streams that has successfully tackle challenges related to explosion of extracted clusters. Inspired from achievements in MStream [11], there are several contemporary proposed techniques, like as: DP-BMM [12] and OSDM [18] have significantly improved the performance of short text stream clustering task with enhancements in the extensive co-occurrence word distribution evaluation in texts. Or the proposals of Si-DPMM [13] and NPMM [14] models in the integration of DPMM with pre-trained word embedding approach to enrich quality of the textual representations to deal with sparsity/noise problem in short text analysis. Although these recent works have gained state-of-the-art performances in text stream clustering, there are several existing limitations related to the capability of capturing rich contextual and syntactic representations of texts in stream clustering which can support to leverage the quality of extracted clusters in the after all. Majorly inspired from notable gains of previous works and motivated by remained challenges, in this paper we propose the utilization of enhanced contextual and syntactical structural text representation approach with DPMM-based framework to improve the performance of text stream clustering task.

III. GOWGCNSTREAM MODEL

In this section, we formally present the methodology and implementation of our proposed GOWGCNStream model. First of all, we propose a novel joint rich contextual and syntactical structural representation learning technique which is an integration between pre-trained BERT and GCN-based architecture. Then, the rich semantic and syntactical representations of input documents are used to facilitate the DPMM-based framework to assist the text stream clustering task.

A. Textual representation learning via BERT and GCN

1) Word and sentence representation learning with BERT

In order to efficiently and effectively capture the rich contextual information from input documents, we applied the pre-trained BERT model to achieve the representations of all (n) words, dented as: $\mathcal{W}_d = \{w_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{n}}$ and (m) sentences, denoted as: $\mathcal{S}_d = \{s_{\mathrm{i}}\}_{\mathrm{i}=1}^{\mathrm{m}}$ in each input given document, as: (d). In general, the BERT is considered as the sentence-level textual encoding approach in which the word embedding vectors in each sentence of a given document is produced as the hidden states of the last k^{th} layer of a given transformer-based architecture, for a specific sentence (s), the set of rich contextual embedding vectors of occurred (z) words, are achieved as: $\left\{\overrightarrow{w_1^{\mathrm{BERT}}}, \overrightarrow{w_2^{\mathrm{BERT}}}, \dots, \overrightarrow{w_{\mathrm{z}}^{\mathrm{BERT}}}\right\} = \mathrm{BERT}^{[\mathrm{k}]}(\{w_1, w_2, \dots, w_{\mathrm{z}}\})$, with: $s = \{w_1, w_2, \dots, w_{\mathrm{z}}\}$ and $w_{\mathrm{i}}^{\mathrm{BERT}}$ presents the word embedding vector of the (i^{th}) word in given sentence (s). Then, to achieve the embedding vectors of sentences in a given document (d) by utilizing the BERT model, denoted as: $\overrightarrow{s^{\mathrm{BERT}}}$, we applied a Bi-LSTM sequential encoder to aggregate the continuous latent representations of all words in each sentence as the output last (k^{th}) hidden states in both directions of the given Bi-LSTM architecture. Similar to that, we also applied another Bi-LSTM based architecture to produce the document-level representation via BERT, denoted as: $\overrightarrow{d^{\mathrm{BERT}}}$ by aggregating sentence-level embedding vectors which have been achieved in the previous steps. The overall steps of this process can be formulated as the following (as shown in equation 1a & 1b):

$$\begin{aligned} \mathcal{H}_s^{[\mathrm{k}],+} &= \mathrm{LSTM}^{\mathrm{sen}}\left(\left\{\overrightarrow{w_1^{\mathrm{BERT}}}, \overrightarrow{w_2^{\mathrm{BERT}}}, \dots, \overrightarrow{w_{\mathrm{z}}^{\mathrm{BERT}}}\right\}\right) \\ \mathcal{H}_s^{[\mathrm{k}],-} &= \mathrm{LSTM}^{\mathrm{sen}}\left(\left\{\overrightarrow{w_1^{\mathrm{BERT}}}, \overrightarrow{w_2^{\mathrm{BERT}}}, \dots, \overrightarrow{w_{\mathrm{z}}^{\mathrm{BERT}}}\right\}\right) \\ \overrightarrow{s^{\mathrm{BERT}}} &= \left[\mathcal{H}_s^{[\mathrm{k}],+}, \mathcal{H}_s^{[\mathrm{k}],-}\right] \end{aligned} \quad (1a)$$

$$\begin{aligned} \mathcal{H}_d^{[\mathrm{k}],+} &= \mathrm{LSTM}^{\mathrm{doc}}\left(\left\{\overrightarrow{s_1^{\mathrm{BERT}}}, \overrightarrow{s_2^{\mathrm{BERT}}}, \dots, \overrightarrow{s_{\mathrm{m}}^{\mathrm{BERT}}}\right\}\right) \\ \mathcal{H}_d^{[\mathrm{k}],-} &= \mathrm{LSTM}^{\mathrm{doc}}\left(\left\{\overrightarrow{s_1^{\mathrm{BERT}}}, \overrightarrow{s_2^{\mathrm{BERT}}}, \dots, \overrightarrow{s_{\mathrm{m}}^{\mathrm{BERT}}}\right\}\right) \\ \overrightarrow{d^{\mathrm{BERT}}} &= \left[\mathcal{H}_d^{[\mathrm{k}],+}, \mathcal{H}_d^{[\mathrm{k}],-}\right] \end{aligned} \quad (1b)$$

In this equation, the $\mathcal{H}^{[\mathrm{k}],(+,-)}$ and $[.,.]$ presents for the last (k^{th}) hidden states of the given Bi-LSTM-based architecture in both (+/-) directions and the concatenation operation. In our approach, we use separated Bi-LSTM architectures for encoding the sentences ($\mathrm{LSTM}^{\mathrm{sen}}$) and documents ($\mathrm{LSTM}^{\mathrm{doc}}$) within the given text corpus. At the end of this process, we can achieve the rich contextual representations of input documents as well as their contained sentences and words in forms of fixed-dimensional embedding vectors.

2) Text graph construction and structural representation learning via GCN

Multi-typed text graph construction. In this step, we present the utilization of graph-of-words (GOW) based textual representation paradigm to preserve the similarity-based and syntactic dependent relationships between words in each input documents. To do this, for each input document (d), we constructed two types of text graph, called as: similarity-based GOW and syntactical GOW, in forms of graph-based structures, denoted as: $\mathcal{G}^{\mathrm{GOWSim}} =$

$\{\mathcal{V}^{w}, \mathcal{E}^{\mathrm{GOWSim}}\}$ and $\mathcal{G}^{\mathrm{GOWSyn}} = \{\mathcal{V}^{w}, \mathcal{E}^{\mathrm{GOWSyn}}\}$, in which the $\mathcal{V}^{w}$ is the vocabulary set or a set of unique words which are occurred in the given document (d). For the similarity-based GOW, its set of edges, as: $\mathcal{E}^{\mathrm{GOWSim}}$, present for the 1-hop co-occurring relationships between two continuous pairwise (a^{th}) and (b^{th}) words in the given document (d), the edge's weight is calculated as the cosine similarity between the BERT-based embedding vectors of the corresponding (a^{th}) and (b^{th}) words which have been achieved previously, denoted as: $\mathrm{cosine}(w_a, w_b) = \frac{\overrightarrow{w_a^{\mathrm{BERT}}}.\overrightarrow{w_b^{\mathrm{BERT}}}}{\left\|\overrightarrow{w_a^{\mathrm{BERT}}}\right\|.\left\|\overrightarrow{w_b^{\mathrm{BERT}}}\right\|}$. For the graph-based syntactical representation of document (d), the set of edges in ($\mathcal{E}^{\mathrm{GOWSyn}}$) present for the syntactical dependent relationships between words which are achieved by utilizing the available dependency parser of Stanford CoreNLP library [31]. Thus, each edge between each pairwise (a^{th}) and (b^{th}) words presents the grammatical syntactic relations (e.g., "det", "nsubj", etc.) between them. At the end of this process, we can achieve two types of GOW-based representations for each input document which are later fed to the GCN-based architecture to effectively learn and capture the structural representation of a given document (d).

Structural representation via GCN. From different constructed text graphs, to efficiently learn and transform these structural features of each document into the low-dimensional latent representations, we apply the multi-layered GCN-based architecture [23] to learn the structural representations of occurred words in the given document (d) via the graph-based aggregation learning process. In more details, for the initial layer of a the given GCN layer, it takes the BERT-based word embedding matrix of a given document (d) which is obtained previous, as: $\mathrm{X}_{w}^{\mathrm{BERT}} \in \mathbb{R}^{n \times d}$ as the initial node features with A^{*} is normalized adjacency matrix of each type of constructed GOW. The normalized adjacency matrix is identified as: $A^{*} = \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$, in which: $\widetilde{A} = A + I$ and $\widetilde{D} = \mathrm{diag}\left(\sum_j \widetilde{A}_{ij}\right)$, where:I, $\widetilde{D}$ and $\widetilde{A}$ are the identity matrix, degree matrix and adjacency matrix with self-connection of a given text graph, respectively. Equation 2a shows the general structure of the given initial layer which produce the first hidden state, as: $\mathcal{H}^{\mathrm{GCN},[0]}$ of a given GCN-based architecture for a specific type of text graph. Then for next (l^{th}) layer through the propagation learning process, the corresponding hidden state is produced as shown in the equation 2b.

$$\mathcal{H}^{\mathrm{GCN},[0]} = \mathrm{ReLu}\left(\mathrm{W}^{\mathrm{GCN},[0]}.\mathrm{X}_{w}^{\mathrm{BERT}}.A^{*}\right) \quad (2a)$$

$$\mathcal{H}^{\mathrm{GCN},[l]} = \mathrm{ReLu}\left(\mathrm{W}^{\mathrm{GCN},[l-1]}.\mathcal{H}^{\mathrm{GCN},[l-1]}.A^{*}\right) \quad (2b)$$

At the end of this process, we achieve the final structural word embedding matrix as the last k^{th} hidden state of the given GCN-based architecture, as: $\mathrm{X}^{\mathrm{GCN}} = \mathcal{H}^{\mathrm{GCN},[k]}$ and $\mathrm{X}^{\mathrm{GCN}} \in \mathbb{R}^{n \times d}$ Thus for each type of text graph, we receive different word embedding matrix, denoted as: $\mathrm{X}_{\mathrm{GOWSim}}^{\mathrm{GCN}} = \mathrm{GCN}(\mathcal{G}^{\mathrm{GOWSim}})$ and $\mathrm{X}_{\mathrm{GOWSyn}}^{\mathrm{GCN}} = \mathrm{GCN}(\mathcal{G}^{\mathrm{GOWSyn}})$. Then, we applied a linear fusion mechanism in form of full-connected neural network architecture to obtain the final unified word representation matrix, denoted as: $\mathrm{X}_{d}^{\mathrm{GCN}} \in \mathbb{R}^{n \times d}$ of each document (d), as formulated in the equation 3:

$$\mathrm{X}_{d}^{\mathrm{GCN}} = \mathrm{W}_{\alpha}^{\mathrm{fuse}}.\mathrm{X}_{\mathrm{GOWSim}}^{\mathrm{GCN}} + \mathrm{W}_{\beta}^{\mathrm{fuse}}.\mathrm{X}_{\mathrm{GOWSyn}}^{\mathrm{GCN}} + \mathrm{b}^{\mathrm{fuse}} \quad (3)$$

In this equation, $\mathrm{W}_{\alpha}^{\mathrm{fuse}}$, $\mathrm{W}_{\beta}^{\mathrm{fuse}}$ and $\mathrm{b}^{\mathrm{fuse}}$ are the trainable weighting and bias parameter matrices which are correspondingly updated during the representation learning process. Finally, we achieve the final rich contextual and structural representation of all words in a given document by applying the concatenation operation on the BERT-based (is achieved previously) and this GCN-based word embedding matrices, denoted as: $\mathrm{X}_{d} = [\mathrm{X}_{w}^{\mathrm{BERT}}, \mathrm{X}_{d}^{\mathrm{GCN}}]$.

B. Cluster inference process via DPMM

To utilize previous archived rich contextual and structural representations of input documents for improving the performance of text stream clustering task, we develop a DPMM-based clustering model which can be integrated with the semantic word and document embedding vectors. In order to efficiently incorporate these latent feature representations of input texts into the DPMM framework, we applied the likelihood function for each input document (d), as: $f_{LLH}(d|\Theta)$, which is majorly inherited from previous works [13] [14] for the cluster inference process as the following (as shown in the equation 4):

$$\mathrm{f}_{\mathrm{LLH}}(d|\Theta) = \mathrm{Multi}(\mathcal{W}_d|\delta).\mathcal{N}(\mathrm{X}_d|\psi).\mathcal{N}\left(\overrightarrow{d^{\mathrm{BERT}}}\middle|\eta\right) \quad (4)$$

In this equation, Θ is the set of model's parameters, as: $\Theta = \{\delta, \lambda, \psi\}$ which will be evaluated and learnt during the model's inference process and $\mathcal{N}(.)$ is the normal distribution. This likelihood function indicate the relationships between the independent distributions and rich semantic representations of words in a given document (d). Following previous works [11] [13] [14], to estimate the given model's parameters, we utilize the collapsed Gibbs sampling technique in which can support to significantly decrease the sample drawing space for the cluster inference process. Upon the DPMM paradigm, each (i^{th}) document in a given text corpus/document's batch ($\mathcal{D}$) will be allocated to a specific (k^{th}) cluster which is explicitly identified the posterior distribution, as: $\mathrm{P}(c_i = k|\mathcal{C}_{\neg i}, \mathcal{D}, \Theta)$, where: $\mathcal{C}_{\neg i}$ is the overall cluster assignments except the (i^{th}) document, which can be achieved by the multiplication of the (i^{th}) document likelihood and the prior distribution of a given (k^{th}) cluster, as the following (as shown in equation 5):

$$\mathrm{P}(c_i = k|\mathcal{C}_{\neg i}, \mathcal{D}, \Theta) = \frac{\mathrm{P}(\mathcal{C}|\Theta)}{\mathrm{P}(\mathcal{C}_{\neg i}|\Theta)}.\frac{\mathrm{P}(\mathcal{D}|\mathcal{C}, \Theta)}{\mathrm{P}(\mathcal{D}_{\neg i}|\mathcal{C}, \Theta)} = \mathrm{P}(c_i|\mathcal{C}_{\neg i}, \Theta).\mathrm{P}(d_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta) \quad (5)$$

DPMM-based cluster update process. By formulating different textual features of each document as the joint cluster distribution, we assume that all model's variables

are independent and utilize the factorization approach for the $P(d_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta)$ as the joint distributions of: $P(\mathcal{d}_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta) \propto P(\mathcal{W}_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta).P(X_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta).P(X_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta)$, in which: $P(\mathcal{W}_i|\mathcal{D}_{\neg i}, \mathcal{C}, \Theta) = P(\mathcal{W}_i|c_i = k, \mathcal{D}_{k,\neg i}, \delta)$, $P(X_i|\mathcal{D}_{\neg i}, C, \Theta) = P(X_i|c_i = k, \mathcal{D}_{k,\neg i}, \psi)$ and $P\left(\overrightarrow{d_1^{BERT}}|\mathcal{D}_{\neg i}, C, \Theta\right) = P\left(\overrightarrow{d_1^{BERT}}|c_i = k, \mathcal{D}_{k,\neg i}, \eta\right)$ [11] [13]. Then, inspired from previous works [11], in order to enable the proposed model can effectively handle the one-pass text clustering task and topic/cluster explosion in which multiple input textual documents are sequentially and rapidly come in to the system as stream, we adopted the cluster update mechanism of MStream model [11] in our proposed GOWGCNStream model. To efficiently obtain the distributions of occurring words along with its corresponding rich semantic embedding vectors (X_i) upon the given BERT-based document embedding ($\overrightarrow{d_1^{BERT}}$) over a specific ($k^{th}$) cluster is identified as the following (as shown in equation 6):

$$P(\{\delta, \psi, \eta\}|\mathcal{D}_{k,\neg i}, \beta) = P(\{\delta, \psi, \eta\}|\mathcal{D}_{k,\neg i}, \alpha, \beta) \propto \frac{\mathcal{D}_{k,\neg i}}{\mathcal{D} - 1 + \alpha\mathcal{D}} \cdot \frac{\prod_{w \in d_i} \prod_{j \in freq(w, d_i)} (freq(w, \mathcal{D}_{k,\neg i}) + \beta + j - 1)}{\prod_{l=1}^{d_i} (\mathcal{D}_{k,\neg i} + \mathcal{W}.\beta + l - 1)} \quad (6)$$

In this equation [11], α and β are the hyper-parameters, the $\mathcal{W}$ and $\mathcal{D}_{k,\neg i}$ are the vocabulary set and a set of documents which have been assigned to the (k^{th}) cluster except the given (i^{th}) document in each document's batch. The $freq(w, d_i)$ and $freq(w, \mathcal{D}_{k,\neg i})$ are the occurrence frequency of a specific word (w) in a given (i^{th}) document and a set of documents belong to (k^{th}) cluster except the given (i^{th}) document.

IV. EXPERIMENTS & DISCUSSIONS

To demonstrate the effectiveness of our proposed GOWGCNStream model in comparing with recent state-of-the-art baselines, we conducted extensive experiments in benchmark text stream based dataset such as Tweet-Set and Google-News. The comparative experimental outputs of the GOWGCNStream model with recent state-of-the-art text stream clustering baselines showed the usefulness of our proposed model and associated ideas in this paper.

A. Dataset & experimental setups

1) Dataset descriptions & pre-processing steps

To evaluate the performance of different methods in text stream clustering task, similar to previous studies [11] [12] [18], we mainly utilized two well-known standard datasets, which are: Tweet-Set and Google-News. The detailed information of these two datasets are:

- **Tweet-Set (Tw/Tw-T)**: is considered as a common dataset for evaluating multiple primitive tasks in NLP domain, including text stream clustering. This dataset contains 30K documents in forms of tweets (comments or micro-blogs in the Twitter social network). The Tweet-Set is considered as a short text corpus in which the average document's length is about 7.72. All documents in this dataset are categorized into 269 topics/clusters. For the experiments with in short text stream clustering task, we also divide this dataset into 16 different document's batches [11] [18]. Then, we also constructed the synthetic version of this dataset, named as: Tw-T, following previous works [11] [18] in which this dataset is randomly divided in to different batches and shuffle all documents in each batch. This dataset can be downloaded at this website[1].
- **Google-News** (GN/GN-T): similar to the Tweet-Set, the Google-News dataset is also popular in experimental studies for short text stream clustering task. This dataset contains about 11K documents in forms of articles/news which are collected from Google News online platform[2]. All documents in this dataset are also considered as short texts with average length is about 6.23 and they are categorized into 152 topics/clusters. Similar to the Tweet-Set, we also divided this dataset into 16 different batches and created a synthetic version of this dataset, named as: GN-T for experiments in short text stream clustering task.

Textual pre-processing steps. For initial textual pre-processing steps, such as: stop-word filtering, word tokenization, etc. we mainly applied the Stanford CoreNLP library [31]. The Stanford CoreNLP library [31] is also applied to extract syntactic relationships between words in each document of these two datasets. Table 1 show general statistics of the Tweet-Set and Google-News datasets for all experiments in our paper after textual pre-processing steps.

Table 1. General summary of used datasets for experiments in this paper

Dataset	No. docs	No. clusters	Vocabulary size	Avg. length
Tw/Tw-T	30,289	269	12,301	7.72
GN/GN-T	11,109	152	8,110	6.23

2) Evaluation methods & configurations

Evaluation method for text clustering task. To evaluate the performance of different text stream clustering methods in our experiments, we mainly applied the common Normalized Mutual Information (NMI) metric which are mainly utilized in recent studies [11] [14] [18]. The NMI accuracy score for a clustering output is identified as: $NMI = \frac{\sum_{c,k} n_{c,k} log\left(\frac{N.n_{c,k}}{n_c.n_k}\right)}{\sqrt{\left(\sum_c n_c log\frac{n_c}{N}\right).\left(\sum_k n_k log\frac{n_k}{N}\right)}}$, in which, the n_c and n_k are the number of documents in a class (c) and number of documents in a cluster (k), respectively. The $n_{c,k}$ and N present for the number of documents in both class (c) and cluster (k) and the total number of documents in a given document's batch. For experiments in text stream clustering task with different models, we conducted the experiments 10 times and reported the average NMI-based score as the final result for each model.

Experimental configurations for GOWGCNStream. We implemented our proposed GOWGCNStream by using Python programming language with the support of

[1] Tweet-Set dataset: http://trec.nist.gov/data/microblog.html

[2] Google News platform: https://news.google.com/

PyTorch machine learning framework[3]. The proposed GOWGCNStream and other comparative baselines are set up in a single server with Intel Xeon SKL-SP 4210 CPU and 64Gb memory. For the implementation of pre-trained BERT model which is used for our contextual text representation learning process, we reused the large/uncased pre-trained BERT version which is provided by Google. The detailed information and implementation of this pre-trained BERT version can be achieved at this GitHub repository[4]. For the setup of our Bi-LSTM architecture which is used to capture the completed representations for sentences and documents from BERT (as described in section III.A.1)), we set the number of LSTM-based cells, as: $k^{\text{Bi-LSTM}} = 400$. For the utilization of GCN based architecture in learning the structural representations of words in each document (as described in section III.A.2)), we configured the number of GCN-based layers for each type of text graph, as: $k^{GCN} = 3$. Table 2 listed other configuration parameters of our proposed GOWGCNStream model for comparative experiments with other text streaming clustering baselines in this paper.

Table 2. General configurations of our proposed GOWGCNStream for experiments in this paper

Parameter of GOWGCNStream	Value
Dimensonality of word embedding vector by using pre-trained BERT (d^{BERT})	768
Number of LSTM-based cells ($k^{\text{Bi-LSTM}}$) for the used Bi-LSTM archiecture.	400
Number of layer (k^{GCN}) for the given GCN-based archiecture.	3
Number of iteration for each document's batch	8
Hyper-parameter α	0.05
Hyper-parameter β	0.05

3) Comparative methods for text stream clustering

To evaluate the performance of our proposed model in comparing with recent state-of-the-art baselines, we implemented several notable text stream clustering methods in our experiment, which are:

- **DTM** [1]: is considered as an earliest text clustering approach which is introduced to efficiently deal with the text clustering task in the context of dynamism. Majorly inherited from the traditional topic modelling approach such as LDA, in DTM model., Blei, D. M., et al. proposed a novel latent topic inference process to efficiently handle the topic evolution within text corpus, however the DTM still suffered limitations regarding with the topic/concept drift challenge and low-quality in textual representation in which can't achieve best results for text stream clustering task.
- **Sumblr** [7]: is considered as the most notable vector space similarity-based model for text stream clustering task. In Sumblr model, Shou, L. et al. [7]: proposed a novel textual feature vectorization and similarity-based measurement techniques to identify appropriated topics/clusters for the input short-length documents like as tweets, comments, micro-blogs, etc. in forms of continuous text streams in social network platforms. Although Sumblr demonstrates remarkable performances in short text clustering, it has still been considered as unable to deal with the topic/cluster evolution challenge due to the number of topics/clusters must be initially predefined at the beginning.
- **MStream** [11]: is a recent well-known DPMM-based model for handling short text stream clustering task. In the MStream model, Jianhua Yin et al. proposed a novel cluster update/forget mechanism [11] along with the cluster inference process to effectively cope with the topic/cluster explosion within a given text stream. Extensive experiments in benchmark datasets show the outperformances of proposed MStream model in comparing with previous text stream clustering baselines.
- **OSDM** [18]: is a recent DPMM-based text stream clustering technique which is majorly inherited from the MStream model [11]. To improve the representations of textual documents in stream through the clustering process, in the OSDM model [18], Kumar, J. et al. proposed a novel word's co-occurrence relationship evaluation during the cluster inference process to leverage the quality of extracted cluster information from stream.

For the unique configurations of all implemented text stream clustering methods which are listed above, we used

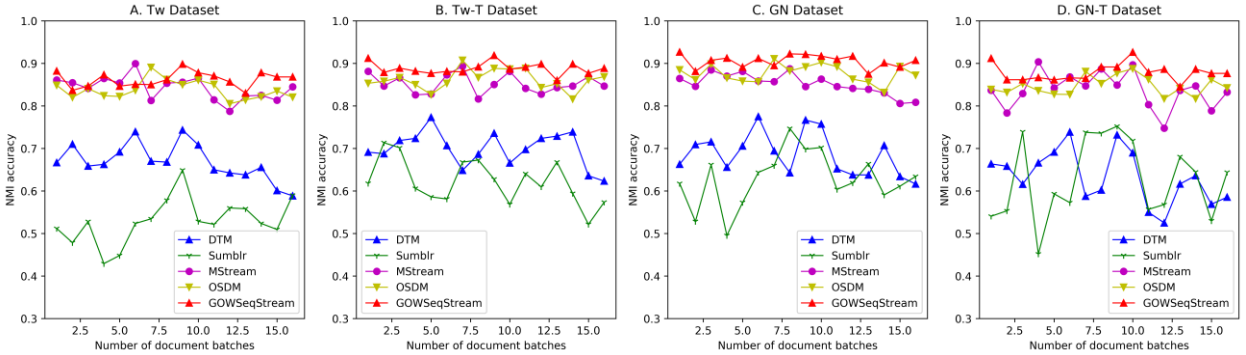

Figure 2. The accuracy performances in terms of NMI metric of different text stream clustering techniques with different document's batches of Tw/Tw-T and GN/GN-T datasets

[3] PyTorch: https://pytorch.org/

[4] Pre-trained BERT (large/uncased version): https://github.com/google-research/bert

the same values of these model's configuration parameters as described in the original papers in which these methods gained the highest performances. For common configurations, we set them as the same with the our proposed GOWGCNStream model as described in Table 2.

B. Experimental results & discussions

In this section, we present experimental outputs for text stream clustering task with different baselines in two benchmark datasets, which are: Tw/Tw-T and GN/GN-T. As shown from the experimental outputs in , out proposed GOWGCNStream achieved better performances than all other text stream clustering methods in both Tw/Tw-T and GN/GN-T datasets. In more details, our proposed GOWGCNStream model achieved explicit better performances than previous traditional topic modelling and vector space similarity-based approaches like as DTM and Sumblr about 32.56% and 48.1%, respectively in terms of NMI metrics for all datasets. For our main competitors, like as MStream and OSDM, our proposed GOWGCNStream model also slightly improve the performance for text stream clustering task in both Tw/Tw-T and GN/GN-T datasets approximately 5.2% and 3.30% in terms of NMO evaluation metric, respectively.

Moreover, as shown from the accuracy performance outputs (in Figure 3) of different text stream clustering methods at different document's batches in Tw/Tw-T and GN/GN-T datasets, our proposed GOWGCNStream model and other DPMM-based methods (MStream and OSDM) achieved the quite stable performances in all document's batches, whereas the traditional topic modelling and vector space similarity-based methods suffered oscillations in different document's batches. This experimental outputs not only demonstrate the stability of our proposed GOWGCNStream but also prove that our proposed model can work efficiently in short text streams with huge number of clusters in each document's batch.

C. Ablation studies

In this section, we present several extensive empirical studies in the influences of model's configuration parameters on the overall accuracy and stability performances of our proposed GOWGCNStream model. In this ablation study section, we evaluate the effects of number of model's iterations per document's batch and two main model's hyper-parameters: α and β.

Number of iterations per document's batch. For the model's cluster inference process, it requires a sufficient number of iterations for the sampling and parameter estimation procedure. The number of iterations is important in which must be carefully taken in consideration to make the proposed model can reach the convergence point with a reasonable computational effort. Thus, in order to evaluate the influence of this parameter in our GOWGCNStream model, we run our model with different number of iterations (in range 1-10) in each document's batch while still kept other model's parameters as the same. Figure 4 shows the average experimental outputs in terms of NMI metric (Figure 4-A) and number of extracted clusters (Figure 4-B) of our proposed GOWGCNStream with different number of model's iterations in each document's batch. As shown from the experimental outputs, our proposed GOWGCNStream model achieved the highest performance and become stable with number of iterations > 7 for all Tw/Tw-T and GN/GN-T dataset which is considered as a reasonable value for this parameter in order to ensure both accuracy and time complexity performance. Moreover, as shown from the Figure 4-B, with the changes on number of model's iterations parameter, the number of extracted clusters in each document's batch still be stable and does not change much which prove that our model can efficiently cope with the cluster explosion challenge in text stream clustering task.

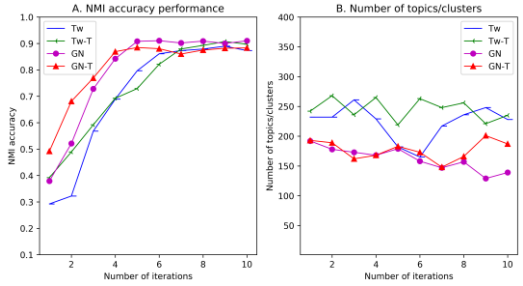

Figure 3. Experimental studies on the influence of number of iterations per document's batch of our proposed GOWGCNStream model

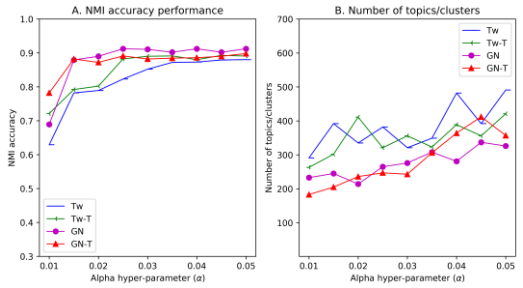

Figure 4. Experimental studies on the influence of α hyper-parameter of our proposed GOWGCNStream model

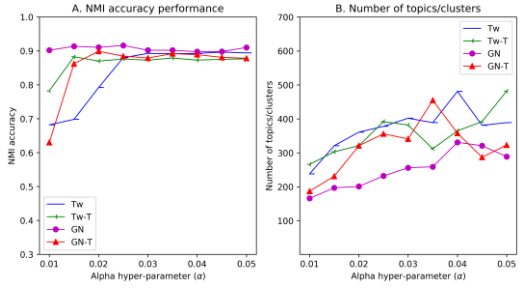

Figure 5. Experimental studies on the influence of β hyper-parameter of our proposed GOWGCNStream model

Hyper-parameter: α and β. In this section we presented extensive experimental studies on the influences of α and β hyper-parameters upon the overall accuracy and stability performances of our proposed GOWGCNStream. Similar to the experiments with the number of model's iterations parameter, we varied the values of these two hyper-parameters within range [0.01-0.05] and reported the accuracy and stability performances in terms of NMI score and quantity of extracted clusters per document's batch.

As shown from the experimental outputs in Figure 4 and Figure 5, our proposed GOWGCNStream is quite insensitive with these two hyper-parameters in which it achieved the convergence point with highest accuracy

performance in terms of NMI metric (as shown in Figure 4-A and Figure 5-A) with values of $\alpha > 0.03$ and $\beta > 0.04$ for both Tw/Tw-T and GN/GN-T datasets. Similar to that with the stability performance evaluation under the reported number of extracted clusters in each document's batch (as shown in Figure 4-B and Figure 5-B).

V. CONCLUSIONS & FUTURE WORKS

In this paper, we propose a novel approach of rich semantic-enhanced short stream clustering model, called as GOWGCNStream. In our proposed GOWGCNStream model, we utilize the combination between pre-trained BERT and GCN to learn the rich contextual and structural representations of input documents from stream in order to assist the DPMM-based inference process for effectively extracting cluster information from a given stream. By combining the DPMM framework with recent advanced deep neural network architectures, like as BERT and GCN for rich semantic textual representation learning process, we can equip our proposed GOWGCNStream model the capability of effectively dealing with the textual data sparsity/noise and topic/concept drift challenges which normally encounter in text stream analysis and mining task. Extensive experiments in benchmark datasets demonstrated the outperformances of our proposed GOWGCNStream model in this paper as well as promising directions for further improvements in text stream clustering task. For our future works, we intend to implement the GOWGCNStream model upon the distributed data stream processing framework like as Spark Stream in order to assist our proposed method can work well in large-scale text streams.

VI. ACKNOWLEDGEMENT

This research is funded by Thu Dau Mot University, Binh Duong, Vietnam.

VII. REFERENCES

[1] David Blei, and Lafferty John D., "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006.

[2] Wang Xuerui, and Andrew McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[3] Wei Xing, Sun Jimeng, and Wang Xuerui, "Dynamic Mixture Models for Multiple Time-Series," *IJCAI*, vol. 7, pp. 2909-2914, 2007.

[4] Aggarwal Charu C, Philip S Yu, Han Jiawei, and Wang Jianyong, "A framework for clustering evolving data streams," in *In Proceedings 2003 VLDB conference*, 2003.

[5] Zhong Shi, "Efficient streaming text clustering," *Neural Networks*, vol. 18, no. 5-6, pp. 790-798, 2005.

[6] Cao Feng, Estert Martin, Qian Weining, and Zhou Aoying, "Density-based clustering over an evolving data stream with noise," in *Proceedings of the 2006 SIAM international conference on data mining*, 2006.

[7] Shou, Lidan, et al., "Sumblr: continuous summarization of evolving tweet streams," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013.

[8] Jianhua Yin and Jianyong Wang, "A Text Clustering Algorithm Using an Online Clustering Scheme for Initialization," in *ACM International Conference on Knowledge Discovery and Data Mining*, 2015.

[9] Yukun Zhao, Shangsong Liang, Zhaochun Ren, Jun Ma, Emine Yilmaz, and Maarten de Rijke, "Explainable User Clustering in Short Text Streams," in *International ACM conference on Research and De- velopment in Information Retrieval*, 2016.

[10] Liang Shangsong, Yilmaz Emine, and Kanoulas Evangelos, "Dynamic Clustering of Stream- ing Short Documents," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowl- edge Discovery and Data Mining*, 2016.

[11] Jianhua Yin, Daren Chao, Zhongkun Liu, Wei Zhang, Xiaohui Yu, and Jianyong Wang, "Model-based Clustering of Short Text Streams," in *ACM International Conference on Knowledge Discovery and Data Mining*, 2018.

[12] Chen, Junyang, Zhiguo Gong, and Weiwen Liu, "A Dirichlet process biterm-based mixture model for short text stream clustering," *Applied Intelligence*, pp. 1-11, 2020.

[13] Duan Tiehang, Lou Qi, Srihari Sargur N, and Xie Xiaohui, "Sequential embedding induced text clustering, a non-parametric bayesian approach," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2019.

[14] Junyang Chen, Zhiguo Gong, and Weiwen Liu, "A nonparametric model for online topic discovery with word embeddings," *Information Sciences*, vol. 504, pp. 32-47, 2019.

[15] Mikolov Tomas, Chen Kai, Corrado Greg, and Dean Jeffrey, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations (ICRL)*, 2013.

[16] Pennington Jeffrey, Richard Socher, and Christopher D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[17] Mikolov Tomas, Grave Edouard, Bojanowski Piotr, Puhrsch Christian, and Joulin Armand, "Advances in Pre-Training Distributed Word Representations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[18] Kumar Jay, Shao Junming, Uddin Salah, and Ali Wazir, "An Online Semantic-enhanced Dirichlet Model for Short Text Stream Clustering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[19] Liu Xien, You Xinxin, Zhang Xiao, Wu Ji, and Lv, Ping, "Tensor graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[20] Zhang Haopeng, and Jiawei Zhang., "Text Graph Transformer for Document Classification," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[21] Jia, Ruipeng, et al., "Neural Extractive Summarization with Hierarchical Attentive Heterogeneous Graph Network," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[22] Hamilton Will, Zhitao Ying, and Jure Leskovec., "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[23] Kipf, Thomas N., and Max Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations*, 2017.

[24] Velickovic Petar, Cucurull Guillem, Casanova Arantxa, Romero Adriana, Lio Pietro, and Bengio Yoshua, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[25] Sutskever Ilya, Oriol Vinyals, and Quoc V. Le., "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, 2014.

[26] Bahdanau Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *3rd International Conference on Learning Representations, ICLR*, 2015.

[27] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Lukasz, and Polosukhin Illia, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[28] Peters ME,Neumann M,Iyyer M,Gardner M,Clark C,Lee K,and Zettlemoyer L, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[29] Radford, Alec, et al., "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[30] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[31] Manning Christopher D, Surdeanu Mihai, Bauer John, Finkel Jenny Rose, Bethard Steven, and McClosky David, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014.

[32] Blei David M., Andrew Y. Ng, and Michael I. Jordan., "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993-1022, 2003.

[33] McAuliffe Jon D., David M. Blei, and Michael I. Jordan., "Nonparametric empirical Bayes for the Dirichlet process mixture model," *Statistics and Computing*, vol. 16, no. 1, pp. 5-14, 2006.

[34] Li, Yuelin, Elizabeth Schofield, and Mithat Gönen, "A tutorial on dirichlet process mixture modeling," *Journal of mathematical psychology*, vol. 91, pp. 128-144, 2019.

CẢI TIẾN HIỆU QUẢ GOM CỤM LUỒNG VĂN BẢN NGẮN THÔNG QUA HỌC BIỂU DIỄN ĐẶC TRƯNG VĂN BẢN TÍCH HỢP ĐỒ THỊ TỐI ƯU NGỮ NGHĨA

Tóm tắt: Gom cụm luồng văn bản đối mặt nhiều thách thức như độ rời rạc, độ nhiễu, độ dài vô hạn và sự thay đổi cụm của các tài liệu đến trên luồng. Trong những năm gần đây, nhiều mô hình chủ đề hỗn hợp, chẳng hạn như: Các thuật toán dựa trên Mô hình hỗn hợp quy trình Dirichlet (DPMM) (ví dụ: MStream, OSDM, v.v.) đã chứng minh những cải tiến đáng kể độ chính xác của tác vụ gom cụm luồng văn bản ngắn. Tuy nhiên, các mô hình dựa trên DPMM mới này vẫn tồn tại những hạn chế liên quan đến khả năng nắm bắt đầy đủ các mối quan hệ phụ thuộc cú pháp tuần tự và phạm vi dài giữa các từ trong văn bản mà điều này có thể hỗ trợ nâng cao chất lượng của các cụm. Để giải quyết những vấn đề này, trong bài báo này, chúng tôi đã đề xuất một mô hình mới dùng kết hợp mạng đồ thị (GCN) với DPMM để gom cụm luồng văn bản, GOWGCNStream. Đây là mô hình kết hợp GCN với BERT nhằm nắm bắt cấu trúc cú pháp chung và biểu diễn ngữ cảnh của văn bản, từ đó làm tiền đề cho DPMM gom cụm luồng văn bản ngắn tốt hơn. Các thử nghiệm sử dụng các tập dữ liệu chuẩn (Tweet-Set và Google-News) đã chứng minh tính hiệu quả của mô hình GOWGCNStream, so sánh với các thuật toán gần đây.

Từ khóa: BERT; DPMM; GCN; Đồ thị từ; Gom cụm luồng văn bản.

AUTHORS' BIOGRAPHIES

Tham Thi Hong Vo received the M.S. degree in Computer Science from University of Information Technology (UIT), Ho Chi Minh, Vietnam in 2009. She received the PhD degree from Lac Hong University, Dong Nai, Vietnam in 2021. She is working at Institute of Engineering and Technology, Thu Dau Mot University, Binh Duong, Vietnam. Her researches focus on Text Mining, Data Stream Analysis, Deep Learning, Information Network Analysis & Mining, Text Representation Learning.