An Effective Vector Representation of Facebook Fan Pages and Its Applications

Viet Hoang Phan[1], Duy Khanh Ninh[1(✉)], and Chi Khanh Ninh[2]

[1] The University of Danang, University of Science and Technology, Danang, Vietnam
hoangvietit15@gmail.com, nkduy@dut.udn.vn
[2] The University of Danang – Vietnam - Korea, University of Information and Communication Technology, Danang, Vietnam
nkchi@vku.udn.vn

Abstract. Social networks have become an important part of human life. There have been recently several studies on using Latent Dirichlet Allocation (LDA) to analyze text corpora extracted from social platforms to discover underlying patterns of user data. However, when we wish to discover the major contents of a social network (e.g., Facebook) on a large scale, the available approaches need to collect and process published data of every person on the social network. This is against privacy rights as well as time and resource consuming. This paper tackles this problem by focusing on fan pages, a class of special accounts on Facebook that have much more impact than those of regular individuals. We proposed a vector representation for Facebook fan pages by using a combination of LDA-based topic distributions and interaction indices of their posts. The interaction index of each post is computed based on the number of reactions and comments, and works as the weight of that post in making of the topic distribution of a fan page. The proposed representation shows its effectiveness in fan page topic mining and clustering tasks when experimented on a collection of Vietnamese Facebook fan pages. The inclusion of interaction indices of the posts increases the fan page clustering performance by 9.0% on Silhouette score in the case of optimal number of clusters when using K-means clustering algorithm. These results will help us to build a system that can track trending contents on Facebook without acquiring the individual user's data.

Keywords: Topic modeling · Latent Dirichlet Allocation · Interaction index · Facebook fan pages · Social network analysis

1 Introduction

Nowadays, social networks have become an essential part of human life. Based on a recent research of Statista, more than 3.5 billion people on earth have at least one account on a social platform in 2019 [1]. With rapid growth of users, comes giant amount of data. This can bring a lot of opportunities for those who can discover patterns inside of the user data and find out meaningful usages of them.

M. Hernes et al. (Eds.): ICCCI 2020, CCIS 1287, pp. 674–685, 2020.
https://doi.org/10.1007/978-3-030-63119-2_55

In Vietnam, Facebook is the social network having the largest number of users [1]. Posts on Facebook can come from individuals (particularly famous figures) or from organizations in a form of what is called a fan page. Because of the great ability of sharing posts to the fans (i.e., people who follow pages), fan pages are playing an important role in spreading information, news, and facts on Facebook. If we can model the topics of posts on popular pages, we will have a good chance to find out trending contents on Facebook.

In recent years, there have been a lot of researches on using Latent Dirichlet Allocation (LDA) to cluster the scientific documents [2, 3] and news articles [4, 5]. For social networks document analysis, there were some studies about modeling the topic on Twitter [6, 7] or favorite topics on Facebook posts [8]. This research focuses on modeling Facebook fan pages by using the method of topic modeling from documents (i.e., the fan page's posts).

In this paper, we propose a solution of modeling the topic of documents with LDA combined with calculating the interaction index of the Facebook posts to find an effective vector representation of Facebook fan pages. Then we apply this representation to analyze topic distribution of each fan page and to find out groups of similar fan pages. The proposed solution shows the effectiveness on clustering the fan pages into subsets by increasing the clustering performance than modeling using just LDA. The fan page representation also helps point out the similarities between fan pages and give us an idea about what is happening on Facebook in a particular period of time.

This paper is organized as follows. Section 2 reviews past studies leading to the motivation of our work. Section 3 describes our proposed solution. Experiments and results are given in Sect. 4. Section 5 presents the conclusion and future work of the current research.

2 Related Work and Motivation

Topic modeling using LDA is not a new technique in Natural Language Processing. LDA uses an unsupervised learning model, therefore it is a good technique for document classification, especially on unlabeled datasets such as social network's textual data. There were several researches taking this advantage of LDA to model and analyze Twitter conversations [6] or favorite topics of young Thai Facebook users [8]. The main focus of these studies is the modeling and mining the topics from the text corpus of social network users. Their proposed methods can help us to obtain the topics in which a part of users interested in, for example educational workers and students at National University of Colombia [6] or students at Assumption University in Thailand [8]. However, when we wish to discover the major contents of a social network on a large scale such as finding trending topics among users of a nation, the available approaches exhibit their limitations, that is it is almost impossible to collect published data (i.e., the posts) of every person on the social network because it goes against privacy rights as well as takes a lot of time and resource to collect and process the data.

This paper tackles this problem by focusing on a class of special users that have much more impact on social networks than other individuals, which are key opinion leaders (KOLs) and popular organizations. A post of a KOL or a well-known organization,

usually on their fan pages, may lead the opinions, represent for the thoughts, and attract the interests of many people which follow them on social networks. Therefore, instead of collecting data from each regular account on a social network, we only need to get and analyze data from a number of influential accounts of KOLs and organizations, thereby achieving the equivalent effectiveness in capturing the trends of the social network on a large scale. In this paper, we selected the most reputable Facebook fan pages in Vietnam for topic mining and other data analyses.

3 A Novel Solution for Facebook Fan Page Modeling

3.1 Observations

To know what a Facebook fan page is talking about, we have to analyze the contents of its posts. In this research, we are only interested in the textual part (called the document hereinafter) and the interactional part (i.e., reactions and comments) of a post. If we can extract the topics of every document in the text corpus of a fan page, we probably can figure out the most popular topics of a fan page.

Assuming that each document in a fan page's corpus has its own topic probability distribution or, in other words, each document can be represented by a fixed-dimensional vector depending on what its content is about. For example, if a document has its topic proportions of 30% about sport, 50% about technology, 20% about politics, the topic distribution vector of this document will be [0.3;0.5;0.2]. In practice, the results of vectorizing documents are not clearly visible like the above example, but usually hidden in the textual data. We need a solution to combine the topic distribution vectors of the documents in a fan page's corpus to find the topic distribution vector representing the fan page.

After studying about Facebook data properties, we realized that different posts (thus their corresponding documents) have different degrees of importance to the topic proportions of a fan page. The posts that receive more interactions from users are likely to contribute more to the composition of the topics of a fan page and to the distinction among fan pages.

3.2 Proposed Process Diagram

Figure 1 presents the proposed process flow. Firstly, the raw data of fan pages are collected from crawlers, from which textual data and interactional data are extracted. After that, the textual data is pre-processed by removing page signatures, special characters, icons, and stop words. Pre-processed documents are then applied LDA-based topic modeling process, returning topic distribution vectors of all documents of each fan page's corpus. Meanwhile, the interactional data is used to calculate the interaction indices of all documents of each fan page. Finally, the vector representation for every fan page is obtained by combining the topic distribution vectors and interaction indices of all documents of the page. How the combination is done is described in details in Sect. 3.4.

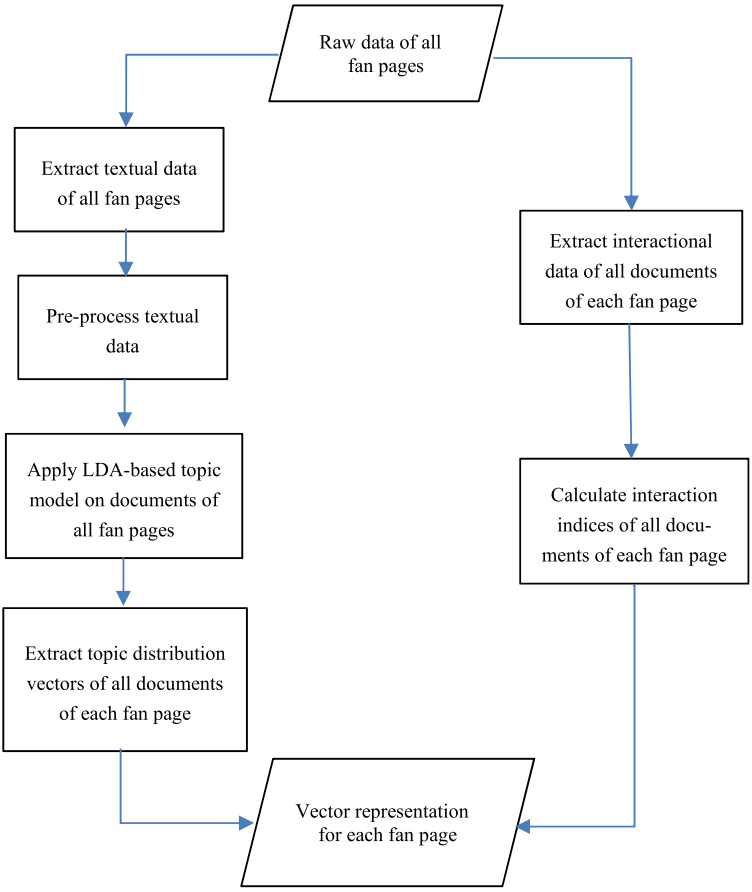

Fig. 1. Process diagram of the proposed solution.

3.3 Topic Modeling Using LDA

LDA is a method widely employed for modeling the topics of documents in a corpus, which was proposed by Blei et al. in 2003 [9]. This method assumes that each document in the corpus is a probability distribution of topics and each topic is a probability distribution of words in the vocabulary of the corpus. Given a corpus D, LDA assumes that the corpus can be generated by the following process [10] (Fig. 2):

Step 1. For each topic k in K topics, draw a distribution over words in the vocabulary:

$$\varphi(k) \sim Dirichlet(\beta)$$

Step 2. For each document $d \in D$:
a) Draw a distribution over topics of the document:

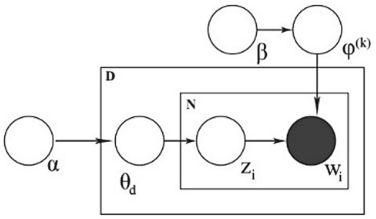

Fig. 2. Graphical representation of LDA model (modified from [9]).

$$\theta_d \sim Dirichlet(\alpha)$$

b) For each word $w_i \in d$:
i. Draw a topic assignment: $z_i \sim Discrete(\theta_d)$
ii. Choose the word: $w_i \sim Disctete\left(\varphi^{(z_i)}\right)$

where K is the number of latent topics in the corpus and α, β are the parameters of the corresponding Dirichlet distributions.

The above process results in the following joint distribution [10]:

$$p(\mathbf{w}, \mathbf{z}, \theta, \varphi|\alpha, \beta) = p(\varphi|\beta)p(\theta|\alpha)p(\mathbf{z}|\theta)p(\mathbf{w}|\varphi_z) \quad (1)$$

where **w** is the vocabulary and **z** is the topic assignment for each word in w.

3.4 Fan Page Representation Using Vector

Figure 3 illustrates the process to obtain the topic distribution vector for a particular document in a fan page's text corpus by using LDA. LDA model gives us two outputs, the cluster of words for each topic and the topic assignment for each word in the corpus. Therefore, we can know exactly how many times a topic appears in the document or, in other words, how many times a topic has been assigned to any word in the document by counting. We then get the topic distribution vector of the document by calculating the probability of each topic being assigned to a word in that document. Consequently, we can generate the topic distribution vector for each fan page in some way.

We propose a simple way to calculate the topic distribution vector for each fan page by summing over the topic distribution vectors of all documents in its corpus. However, each document in the sum should be associated with a weight reflecting how interactive its corresponding post is, as presented in Sect. 3.1. Therefore, we additionally propose to use the number of reactions (e.g., like, haha, angry, etc.) and the number of comments on each post as the parameters to compute the weight of that post, thus its document, in making of the topic distribution of a fan page.

Let $V = \{v_1; v_2; \ldots; v_n\}$ the set of topic distribution vectors of the documents of a fan page's corpus; n is the number of documents of the corpus; t_i, r_i, c_i are respectively

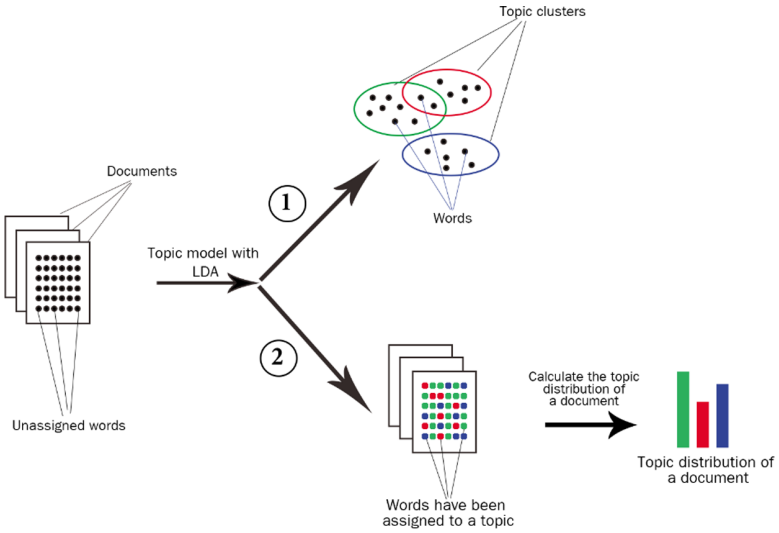

Fig. 3. Process to obtain the topic distribution of a document by using LDA.

the interaction index, number of reactions and number of comments of the ith document ($1 \leq i \leq n$). The interaction index of the ith document can be calculated as

$$t_i = \eta r_i + \mu c_i, \tag{2}$$

where η, μ respectively represents the relative importance between reactions and comments in the interaction index. Since comments are considered more valuable than reactions in terms of the degree of interaction, we experimentally set $\eta = 1$ and $\mu = 3$.

Let P the topic distribution vector represented a fan page. P can be calculated as the weighted sum of topic distribution vectors of all documents of the page, i.e.

$$P = w_1 v_1 + w_2 v_2 + \ldots + w_i v_i + \ldots w_n v_n, \tag{3}$$

where the weight of each document is its interaction index normalized among all documents of the fan page, i.e.

$$w_i = \frac{t_i}{\sum_{i=1}^{n} t_i}. \tag{4}$$

Finally, we can rewrite the vector representation of the page as:

$$P = \frac{\sum_{i=1}^{n} (t_i \times v_i)}{\sum_{i=1}^{n} t_i}. \tag{5}$$

4 Experiments and Results

4.1 Data

The data for this project was crawled from the top fan pages that have the biggest fanbase in the Media category of Vietnamese Facebook (according to the ranking of socialbakers.com in October, 2019 [11]). Details of the dataset are described as below:

- Number of fan pages: 100
- Number of posts (documents): 27,226
- Number of unique words (segmented by the pyvi toolkit [12]): 56,135
- Total number of words: 743,725
- Timespan: during October, 2019

4.2 Experimental Settings

All experiments were conducted using the scikit-learn toolkit [13]. We used LDA model for the document topic modeling process with the following parameters:

- Number of topics: $K = 20$
- Parameters of Dirichlet distributions: $\alpha = \beta = \frac{1}{K} = 0.05$

If the number of topics is too small, there will be little diversity among topic distributions of the corpus. On the contrary, if the number of topics is too big, it will be difficult to interpret what the topics are about since the topics are not obvious anymore. Therefore, we set the number of topics K to 20 in the experiments.

4.3 Topic Modeling Results

Table 1 presents the topic modeling results based on LDA method by showing the top 10 keywords of 20 topics. We can observe that several topics represent quite well about hot events or issues happening in October, 2019. For example, Topic 8 is clearly about the football match between national teams of Vietnam and Malaysia inside the World Cup 2022 qualification round with keywords such as "việt_nam" (*Vietnam*), "malaysia", "trận" (*match*); Topic 5 can be associated with the protest escalation in Hong Kong with the keywords such as "hồng_kông" (*Hong Kong*), "biểu_tình" (*demonstrate*), "dân_chủ" (*democracy*); or Topic 3 can be identified as the air pollution spike in Hanoi due to the keywords such as "không_khí" (*air*), "ô_nhiễm" (*pollution*), "hà_nội" (*Hanoi*), etc. Other topics about daily issues also can be easily identified from their keywords such as Topic 0 (about fashion and music), Topic 6 (about technology and cellphone), Topic 14 (about food and restaurant), to name just a few.

4.4 Fan Page Modeling Results

Based on the outputs of LDA, we can infer the topic distribution vector of a document. Since a fan page can publish multiple documents with different topics, we can represent

Table 1. Top 10 keywords of 20 topics found by LDA (English translation in parentheses).

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
đẹp *(beauty)*	đồng *(currency)*	công_an *(police)*	không_khí *(air)*	tập *(episode)*
thời_trang *(fashion)*	công_ty *(company)*	vụ *(case)*	ô_nhiễm *(pollution)*	phim_truyện *(drama)*
nghệ_sĩ *(artists)*	tiền *(money)*	án *(crime)*	bệnh *(disease)*	khám_phá *(discovery)*
nữ *(female)*	đầu_tư *(invest)*	điều_tra *(investigate)*	hà_nội *(Hanoi)*	thế_giới *(world)*
diễn_viên *(actor/actress)*	hàng *(product)*	cảnh_sát *(police)*	môi_trường *(environment)*	tập_h *(episode)*
ca_sĩ *(singer)*	giá *(price)*	bắt *(arrest)*	bác_sĩ *(doctor)*	vtv *(stands for Vietnam Television)*
nàng *(she/her)*	trung_quốc *(China)*	tp *(stands for city)*	sức_khỏe *(health)*	tiếng *(language)*
mv *(stands for Music Video)*	thương_mại *(commerce)*	tỉnh *(province)*	bệnh_viện *(hostpital)*	việt_nam *(Vietnam)*
nhạc *(music)*	kinh_tế *(economy)*	đối_tượng *(criminal)*	ung_thư *(cancer)*	phim *(movie)*
khán_giả *(audience)*	triệu *(million)*	thông_tin *(information)*	thuốc *(pharmacy)*	đi *(go)*
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**
trump	việt_nam *(Vietnam)*	trung_quốc *(China)*	việt_nam *(Vietnam)*	thvl *(stands for Vinh Long Television)*
tổng_thống *(president)*	game	việt_nam *(Vietnam)*	trận *(match)*	kênh *(channel)*
trung_quốc *(China)*	ứng_dụng *(application)*	tàu *(ship)*	malaysia	đội *(team)*
đảng *(party)*	công_nghệ *(technology)*	biển *(sea)*	đội *(team)*	full
biểu_tình *(demonstrate)*	sự_kiện *(event)*	quốc_tế *(international)*	trận_đấu *(match)*	tình_yêu *(love)*
hồng_kông *(Hong Kong)*	thông_tin *(information)*	khu_vực *(region)*	hlv *(coach)*	fanpage
chính_trị *(politics)*	diễn *(perform)*	biển_đông *(East Sea)*	cầu_thủ *(footballer)*	youtube
dân_chủ *(democracy)*	iphone	nga *(Russia)*	đội_tuyển *(team)*	lỡ *(miss)*
điều_tra *(investigate)*	trải_nghiệm *(experience)*	mỹ *(the US)*	bóng *(ball)*	cực *(very)*
hoa_kỳ *(the US)*	điện_thoại *(cellphone)*	máy_bay *(aircraft)*	sân *(stadium)*	phát_sóng *(broadcast)*
Topic 10	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**
mẹ *(mother)*	yêu *(love)*	trường *(school)*	xe *(vehicle)*	đi *(go)*
đi *(go)*	đi *(go)*	học *(study)*	đường *(street)*	món *(food)*
vợ *(wife)*	chăng *(not)*	trẻ *(kid)*	đi *(go)*	rau *(vegetable)*
tao *(me)*	đừng *(don't)*	học_sinh *(student)*	cháy *(burn)*	bánh *(cake)*
tiền *(money)*	sống *(live)*	lớp *(class)*	hà_nội *(Hanoi)*	lắm *(very much)*
chồng *(husband)*	phụ_nữ *(woman)*	đại_học *(university)*	dân *(resident)*	ngon *(delicious)*
bé *(baby)*	đàn_ông *(man)*	tiếng *(language)*	chạy *(run)*	đồ *(food)*
đứa *(kid)*	hai *(two)*	giáo_dục *(education)*	giao_thông *(traffic)*	thịt *(meat)*
thuốc *(medicine)*	hạnh_phúc *(happy)*	phụ_huynh *(parents)*	tàu *(ship)*	gà *(chicken)*
đời *(life)*	cuộc_sống *(life)*	bé *(baby)*	xảy *(happen)*	quán *(food stall)*
Topic 15	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**
thi *(compete)*	đón *(wait)*	tỉnh *(province)*	phim *(movie)*	ảnh *(picture)*
giải *(prize)*	tập *(episode)*	dự_án *(project)*	đi *(go)*	vàng *(yellow)*
tham_gia *(participate)*	htv *(stands for Ho Chi Minh City Television)*	xây_dựng *(construction)*	dám *(dare)*	mùa *(season)*
quà *(gift)*	full	tp *(city)*	tốt_đẹp *(nice)*	đi *(go)*
chương_trình *(program)*	hấp_dẫn *(hot)*	ubnd *(stands for People's Committee)*	tặng *(give)*	hoa *(flower)*
câu *(question)*	đừng *(do not)*	quy_định *(rule)*	xe *(car)*	thời *(time)*
việt_nam *(Vietnam)*	chủ_nhật *(sunday)*	thông_tin *(information)*	cướp *(robbery)*	lịch_sử *(history)*
trao *(give)*	chồng *(husband)*	dân *(resident)*	đầu_bếp *(chief)*	hàng *(store)*
giải_thưởng *(award)*	link	đầu_tư *(investment)*	kết_quả *(result)*	phố *(street)*
may_mắn *(lucky)*	gameshow	tổ_chức *(organization)*	chuyên_nghiệp *(professional)*	thu *(autumn)*

the page based on the topic distribution vectors of its documents. The page's topic distribution vector is defined as the weighted sum of the document vectors as described in Sect. 3.4. Thus it has the same dimension of 20 with the document vectors (due to $K = 20$).

As an example, the resulting topic distribution vector of the fan page for "Báo Đời Sống Pháp Luật" (*Law and Life Journal*) is displayed in Fig. 4. As can be observed, the topic probability distribution attains notable peaks at three topics, which are: Topic 2 – a justice-related topic with the keywords such as "cảnh sát" (*police*), "vụ" (*case*), and "điều tra" (*investigate*); Topic 10 – a family-related topic with the keywords such as "mẹ" (*mother*), "vợ" (*wife*), and "tiền" (*money*); Topic 13 – a transportation-related topic with the keywords such as "xe" (*vehicle*), "đường" (*street*), and "giao thông" (*traffic*).

This result is quite reasonable because justice, family, and transportation are the most concerns of this journal.

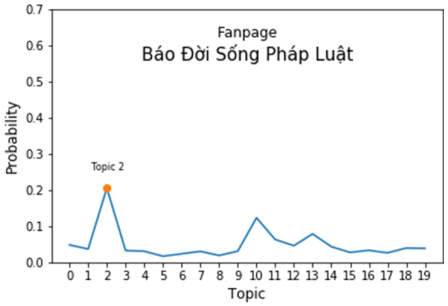

Fig. 4. Topic distribution of fan page "Báo Đời Sống Pháp Luật" (*Law and Life Journal*).

4.5 Fan Page Clustering Results

With the resulting vector representations of fan pages, we can group them into different clusters so that the pages in each cluster have similar topic distributions and the resulting clusters are well separated each other. We have tried to cluster the topic distribution vectors of all fan pages in the dataset with the K-mean Clustering algorithm. With the optimal number of clusters of 12 (see the results in Table 2), we got several example results as follows.

Table 2. Silhouette scores comparison between two methods of fan pages modeling.

Method	# of clusters						
	2	4	6	8	10	12	14
LDA with interaction indices	0.1398	0.1606	0.1896	0.2328	**0.2669**	**0.3008**	**0.2748**
LDA only	**0.1404**	**0.1821**	**0.1965**	**0.2345**	0.2370	0.2759	0.2523
Method	# of clusters						
	16	18	20	22	24	26	28
LDA with interaction indices	**0.2665**	0.2462	**0.2503**	**0.2756**	**0.2580**	**0.2556**	**0.2324**
LDA only	0.2623	**0.2525**	0.2403	0.2279	0.2273	0.2029	0.2322

Cluster 1 includes several fan pages such as "Giải trí TV" *(Entertainment TV)*, "HTV3 - DreamsTV", "Kênh Nhạc Việt" *(Vietnamese Music Channel)*, "VTV Giải trí VTV6" *(VTV Entertainment VTV6)*. All of these are the pages of entertainment channels (Fig. 5).

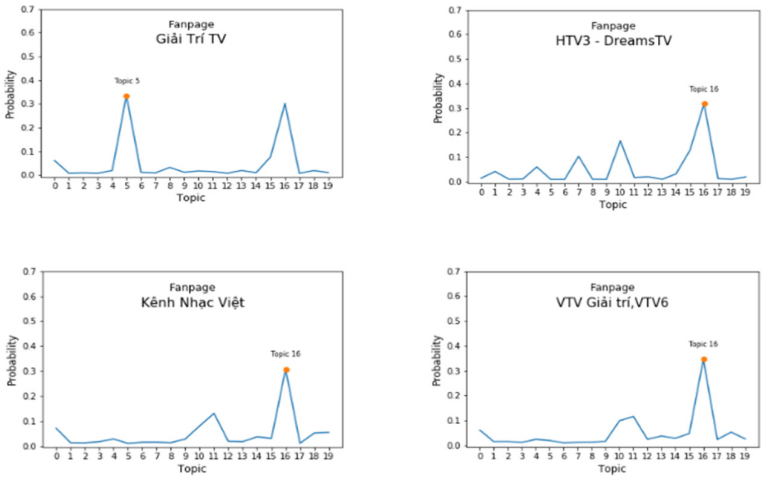

Fig. 5. Topics distribution vectors of the fan pages belonging to Cluster 1.

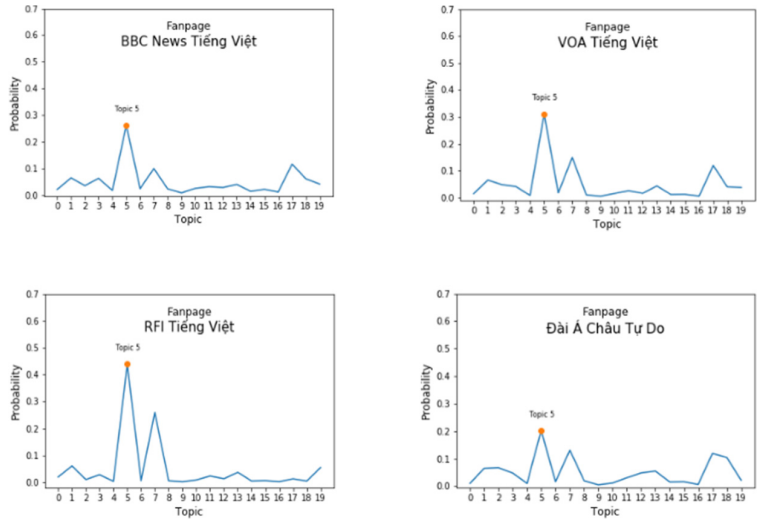

Fig. 6. Topics distribution vectors of the fan pages belonging to Cluster 2.

Cluster 2 includes fan pages of broadcasters about news and politics such as "BBC Tiếng Việt" *(BBC Vietnamese)*, "Đài Châu Á Tự Do" *(Radio Free Asia)*, "RFI Tiếng Việt" *(RFI Vietnamese)*, "VOA Tiếng Việt" *(VOA Vietnamese)* (Fig. 6).

As can be seen on Fig. 5 and Fig. 6, those fan pages having similar topic distributions were grouped quite well thanks to their vector representations.

To quantitatively evaluate the clustering performance, we used Silhouette score [14] to measure how well the clusters are separated to each other. The higher the score, the better clustering process. We compared the clustering performance between the two

vector representations of fan pages: our proposed method (LDA-based topic distributions combined with interaction indices, i.e., each document has a different weight in Eq. (3)) and conventional one (LDA-based topic distributions only, i.e., all documents have the same weight in Eq. (3)). The results in Table 2 show that when the number of clusters is high enough (more than 8), our proposed method outperforms the conventional one on Silhouette score. In particular, both of the two fan page representation methods achieve optimal clustering performance when the number of clusters is set to 12. In that case, our proposed method improves 9.0% on Silhouette score compared to the conventional one (0.3008 vs. 0.2759).

5 Conclusion

In this paper, we have proposed a method to represent a fan page by a vector using LDA-based topic modeling on all fan pages in the corpus combined with interaction index analysis of their posts. Experiment results showed that this representation can be used to cluster a set of fan pages effectively and obtained better clustering performance than the conventional one just based on LDA. The proposed vector representation of fan pages also showed its effectiveness in figuring out hot topics as well as regular issues posted on Facebook in a fixed period of time. The main benefit of our approach to fan page modeling and mining is that it helps us to follow trending contents on this social platform on a large scale without collecting the data of regular individual users. In the future, we will apply other models that focus more on the segmentation of documents such as lda2vec [15] to find out how positive or negative different fan pages talk about the same topic. We also want to extend the proposed method so that the time factor is included to reflect how the relationship between fan pages changes over time.

Acknowledgments. This research is funded by the University of Danang – University of Science and Technology under grant number T2017-02-93.

References

1. Datareportal, "Social Media Users by Platform," 2019. Available: https://datareportal.com/social-media-users. Accessed 19 Nov 2019
2. Yau, C.-K., Porter, A., Newman, N., Suominen, A.: Clustering scientific documents with topic modeling. Scientometrics **100**, 767–786 (2014)
3. Kim, S.-W., Gil, J.-M.: Research paper classification systems based on TF-IDF and LDA schemes. Hum. centric Comput. Inf. Sci. **9**(1), 1–21 (2019). https://doi.org/10.1186/s13673-019-0192-7
4. Pengtao, X., Eric, P.X.: Integrating document clustering and topic modeling. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Virginia, United States, pp. 694–703 (2013)
5. Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I.: A hybrid framework for news clustering based on the DBSCAN-martingale and LDA. MLDM 2016. LNCS (LNAI), vol. 9729, pp. 170–184. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41920-6_13
6. Eliana, S., Camilo, M., Raimundo, A.: Topic modeling of Twitter conversations (2018)

7. Zhou, T., Haiy, Z.: A text mining research based on LDA topic modelling. In: International Conference on Computer Science, Engineering and Information Technology, pp. 201–210 (2016)
8. Jiamthapthaksin, R.: Thai text topic modeling system for discovering group interests of Facebook young adult users. In: 2016 2nd International Conference on Science in Information Technology (ICSITech), pp. 91–96. IEEE 2016
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
10. Darling, W.M.: A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pp. 642–647 (2011)
11. Social Bakers Vietnamese Statistics. Available: https://www.socialbakers.com/statistics/facebook/pages/total/vietnam. Accessed 19 Nov 2019
12. pyvi toolkit. Available: https://pypi.org/project/pyvi/. Accessed 19 Nov 2019
13. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
14. Rousseeuw, P.J.: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
15. Moody, C.E.: Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv arXiv:1605.02019 (2016)