Mining Significant Terminologies in Online Social Media Using Parallelized LDA for the Promotion of Cultural Products*

Richard de Groof[1], Haiping Xu[2], Jurui Zhang[3], and Raymond Liu[4]
[1,2]Computer and Information Science Department
University of Massachusetts Dartmouth, Dartmouth, MA, USA
[3,4]Department of Marketing, University of Massachusetts Boston, Boston, MA, USA

Abstract - *Despite the growing popularity of online social media, there are very few research efforts to use online social media to study market strategies for the promotion of cultural products. With online content being largely unregulated, Latent Dirichlet Allocation (LDA) provides a useful mechanism for organizing textual data and deriving conclusions about the subject matter. In this paper, we introduce a parallelized LDA, called pLDA, to analyze clustered textual data in online social media. We use pLDA to infer the posterior of latent topics over documents and words, and identify significant terminologies that describe the vast number of posts. Making use of sentiment analysis, we are able to further make suggestions about the relevant topics for promoting cultural products. Finally, we use a case study of the music industry to demonstrate how the most relevant aspects to artist popularity can be derived.*

Keywords: Cultural products; online social media; text mining; latent Dirichlet allocation; topic modeling

1 Introduction

The proliferation of online social media technologies has resulted in a tremendous amount of information becoming readily available. Social media websites such as Twitter and Facebook, provide the users an opportunity to speak their mind on pages hosted by everyone from celebrities to artists, from young kids to pop shop owners. Due to its popularity, the potential utility of online social media in promoting cultural products is being increasingly recognized. The social media sites typically require artists such as musicians, to establish an online presence, by which they may disseminate their brand names. Much of the feedback posted online could be useful to identify trends and understand what is important in cultural products such as those produced by musicians. However, mining social media has been a difficult task because so much of the information is in the form of unstructured free text. Previous work has focused on the application of text mining techniques that require manual interpretation. Since there is too much information to provide manual labels, there is a pressing need to develop tools to automatically categorize text and provide meaningful interpretation. To achieve this, it is necessary to identify the subject matter involved. One of the most popular topic modeling techniques is called Latent Dirichlet Allocation (LDA), which is a powerful method for learning topic distributions in text [1]. LDA is an unsupervised topic modeling technique for mining text data and deriving latent topic distributions. Like other unsupervised techniques, it does not require a labeled training set for its operation. This makes it very useful in the context of a large amount of uncategorized data, such as musicians' Facebook pages. However, to make meaningful classifications, a user must inspect the results to determine the associations between hidden topics and documents. In addition, the computational complexity of this methodology may render its use limited for massive documents.

In this paper, we introduce an LDA-based approach to mining significant terminologies in online social media for the promotion of cultural products. Our unique process uses a post-processing step to LDA, which allows us to broadly categorize the posts with less manual intervention. Once we know the subject matter of the posts, we may use Stanford Core NLP to provide sentiment analysis deriving the negative or positive orientation of each. Using meaningful indicators, such as the Facebook numbers of followers, and solving a system of equations characterizing each artist, we may derive the relationship between these topics and artist popularity. Generally, a linear system of equations may indicate relevance of one or more common attributes, calculated sentiment scores of commonly occurring subject matter attributed to each artist. Based on various factors that are important to culture products, artist popularity can be considered as the dependent factor, and brand niche and audience consensus, amongst others, as independent factors using such equations [2]. We focus on the text content of the online social media and use sentiment analysis to determine the orientation and thereby relevance of each independent factor to artist popularity. To improve the efficiency of our approach, we introduce a parallelized LDA procedure, called pLDA, to mine text-based online social media. In a case study of mining online social media for the music industry, we show that our approach can not only effectively identify the most relevant aspects to artist popularity according to the sentiments expressed by the listening audience, but also run faster than its sequential version of the LDA mechanism.

* This material is based upon work supported by the President's 2017 Creative Economy (CE) Fund, University of Massachusetts.

2 Related Work

Researchers have increasingly recognized that online social media offers a breadth of information related to the promotion of cultural products. Goh *et al.* studied the effect of user-generated and marketer-generated content in social media on consumers' repeated purchase behaviors [3]. They used commercial text mining applications to analyze text gathered from Facebook, and found that user-generated content had a more significant impact. Similarly, Kim *et al.* analyzed text reviews of hotels on Trip Advisor to discover satisfiers and dissatisfiers, as well as the reasons why people leave positive or negative reviews [4]. They also leveraged a feature on Trip Advisor, which allows the user to leave a categorical rating. He *et al.* studied the pizza industry using text posts from Facebook and Twitter in an effort to gain marketing insight [5]. Using existing text mining tools, they identified themes in the data that they used to categorically compare three major pizza chains. Unlike the above work, we introduce our novel parallelized text mining approach, and our unique procedure allows the identification of topics to be largely automatic.

There are also a few previous research efforts focused on the use of LDA to analyze online social media. Qiang *et al.* incorporated features such as "geo tracking" to aid in the identification of geographical topics from social media [6]. Their method is based on the LDA model using generation probabilities, which generates each keyword from either a local or a global topic distribution. LDA has also been applied to musical recommendation systems. Kinoshita *et al.* proposed a system to describe musical preferences by considering different tags associated with artist genre and user preferences [7]. They used Collaborative Filtering (CF)-based similar user selection to recommend music products to users with similar tastes to a target artist. Such text mining methodologies have been used in a variety of contexts but often in their original formulation, applying the returned topic distributions directly [8-9]. The returned matrices are typically interpreted as clusters, which represent various combinations of underlying topics. In contrast, our approach derives significant terminologies conditioned on documents, which are highly reflective of the underlying topics. In our approach, we apply the probability of word given document to the identification of latent topics, and consider freely-formed text rather than information derived from tags.

Recent work has discussed the parallelization of the LDA process. Newman *et al.* introduced a parallel process for LDA, which essentially divides the document set into sections and distributes it across computation units [10]. Similarly, Wang *et al.* introduced Plda, which operates in the same fashion but incorporates Hadoop functions [11]. Liu *et al.* introduced Plda+, which works with a pipeline system to perform the same task [12]. A critical aspect of these parallel algorithms is synchronization. In previous work, inference techniques like Collapsed Gibbs Sampling (CGS) were used on partitions of the data sets [10]. After each processor had performed a single iteration, the results were aggregated in a separate step before continuing. Such results are considered an approximation to outputs of the original LDA algorithm. Since the same word may occur across multiple documents, the true results must consider the topic assignment to each in order to attain the true distribution. Different from the above approaches, our method uses a real-time data synchronization mechanism for better accuracy, not only making the results more comparable to a sequential implementation, but also providing a speed-up due to parallelization.

Other related work has focused on the application of information derived from social media towards the music industry. For example, MuSeNet, a network of music artists around the world linked by professional relationships, was developed as an examination of the social network aspect of sites like Facebook, used in the music artist industry [13]. Relationships amongst subscribers may be assembled in a graph to understand underlying phenomena. As such approaches provide useful insights about online social media data, they are complementary to our research efforts on analyzing the text-based social media for the promotion of cultural products.

3 Significant Terminology Identification

Figure 1 shows the procedure by which text data from an online social media such as Facebook pages are processed. Popular and less-known local artists were identified, and review text is extracted using browser scripting mechanisms. Using an online tagging service, such as Last.fm, popular tags are extracted for each artist. Assembling the tags into vectors for each artist, *k*-means clustering can be used to separate the artists into similar categories, i.e., artists with similar proportions of identical tags. We do this for two reasons, to reduce the data size for running pLDA, and also to support our conclusions that different types of artists receive comments reflecting different subject matter and useful information. The outputs of the pLDA process are two matrices, namely θ representing the topic distribution for each document input into pLDA, and φ representing the distribution of words over topics. For pLDA data processing, each post represents a document, which usually discusses a single subject. For those instances in which one subject is discussed, clustering by topic distributions associates comments discussing the same topic together. Also, comments addressing the same multiple topics are also clustered together according to their unique topic distributions.

Once the posts and comments have been extracted and clustered according to artist tags, they are processed using pLDA. Adapted from [14], the matrices θ and φ can be calculated using CGS as in Eq. (1) and (2).

$$\theta_{m,k} = (n_{m,(.)}^{k,-(m,n)} + \alpha_k) \tag{1}$$

$$\phi_{k,n} = \frac{(n_{(.),n}^{k,-(m,n)} + \beta_n)}{\sum_{r=1}^{N}(n_{(.),r}^{k,-(m,n)} + \beta_r)} \tag{2}$$

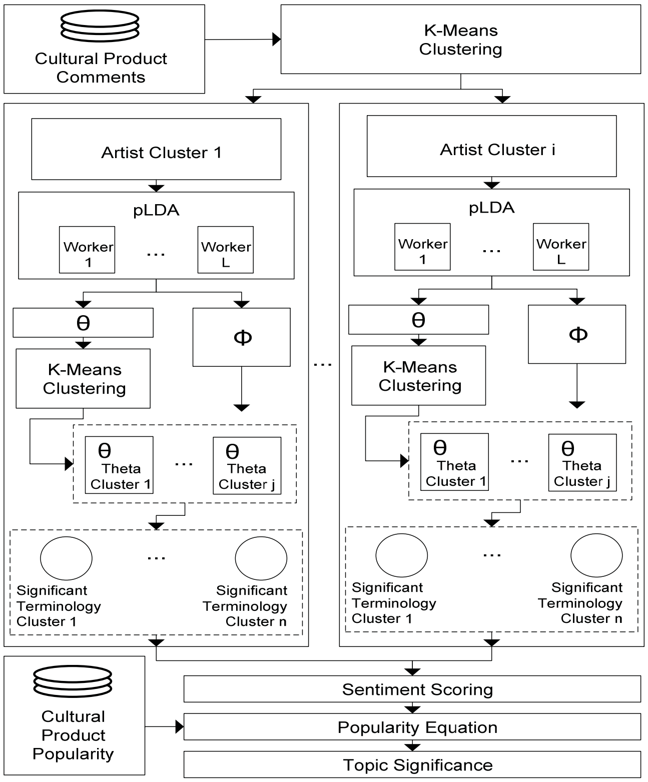

Figure 1. Trends identification in cultural product popularity

The basic idea of Eqs. (1) and (2) shows that LDA defines a probabilistic model, which could be derived from word counts and the hyper-parameter vectors α and β. In the equations, n represents the counts of topics assigned to words and documents. The superscript k represents the topic index, and $-(m,n)$ represents the exclusion of topic for document m and word n as required by CGS. Matrices θ and φ are calculated using document-topic pairs and topic-word pairs, respectively. Each resulting θ from the pLDA process on artist clusters is then input to the k-means algorithm, aggregating into clusters that represent similar distributions of the underlying topics. This step separates the documents into groups representing similar topics. As such, the comments are classified according to content, aiding in the identification of subject matter. These clusters are then multiplied with φ yielding the significant terminologies, which are the basis for classification of comments by topics (the detailed procedure is described in Section 4). The corresponding sentences constituting the documents labeled by topics were then classified according to sentiment scores. The results are inputs of popularity equations, which represent the total sentiment scores by topic accumulating to artist popularity. Assuming that the dependent factor artist popularity relies on the sentiment expressed by all topics, the popularity equations can be solved to derive coefficients representing the topic significance for artist popularity. We then present the results of these simultaneous equations in a case study (described in Section 6) to draw conclusions about what topics are highly relevant in the music industry.

4 Derivation of Significant Terminologies

4.1 Preprocessing of Text Data

Before the online media posts are processed with pLDA, they are clustered according to artist genre tags. As an example shown in Fig. 2, popular tags for each artist were collected from Last.fm, with the artists and the artist clusters displayed inside the boxes. Each tag has a count attributed to it, which represents the number of Last.fm users who have applied that tag. Counts of commonly occurring tags are assembled into a vector for each artist, where each vector is normalized with the summation of its elements to 1. These vectors are then clustered with k-means to produce artist clusters composed of similarly classified musicians according to artist genre tags.

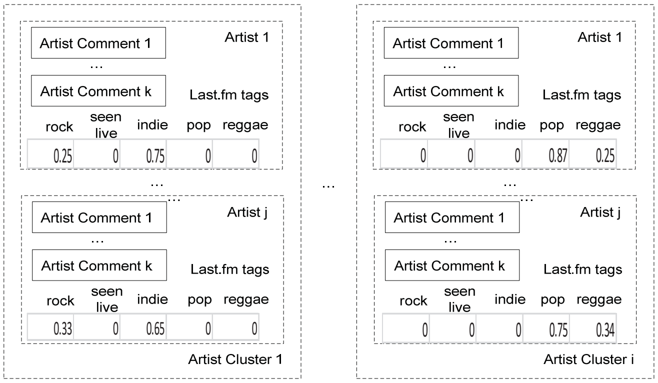

Figure 2. An example of clustering artist comments

4.2 Parallalized LDA

Algorithm 1 shows the pLDA routines, where partitions are made of the data set across documents. Since the documents themselves are disjointed, the only required synchronization among pLDA-Worker processes is the assignments of topics to the shared words in the worker processes. In each iteration, the topic assignments are recorded in matrix *prevAssign*, whose values are used in the next iteration. The shared matrices γ and η represent the counts of topic assignments to words, and the counts of topic assignments to documents, respectively; while *sync* is a semaphore that controls access to γ.

Assume there are L processes of pLDA-Worker that perform the CGS computations in parallel. The algorithm begins by initializing the shared variables θ and φ, and creating the worker processes. Note that to make the algorithm easy to read, we assume the number of documents M is divisible by L. The matrix *prevAssign*, representing the current assignment of topic to document and word, is initialized with random topic assignments. The counts in γ and η are then updated with the topic assignments in *prevAssign*. Barrier synchronization is used to assure that all processes start the sampling processes only after the variables have been initialized (line 6). The matrix *prevAssign* is required by CGS because the current topic is sampled considering the conditional probability of all other assignments except for the current assignment. The algorithm proceeds in a loop across documents, words over iterations until convergence, the point at which modifications to the topic counts stabilize and the results are approaching the true distribution.

Algorithm 1: pLDA Main Process

Shared Variables: γ is a $K \times N$ matrix representing the counts of topic k assigned to word w, where K is the number of topics and N is the number of words; η is an $M \times K$ matrix representing the counts of topic k assigned to document m, where M is the number of documents; *sync* is a semaphore to synchronize writes to γ.

Input: Ψ is an $M \times N$ matrix representing the counts of each word in each document.

Output: θ and φ are $M \times K$ and $K \times N$ matrices that represent the distribution of topics over documents and the distribution of words over topics, respectively.

```
1.  Initialize $\theta$ and $\varphi$ to 0.
2.  process pLDA-Worker[$p$ = 1 to $L$]  // create $L$ worker processes
3.    Let prevAssign be an $M \times N$ matrix representing the topic for
      document $m$ and word $w$.
4.    Initialize prevAssign with random numbers in [1, $K$].
5.    Increment topic assignments from prevAssign in $\gamma$ and $\eta$.
6.    barrier synchronization                  // the lock step
7.    repeat until convergence
8.      first = ($p$-1)*($M$ / $L$) + 1; last = $p$*($M$ / $L$)
9.      for $m$ = first to last
10.       for $w$ = 1 to $N$
11.         preTopic = prevAssign[$m$][$w$]
12.         if $\Psi$[$m$][$w$] > 0
13.           $\eta$[$m$][preTopic]--
14.           P(sync); $\gamma$[preTopic][$w$]--; V(sync)
15.           newTopic = topic assignment to ($m$, $w$) using CGS
16.           $\eta$[$m$][newTopic]++
17.           P(sync); $\gamma$[newTopic][$w$]++; V(sync)
18.           prevAssign[$m$][$w$] = newTopic
19.     barrier synchronization                // the lock step
20. end process
21. Calculate $\theta$ and $\varphi$ as in Eq. (1) and (2) using $\gamma$ and $\eta$.
```

Note that the algorithm selects the next topic assignment using the CGS sampling process (line 15), and increments the counts of topic assigned to both documents and words (line 16-17). At the end of each iteration, the pLDA-Worker processes are synchronized by the barrier synchronization mechanism again to allow all pending updates to occur (line 19). In this way, updates to the common counts of topic assignments to words are continuous, and the results will be more accurate than those from previous implementations. Lastly, θ and φ are calculated as in Eqs. (1) and (2) by summing across topic counts in documents and topic counts in words using η and γ, respectively.

4.3 Significant Terminologies

The complete joint probability of the LDA model can be calculated as in Eq. (3) [15].

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{j=1}^{M} [P(\theta_j; \alpha) \prod_{t=1}^{N} [P(z_{j,t} \mid \theta_j) P(w_{j,t} \mid \phi_{z_{j,t}})]] \tag{3}$$

where W is a set of word assignments to documents, Z is a set of topic assignments to documents and words, M is the number of documents, N is the number of words, and K is the number of topics. The vectors α and β are Dirichlet parameters and the semi-colon indicates that φ and θ are dependent upon them for the calculation.

Due to the complexity of the model, it is not feasible to directly calculate the exact distributions; instead, we first integrate out θ and φ, and using properties of the Dirichlet and multinomial distributions to reduce Eq. (3) to Eq. (4).

$$P(Z_{(m,n)} = k, Z_{-(m,n)}, W; \alpha, \beta) \propto (n_{m,(.)}^{k,-(m,n)} + \alpha_k) \frac{(n_{(.),n}^{k,-(m,n)} + \beta_n)}{\sum_{r=1}^{N} (n_{(.),r}^{k,-(m,n)} + \beta_r)} \tag{4}$$

Eq. (4) serves as the starting point for Gibbs sampling, where n represents counts of topics assigned to documents and words. Gibbs sampling can be used to estimate a distribution using samples from that distribution [14-15]. The sampling process is repeated until it converges when the samples represent the underlying distribution. In CGS, this is performed by removing the measurement of the current value being sampled and making a random draw based on all other stationary values. The notation $-(m, n)$ indicates that the sample for document m and word n has been subtracted from the current sample set.

The joint probability represents that of the current topic and the set of word assignments to documents. With Eq. (4) derived by marginalizing out θ and φ from Eq. (3), the joint probability can be represented as in Eq. (5) [14].

$$P(z \mid d) P(w \mid z) = \frac{P(d \mid z) P(z)}{P(d)} P(w \mid z) \tag{5}$$

An assumption used in deriving the LDA formulation is that of conditional independence between documents and words over topics; thus, we rewrite Eq. (5) as in (6).

$$P(z \mid d) P(w \mid z) = \frac{P(w, d \mid z) P(z)}{P(d)} \tag{6}$$

By the definition of conditional independence, we further rewrite Eq. (6) into (7).

$$P(z \mid d) P(w \mid z) = \frac{P(w, d, z) P(z)}{P(z) P(d)} = \frac{P(w, d, z)}{P(d)} = P(w, z \mid d) \tag{7}$$

Thus the relationship described in Eq. (8) must hold.

$$P(Z_{(m,n)} = k, Z_{-(m,n)}, W) = P(w, z \mid d) = P(z \mid d) P(w \mid z) \tag{8}$$

Definition 1: A *significant terminology* of a corpus is a keyword that represents a major topic of the documents contained in the corpus. A significant terminology, which represents a new metric for the probability of major topics, can be derived from the multiplication of matrices θ and ϕ.

Explanations: The outputs of the LDA process using CGS, i.e., the matrices θ and φ, can be calculated in a single iteration using Eqs. (1) and (2) across documents and topics or topics and words. Therefore, the multiplication of θ and φ can be represented by $P(z|d)P(w|z)$. By the Bayes' rule, θ can be represented equivalently as in Eq. (9).

$$P(z \mid d) = \frac{P(d \mid z)P(z)}{P(d)} \tag{9}$$

By performing standard matrix multiplication of θ and φ, each entry is multiplied and summed across topics, i.e., the matching inner dimension between the matrices. So, the operation proceeds as in Eq. (10).

$$\sum_{k=1}^{K} \frac{P(d \mid z_k)P(z_k)}{P(d)} P(w \mid z_k) = \sum_{k=1}^{K} \frac{P(d,w \mid z_k)P(z_k)}{P(d)} \tag{10}$$

where k is the topic index. Eq. (10) holds because of conditional independence of words and documents over topics. By the definition of conditional independence and the iterations over k, we can derive the result as in Eq. (11).

$$\sum_{k=1}^{K} \frac{P(d,w,z_k)P(z_k)}{P(z_k)P(d)} = \sum_{k=1}^{K} \frac{P(d,w,z_k)}{P(d)} = \frac{P(d,w)}{P(d)} = P(w \mid d) \tag{11}$$

We interpret the conditional probability $P(w|d)$ as the probability of word significance given documents. It is the probability of words according to hidden topics, which are prominent in their respective documents.

5 System of Popularity Equations

The significant terminology clusters are defined as in Eq. (12). According to Definition 1, the multiplication of θ and φ yields the significant terminologies. Once the results of pLDA have been clustered using k-means, we have groups of documents, representing similar underlying topic distributions. Each theta cluster is multiplied with φ from that batch of pLDA using standard matrix multiplication. This operation iterates over topics consistent with the procedure outlined in Eq. (10-11).

$$\theta \times \varphi = \left\{ \begin{array}{c} \begin{bmatrix} \theta_{(1,1)} & \cdots & \theta_{(1,K)} \\ \vdots & \ddots & \vdots \\ \theta_{(M,1)} & \cdots & \theta_{(M,K)} \end{bmatrix} \\ \mathit{Theta\_Cluster\_1} \\ \vdots \\ \begin{bmatrix} \theta_{(1,1)} & \cdots & \theta_{(1,K)} \\ \vdots & \ddots & \vdots \\ \theta_{(M,1)} & \cdots & \theta_{(M,K)} \end{bmatrix} \\ \mathit{Theta\_Cluster\_N} \end{array} \right. \times \begin{array}{c} \begin{bmatrix} \varphi_{(1,1)} & \cdots & \varphi_{(1,V)} \\ \vdots & \ddots & \vdots \\ \varphi_{(K,1)} & \cdots & \varphi_{(K,V)} \end{bmatrix} \\ \mathit{Topics\,/\,Words} \end{array} \tag{12}$$

Once the significant terminologies have been calculated and the documents have been clustered, we can derive patterns in the text. For example, by reviewing the posts for musicians, we may find that they tend to fall broadly into three categories: descriptions of live shows, descriptions and recommendations of new albums, and discussion of streaming services on which the artists appear. The clusters are readily identified by keywords occurring throughout the clustered documents according to word significances. The probabilities of word significances for the various words in a cluster are summed across the set of documents to determine the greatest probabilities. The top occurring words as measured by word significance are the significant terminologies for the cluster.

Figure 3 shows an example of methods for assigning topics to comments by significant terminologies in the domain of the music industry. As shown in the figure, a significant terminology for the live shows that appears consistently is "show". The word "album" appears in that context of album. The words "spotify", "video" and "bandcamp" appear in the streaming category.

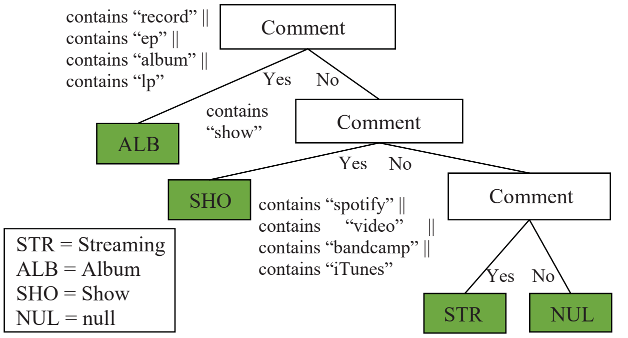

Figure 3. Classification by occurrence of significant terminologies

Using this approach, the clusters can be identified readily by the occurrence or non-occurrence of the significant terminologies. As such, it greatly simplifies the classification process, and hundreds of comments may fall into one category or another quickly and accurately by the inclusion or exclusion of the significant terminologies.

Furthermore, Stanford Core NLP can be used to classify sentiment on a positive/negative spectrum ranging between 0 and 4, with 0 being the most negative and 4 being the most positive. The identified comments for artists are classified by sentiment, and the results summed across musicians cluster and the identified topic as in Eq. (13).

$$\begin{aligned} streaming_i &= \sum_j SentScore\left(Comment_{i,j}^{streaming}\right) \\ album_i &= \sum_j SentScore\left(Comment_{i,j}^{album}\right) \\ shows_i &= \sum_j SentScore\left(Comment_{i,j}^{shows}\right) \end{aligned} \tag{13}$$

where the subscript i indicates that this is the i-th artist. The superscript over comment indicates that the comment has been labeled as that topic. The subscripts i and j under comment indicate that this is the j-th comment belonging to the i-th artist labeled by the superscript. The method *SentScore* is a Stanford Core NLP process that returns the corresponding sentiment score value. Once the sentiment scores are summed, they are averaged by the number of comments attributed to that topic. This is to provide a relative value, as a large number of comments may artificially inflate the value, while an average would better represent the overall score.

As part of the data collection process, the number of followers of each artist posted on their social media pages are retrieved and, with the sentiment scores, assembled in a system of linear equations as shown in Eqs. (14):

$$\begin{aligned} ArtistPopularity_i &= \alpha_k streaming_i + \alpha_l album_i + \alpha_m shows_i \\ &\ldots \\ ArtistPopularity_j &= \alpha_k streaming_j + \alpha_l album_j + \alpha_m shows_j \end{aligned} \qquad (14)$$

This is a linear system with an equation for each artist *i* through *j*. The artist popularity is defined as the number of social media followers (e.g., Facebook followers) in our case study (Section 6). The variables in Eqs. (14), such as $streaming_i$, $album_i$ and $show_i$ are independent ones. We use a process like the least squares to solve for coefficient vector α, one for each independent variable. Weighted least squares with the weights being the number of comments for each artist may also be used to solve this system and could be a better option [6]. Note that regular least squares is not a suitable choice because the error occurring between musicians in each category may vary according to the number of comments, as LDA tends to be more accurate with a larger amount of data.

6 Case Study

In order to draw conclusions about the music industry, we collected 95,265 comments from 879 artists featured on WUMB, a radio station of University of Massachusetts Boston that broadcasts an Americana/Blues/Roots/Folk mix, as well as from a list of Boston, Massachusetts local artists featured on thedelimagazine.com, a daily updated website covering 11 North American music scenes through 12 dedicated, separate blogs. Following the procedure outlined in Section 4, we classified the comments, produced the significant terminologies, and derived the topic significance with regards to artist popularity. Amongst the different types of music, different factors appear to contribute to artist success. Some bands are known more for live performances, and tend to promote and discuss these on their Facebook pages. Other bands are usually promoting a new album when they post to online social media.

6.1 Artist Cluster Level Analysis

Figure 4 shows 4 artist clusters from both the WUMB and thedelimagazine.com lists. Different types of artists have higher correlations with one topic or another. The WUMB artist cluster representing blues and folk artists have a higher correlation with the streaming services and less correlation with album. This indicates that these artists are not using online media to promote their albums sold in record stores to the same degree that they are promoting streaming services, which provides online access to their music. Similarly, the thedelimagazine.com artist group classified as hard rock, indie follows the same pattern. WUMB artist cluster folk, singer-songwriter, however, does not have any correlation with streaming. These artists mostly promote new albums in the conventional fashion. In addition, the thedelimagazine.com artist cluster rock, indie mostly promotes shows but has no mention of streaming.

Moving forward to analyze a particular type of music, in Fig. 5, we show two artist clusters with topic significance, which are both described as folk by Last.fm, and featured on the WUMB and thedelimagazine.com lists. The local folk scene found on thedelimagazine.com cites online sites more frequently and has a higher streaming coefficient. They are more likely to mention singles published on iTunes, their channel on Spotify or a feature on bandcamp, a site which promotes musical artists and has been gaining popularity in recent years. To promote popularity, it is worthwhile for artists at WUMB to fall in this category taking advantage of the streaming services via online social media. This is evidenced by the high coefficients for the WUMB blues and folk cluster shown in Fig. 4, but with extremely low streaming coefficient for folk artists from WUMB as shown in Fig. 5.

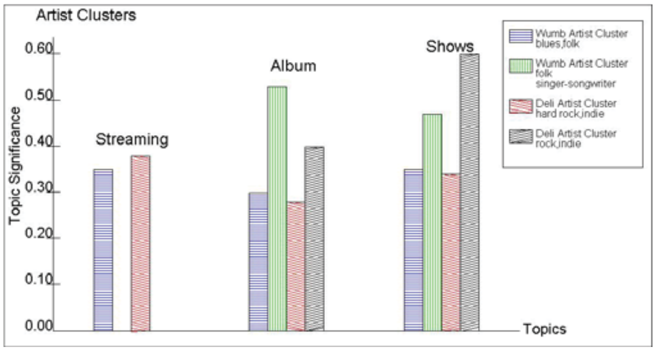

Figure 4. Artist cluster level topic significance by popularity

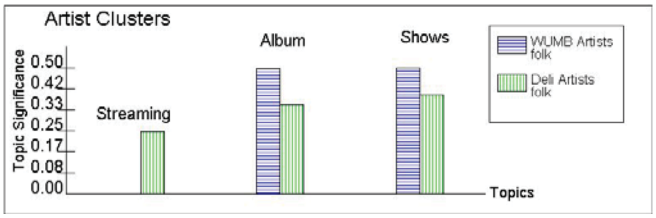

Figure 5. Artist clusters with topic significance for folk music

6.2 Artist Level Analysis

Table 1 shows three artists drawn from three different artist clusters. Artist X belongs to the WUMB folk cluster as shown in Fig. 5, and artist Y and Z belong to the WUMB artist cluster "blues, folk" and the thedelimagazine.com artist cluster "hard rock, indie", respectively, both shown in Fig. 4. With a comparable number of comments, however, artist X is less than half as popular as Artist Y. Similarly, artist Z is more than twice as popular as artist Y. This is majorly due to their streaming scores, which are the cumulative sentiment scores for artists in the streaming category. For example, in the case of artist Y and Z, the streaming scores increase at a rate greater than double for artist Z, despite there being fewer than double the comments between artist Y and Z.

Table 1. Comparison of artists from different clusters

Artist ID	*#Followers*	*Streaming Score*	*#Comments*
Artist X	4334	0	140
Artist Y	9777	34	151
Artist Z	23998	85	270

6.3 Efficiency Analysis

Figure 6 shows the running time of both the sequential version of the LDA implemented via CGS and the pLDA implemented via CGS using 5 parallel threads with increasing numbers of iterations. Compared with the sequential CGS, not only does the pLDA consistently run faster, its running time does not increase linearly, but increases with a decreasing rate. This is because the overhead time for initialization, partitioning the dataset and synchronization is constant. Thus, with more and more iterations going on, the impact of overhead time becomes less and less significant.

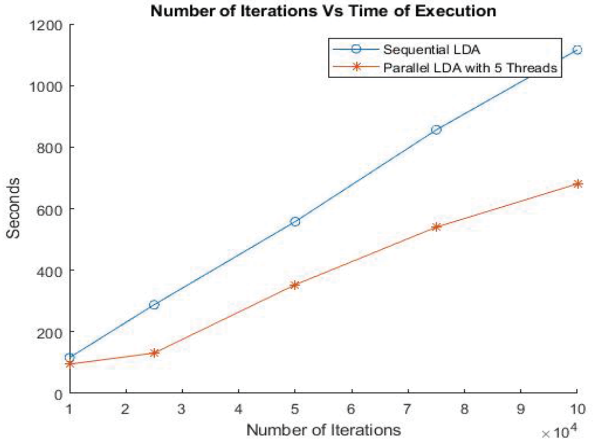

Figure 6. The running time of parallel and sequential LDA

7 Conclusions and future work

Our methodology of deriving significant terminologies allows LDA to be applied to the task of topic categorization quickly and accurately. Our approach is more automated than traditional usages of LDA approach, which require manual interpretation of the document clusters, as opposed to the use of significant terminologies that tend to be more reflective of underlying topics. As a potential additional step to our approach, the clusters of significant terminologies can be well identified using a supervised approach, like decision trees. The inclusion of significant terminologies lends itself well to a tree structure, and may make the process of identifying clusters more accurate.

Using our approach, we have found that, in the folk music scene and elsewhere, online streaming services are highly relevant to artist popularity. These artists are turning to such new venues instead of the conventional approach of promoting albums through record labels. The results are particularly relevant to the WUMB radio station, as their artists are mostly categorized as folk music, where we observed how they use online sources for promotion.

For future work on the promotion of cultural products, we would like to develop a more deterministic approach in LDA, which boasts the same efficiency as stochastic approaches like CGS. We also plan to explore social networks using graph theory to better understand how the demographics may impact on artist popularity.

8 References

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, January 2003, pp. 993-1022.

[2] Z. Jurui and R. Liu. "Popularity of Digital Products in Online Social Tagging Systems," *Journal of Brand Management,* Vol. 24, No. 1, January 2017, pp. 105-127.

[3] K. Goh, C. Heng and Z. Lin, "Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User-and-Marketer-Generated Content," *Information Systems Research*, Vol. 24, No. 1, March 2013, pp. 88-107.

[4] B. Kim, S. Kim and C. Y. Heo, "Analysis of Satisfiers and Dissatisfiers in Online Hotel Reveiws on Social Media," *International Journal of Contemporary Hospitality Management*, Vol. 28, No. 9, September 2016, pp. 1915-1936.

[5] W. He, S. Zha and L. Li, "Social Media Competitive Analysis and Text Mining: a Case Study in the Pizza Industry," *International Journal of Information Management*, Vol. 33, No. 3, June 2013, pp. 464-472.

[6] S. Qiang, Y. Wang and Y. Jin, "A Local-Global LDA Model for Discovering Geographical Topics from Social Media," *Proceedings of the Asia-Pacific Web (APWeb) and Web-age Information Management (WAIM) Joint Conference on Web and Big Data,* Beijing, China, July 2017, pp. 27-40.

[7] S. Kinoshita, T. Ogawa and M. Haseyama, "LDA-Based Music Recommendation with CF-based Similar User Selection," *Proceedings of the IEEE 4th Global Conference on Consumer Electronics,* Osaka City, Japan, October 2015, pp. 215-216.

[8] Y. Xu and Z. Xianli, "A LDA Model Based Text-Mining Method to Recommend Reviewer for Proposal of Research Project Selection," *Proceedings of the IEEE 13th International Conference on Service Systems and Service Management (ICSSSM)*, Piscataway, NJ, USA, June 2016*,* pp. 1-5.

[9] M. Wu, Z. C. Dong, L. Weiyao and W. Q. Qiang. "Text Topic Mining Based on LDA and Co-Occurrence Theory," *Proceeding of the 2012 7th International Conference on Computer Science & Education*, Melbourne, Australia, July 2012, pp. 525-528.

[10] D. Newman, A. Asunction, P. Smyth and M. Welling, "Distributed Inference for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems,* 2008, pp.1081-1088.

[11] Y. Wang, H. Bai, M. Stanton, W. Chen and E. Y. Chang, "Plda: Parallel Latent Dirichlet Allocation for Large-Scale Applications," *Proceedings of the International Conference on Algorithmic Applications in Management (ICAAM),* Berlin, Heidelberg, June 2009, pp. 301-314.

[12] Z. Liu, Y. Zhang and E. Y. Chang, "Plda+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing," *ACM Transactions on Intelligent Systems and Technology (TIST),* Vol. 2, No. 3, April 2011, pp. 26-44.

[13] B. Gabriel, A. Topriceanu, and M. Udrescu. "MuSeNet: Natural Patterns in the Music Artists Industry," *Proceedings of the IEEE 9th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, Romania, May 15-17, 2014, pp. 317-322.

[14] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth and M. Welling, "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation," *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Las Vegas, NV, USA, August 2008, pp. 569-577.

[15] R. de Groof and H. Xu, "Automatic Topic Discovery of Online Hospital Reviews Using an Improved LDA with Variational Gibbs Sampling," *Proceedings of the 2017 IEEE International Conference on Big Data (IEEE BigData 2017)*, Boston, MA, USA, December 11-14, 2017, pp. 3940-3947.