Jiang, Yuting; Fu, Mengyao; Fang, Jie; Rossi, Matti

Applying topic modeling with prior domain-knowledge in information systems research

Published in:
PACIS 2023 Proceedings

Published: 01/07/2023

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Jiang, Y., Fu, M., Fang, J., & Rossi, M. (2023). Applying topic modeling with prior domain-knowledge in information systems research. In *PACIS 2023 Proceedings* Article 1582 (Pacific Asia Conference on Information Systems). Association for Information Systems. https://aisel.aisnet.org/pacis2023/143/

7-8-2023

Applying Topic Modeling with Prior Domain-Knowledge in Information Systems Research

Yuting Jiang
Department of Information and Service Economy, Aalto University School of Business, yuting.jiang@aalto.fi

Mengyao Fu
The Laboratory for AI-Powered Financial Technologies Limited, mengyaofu@hkaift.com

Jie Fang
University of Nottingham Ningbo China, jie.fang@nottingham.edu.cn

Matti Rossi
Information and Service Economy, matti.rossi@aalto.fi

Applying Topic Modeling with Prior Domain-Knowledge in Information System Research

Completed Research Paper

Yuting Jiang
Aalto University
Helsinki, Finland
Yuting.jiang@aalto.fi

Mengyao Fu
Laboratory for AI-Powered Financial Technologies Limited
Hong Kong, China
mengyaofu@hkaift.com

Jie Fang
University of Nottingham Ningbo China
Ningbo, Zhejiang, China
jie.fang@nottingham.edu.cn

Matti Rossi
Aalto University
Helsinki, Finland
matti.rossi@aalto.fi

Abstract

Topic modeling is gaining traction in Information Systems (IS) research as more textual material becomes accessible online and computational tools for analyzing large textual datasets are getting more powerful. This paper advances a new two-step correlation explanation topic modeling (Corex) method with prior domain knowledge to improve the interpretability of topic modeling to meet the needs of current IS research. The proposed method combines the traditional Latent Dirichlet Allocation topic model and the Anchored correlation explanation topic model. In the first step, the approach allows for the rapid and maximum acquisition of topic words related to domain knowledge. These anchor words are then inputted into the second-step CorEx topic model. We further applied and verified the effectiveness of the two-step Corex method to a textual dataset containing 4,290,484 users' personal profiles, thereby illustrating the utility of applying this innovative topic-modeling method in information systems research.

Keywords: Topic modeling, Latent Dirichlet Allocation, Anchored Correlation Explanation, two-step CorEx topic model

Introduction

The rise of social media and content generation platforms has increased the amount of user-generated text data available for analysis, such as vast amounts of tweets posted by users on social media (Twitter, 2023), large number of comments posted by consumers on various service or product consumption platforms such as Amazon.com (Jeong, 2023) and TripAdvisor.com (Olorunsola et al., 2023), and descriptions of service products. These textual data bring more opportunities for information systems (IS) researchers, and massive text data make IS researchers rely more and more on automatic text analysis methods, especially topic modeling. This method aims to determine the structure of the underlying document collection, making it possible for scholars to understand the large amount of text content in information systems. Examples available are for example retrieving aspects from online comments (Jabr et al., 2018b) or identifying opinions from social media (Mukkamala & Beck, 2018). Mining these text data can help researchers explore the hidden content in the text data, including people's emotions, opinions, and attitudes on digital platforms. In addition, it helps people explore how these factors relate to each other and why they behave in certain ways, which helps to identify factors or structures that can be further used in developing theoretical models of information systems (Kar & Dwivedi, 2020). Although massive text data brings more possibilities to IS research, it also brings significant challenges.

As the source and quantity of text data become more and more abundant, the method of extracting text information through manual coding has become more and more challenging to meet the current demand (Lee et al., 2010). It has become an important topic in information system research to use text mining method to automatically extract the characteristics of text data. Topic models such as LDA and LSA, as common text mining techniques, have become increasingly popular because they can quickly obtain the topic overview of the content involved in a large amount of text information (Maier et al., 2018), have very mature implementation tools and evaluation methods (Debortoli et al., 2016), and can quickly identify potential semantic connections, containing multiple topics (Grimmer & Stewart, 2013). Especially in IS research, many studies have applied it to understand the representative and significant topics behind the massive amount of text information. For example, extracting different aspects of the service from the online service comments, identifying the potential topic mentioned in the disaster-related tweets, etc. Based on our review of extant literature, we discovered that the purpose of using topic modeling methods in most IS research is to explore the impact of digital content and the role of topic modeling is largely descriptive in nature. Under such circumstances, the topic model is required to have high interpretability (Debortoli et al., 2016), which in turn calls for a need to bolster the interpretability of topic modeling.

However, unsupervised topic models such as LDA, commonly used at present, show some weaknesses in model interpretability. Specifically, the first problem is topic overlap (Boschetti, 2015). That is, different topics have the same topic words, which makes it difficult for researchers to summarize each topic accurately, and the interpretation of the topic model is weak. Secondly, there may be some topics that cannot be summarized or are challenging to explain (Debortoli et al., 2016), thus weakening the interpretability of the topic model. Third, an unsupervised topic model may not generate the topics we expect to occur because the distribution characteristics of vocabulary cannot be controlled. However, if some uncommon words are deleted directly from the text analyzed, or the weight of some keywords related to the relevant topic is changed, the quality of other topics may be damaged (Debortoli et al., 2016; Denny & Spirling, 2018). In addition to the issues of interpretability, the current generic topic model cannot generate topics driven by domain knowledge. IS researchers encourage the topic model to identify different aspects of the service in the comment (Titov & McDonald, 2008), such as different aspects of the service in the comment and different aspects of the tweet related to the disaster. The expected topic modeling results of these studies are required to have strong domain characteristics. For example, in our example, the former requires the output of service-related topics, while the latter requires the output of disaster-related topics.

In this context, we focus on providing an improved CorEx method towards the above challenges that IS researchers working on text analysis face. To solve the limitations of the current topic models in IS research, this paper proposes a two-step CorEx topic model with prior domain knowledge applied to IS research. We developed a new method to deal with this significant problem:

Research Question: How does the two-stage CorEx approach improve interpretability and achieve domain-knowledge driving?

This article focuses on a correlation explanation topic modeling method called CorEx that can efficiently run as a semi-supervised model. In this model, the probability distribution of topic words is improved by providing some anchor words that are notified before the model (Gallagher et al., 2017). CorEx's anchored words are typically generated using the automated method proposed by Jagarlamudi et al.(2012), which identifies words with the highest interactive information with tags. However, considering that the generation of anchor words cannot integrate domain knowledge comprehensively, we improve the method of generating anchor words. Specifically, we first use the traditional LDA model for unsupervised topic modeling, and then based on the generated topic words, researchers can spontaneously summarize different categories and corresponding anchor words based on domain knowledge. On this basis, CorEx correlation explanation topics are modeled to generate topics. This two-step CorEx topic model can effectively construct a topic model embedded with prior domain knowledge. As a result, this two-step CorEx topic model can solve the interpretability problem of the current topic model for the IS domain and generate domain knowledge-driven topics. We apply this two-step CorEx topic model to an online user-generated self-description corpus to generate topics with prior domain knowledge. Finally, we evaluate the semantic consistency and validity of this model carefully.

The structure of this paper is as follows. In the second section, we summarize the methodology and application of topic modeling in the relevant IS research and discuss the limitations of these studies. In Section 3, we describe the source and structure of the data set used in this article and our approach to the two-step CorEx topic model. Next, we show the application of this approach to a user-generated data set obtained from Zhihu.com and examine the semantic coherence and total correlation score of the topic model. Then, we summarized the results of the research. On this basis, we discuss the significance of this study. Finally, we discuss the limitations of this paper and possible future research directions based on this study.

Research Background

The probabilistic topic model is a general textual analysis method in IS research. This unsupervised learning method has become popular in IS research because it only relies on a few assumptions and has the lowest cost of data analysis (Debortoli et al., 2016). In addition, this unsupervised learning method has many mature software libraries in Python and R (Debortoli et al., 2016), so it is very convenient for IS researchers to implement without requiring additional NLP research background. The basic idea of the unsupervised learning method comes from the assumption of text distribution (Harris, 1954), and the most commonly used text distribution methods in IS research include Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) (Debortoli et al., 2016).

LSA is used to identify the topic-based semantic relationship between text and words through matrix decomposition, but the factor load calculated by LSA has no clear interpretation, which results in low interpretability (Debortoli et al., 2016). Then, LDA methods were developed to improve this issue (Blei et al., 2003). In this method, the topic of each document is given in the form of a probability distribution. However, the LDA approach is not perfect. One of its major drawbacks is topic overlap (Boschetti, 2015), meaning that the same word can be found on multiple topics. Therefore, the topics generated by LDA are not independent and orthogonal, which weakens the interpretability of the topics. Another major fundamental problem with LDA is the inefficiency of LDA on complex text. When analyzing very messy textual data, such as video bullet screen comments and personalized self-description texts, LDA results are highly inaccurate and unreproducible (Lancichinetti et al., 2015).

We refer to the literature on topic modeling in management information system research summarized by Eickhoff and Neuss (2017) and focus on the methodology and application of topic modeling in IS research.

Methodological Research on Topic Modeling within Information Systems

In this section, we summarize extant literature on topic modeling in IS research and list their research modes and specific topic modeling methods in Table 1. Based on an analysis of this literature, we observe that there is not much research on topic modeling methodology in IS research. Most topic model research discusses probabilistic models such as LSA and LDA, which are mainstream topic modeling models in IS research. As we discussed earlier, an important reason is that these models already have easy-to-use software or mature model libraries (Debortoli et al., 2016) that are easy for IS researchers to implement. However, the traditional probability model has become increasingly challenging to meet the current demand.

Key Contribution	Research Approach	Topic Modeling Method
Overview of LSI/LSA (Dumais, 2004)	Model	LSA/LSA
Two hierarchical agglomerative clustering (HAC) techniques (Wei et al., 2006)	Model	HAC
Comparison of structured priors for LDA (Wallach et al., 2009)	Comparative	LDA
Comparison of four topic modeling methods (Lee et al., 2010)	Comparative	LSA/PLSA/LDA/CTM
Labeled LDA for tweet user characteristics (Ramage et al., 2010)	Model	LDA

Methodological recommendations for LSA studies (Evangelopoulos et al., 2012)	Model	LSA
Interval semi-supervised topic model (ISLDA) and coherence metric (Nikolenko et al., 2017)	Model & Validation	ISLDA
Deconstruct reviews into the aspects to evaluate the temporal evolution of user satisfaction with these aspects at a granular level (Jabr et al., 2018b)	Model	LDA
Using topic labeled gold-standard sets to evaluate topic modeling interpretability (Palese & Piccoli, 2020)	Validation	LDA
Using an semi-automated approach to perform systematic literature reviews (Denzler et al., 2021)	Model	LDA

Table 1. Methodological Research in Information Systems

To enhance the interpretability of topic modeling, avoid topic overlap, and generate anchor word-driven topics, IS scholars have tried to optimize the traditional probabilistic topic model in different ways. For example, Nikolenko et al. (2017) proposed an interval semi-supervised LDA topic model to dig specific topics in qualitative research and guide specific words to be sown to a given topic by fixing the Z-values of some keywords related to relevant topics. Nevertheless, one problem with this setup is that the choice of anchor words is complicated because it takes extra effort to determine when words with different meanings appear. Andrzejewski et al. (2009) proposed using logical constraints to enforce separability between topics, but this approach still does not solve the problem of how to anchor vocabulary. In addition, Jabr et al. (2018) proposed some improved LDA algorithms to identify various aspects of product reviews. This approach seems to improve the interpretability of the topic model, but this approach requires additional deletion and lexical extraction of the original corpus, which may affect the quality of the generated topics.

Application Research on Topic Modeling within Information Systems

Second, we summarize the literature on applying topic modeling in IS research. We list their specific application scenarios and topic modeling methods in Table 2. According to our summary of the applied research literature, it can be found that the unsupervised probabilistic topic model is the most commonly used method in IS, such as LDA. The purpose of using the probabilistic topic model is mainly to extract the different aspects involved in the text. We can find a common feature of topic models in these research, which all need to extract topics from the corpora of specific events or domains. For example, for the text content of an online product, the various aspects related to the online product need to be extracted (Titov & McDonald, 2008). For hotel services, it is required to extract all aspects related to hotel services (Io & Lee, 2020). Moreover, it is required to extract the disaster-related aspects of disaster-related tweets (Mukkamala & Beck, 2018). And these aspects or anchor words representing the underlying topic in the corpus of events or domains are called prior domain knowledge. Incorporating domain knowledge into the basic topic model can be used to investigate embedded topics in textual data from different domains. This method directs the topic model, making it easy to explore how the text relates to a particular domain, such as disaster or service.

Description	Purpose of Topic Modeling	Topic Modeling Method
Extract aspects from product reviews using Multi-Grain LDA (Titov & McDonald, 2008)	Extract review aspects	MG-LDA
Review helpfulness and emotions shown in review (Ahmad & Laroche, 2015)	Measure emotions	LSA
Call for use of LDA for theory generation (Rai, 2016)	Theory generation	LDA
Extract different aspects of satisfaction from reviews on shopping websites (Jabr et al., 2018b)	Extract review aspects	LDA

Understand if Amazon reviews discuss different product aspects at different points in time. (Jabr et al., 2018a)	Extract review aspects	LDA
Identify topics of natural disaster tweets on social media and manually summarize them (Mukkamala & Beck, 2018)	Identify underlying topics	LDA
Analyze reactions to the news that robots are replacing human services from hotel reviews (Io & Lee, 2020)	Extract perceptions of robots	LDA
Systematic literature review using semi-automatic methods (Denzler et al., 2021)	Identify underlying topics	LDA
Examine the determinants of the Top 500 companies (Banker et al., 2022)	Interpret factors	-

Table 2. Applied Topic Modeling in Information Systems Research

We believe that there is still space for a new topic model approach. First, considering the difficulty of the current method to obtain anchor words, IS researchers need a new topic model approach to easily identify anchor words and anchor these words to various topics, thus strengthening the separation between topics and improving the interpretability of topic models. IS research usually uses unsupervised probabilistic topic modeling to extract topics. However, considering the coarse granularity of unsupervised topic classification, it may not be able to generate topics that researchers want to appear and may generate topics that are difficult to explain.

Second, to meet the needs of topic extraction from texts in different fields in IS research, this new topic model should be embedded with prior domain knowledge. Semi-supervised topic modeling, which is modeled by correlation interpretation, can effectively solve this problem. Correlation Explanation does not make any structural a priori assumptions about the data generation but instead is similar to information gain (Steeg & Galstyan, 2014), using the difference between total correlations to find the topic that best explains the correlation of the data.

Development of New Topic Modeling Approach

In this section we develop the idea of the new topic modeling approach named two-step CorEx topic model.

Two-Step CorEx Topic Modeling

We propose a two-step CorEx topic model to identify aspects of personal information disclosed in text descriptions. Specifically, by anchoring vocabulary and information bottlenecks (Gallagher et al., 2017), a highly interpretive distribution of topics is obtained. This study develops a two-step CorEx topic modeling followed by semi-supervised topic modeling. Step 1, we obtain multiple topic words through traditional LDA topic modeling. LDA is used to find anchor words due to its efficiency in the fast identification of hidden topics and their distribution from many documents. Therefore, this method can maximize and quickly acquire merged anchor words, such as hobby, game, and run music (examples are provided in Table 5), which are representatives of underlying topics in the corpus and therefore also known as domain knowledge. Researchers then determine topic classifications based on their domain knowledge and research objectives, map these anchor words into corresponding categories, and input them into the Correlation explanation semi-supervised topic model in step 2, and then output the anchored word-driven topic. The graphical structure of the two-step CorEx topic model is shown in Figure 1.

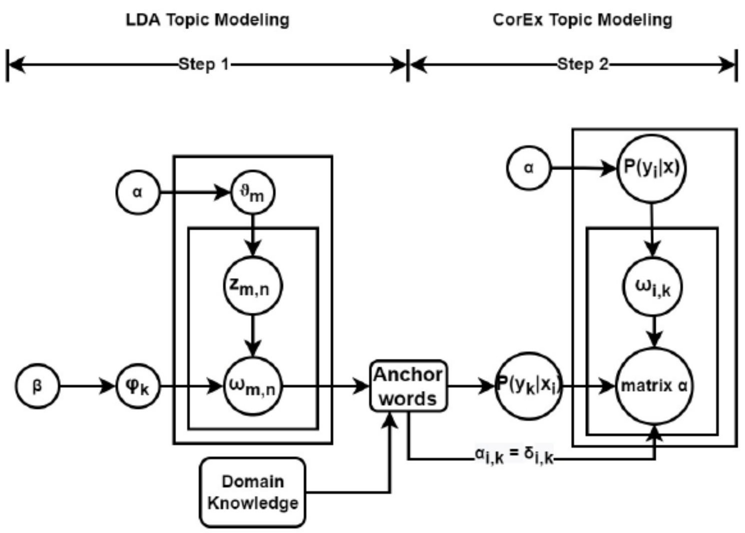


Figure 1. Two-Step CorEx Topic Model

We use the LDA model to generate words that may contain domain knowledge. LDA applies probabilistic cryptic meaning analysis based on the Bayesian inference mechanism (Hofmann, 1999) and Dirichlet distribution is a probability distribution of multiple continuous random variables. Using LDA for topic analysis is to infer all model parameters by estimating a posterior probability distribution for a given text set to learn the topic distribution of each text and the word distribution of each topic. LDA believes that each topic will correspond to a lexical distribution, and each document will correspond to a topic distribution. LDA can give each topic in the form of a probability distribution to conduct topic clustering or text classification by analyzing the distribution of topics in the document. For example, if a given document contains keywords such as "basketball," "football," "tennis," and "Biden," there is a 3/4 probability that the document is "sports" and a 1/4 probability that it is "politics," so the LDA model would mark the document as "sports."

Next, we discussed how the LDA works. Suppose God had two POTS of dice, one containing topic-word dice and the second containing doc-topic dice. LDA topic modeling involves two steps. In the first step, for each document m ∈ {1..., M}, generates a topic distribution vector $\vartheta_m \sim$ Dirichlet (α). Specifically, God randomly draws a doc-topic dice from the first jar and rolls the doc-topic dice to generate the topic number Z_{mn} for the N_{th} word in the document.

The Dirichlet prior distributions of the hyperparameters α is as follows (Bao et al., 2023):

$$\mathrm{P}(\vartheta|\alpha) = \frac{\Gamma(\sum_m \alpha_m)}{\prod_m \Gamma(\alpha_m)} \prod\nolimits_m \vartheta^{\alpha_m - 1} \tag{1}$$

Then, for each topic k ∈ {1..., K}, generates a word distribution vector $\varphi_k \sim$ Dirichlet (β). Specifically, God randomly draws k topic-word dice φ_k independently from the second jar. Then choose k topic-word dice numbered Z_{mn} from φ_k to roll, thus generating vocabulary ω_{mn}.

The Dirichlet prior distributions of the hyperparameters β is as follows (Bao et al., 2023):

$$\mathrm{P}(\varphi|\beta) = \frac{\Gamma(\sum_k \beta_k)}{\prod_k \Gamma(\beta_k)} \prod\nolimits_k \varphi^{\beta_k - 1} \tag{2}$$

As a result, we extract words that represent or are related to domain knowledge from the set of topic words generated by the LDA topic modeling and then anchor these selected topic words to multiple topics that we expect to occur. For example, we can anchor "doctor," "teacher," and "programmer" to a topic. Anchor "medicine," "computer science," and "economics" to another topic to help separate and merge discussion of the "career" topics and "major" topics.

After preparing the anchor words, we proceed to the second step, the domain knowledge-driven correlation explanation topic model (CorEx topic model). The core principle of the second step is to improve the topic word probability distribution by providing some anchor words integrated with domain knowledge before the model. Unlike previous practice, the anchor words incorporated in this paper are not those with the highest interactive information with tags but the most apparent topics-related topic words generated by the unsupervised topic model result combined with the domain knowledge experience and research objectives. Merged anchor words can enhance the document topic distribution by favoring the document to draw topics related to the anchor words it contains (Jagarlamudi et al., 2012). Furthermore, by setting the anchoring intensity, this method changes the degree to which the anchoring words affect the topic model. Anchor strength controls how much weight the CorEx topic model maximizes on the multivariate mutual information between anchoring words and their respective topics (Gallagher et al., 2017). The greater the intensity, the greater the influence of anchor words on the topic model. Generally, when anchor strength is set to a value less than 5, the effect is not mandatory but encourages topic word distribution to be generated according to the topic related to the anchor word.

We then explained the implementation of the CorEx topic model. First, we use the form of KL divergence to define Total Correlation (TC) or multivariate mutual information (Kraskov et al., 2005) as:

$$TC(X_G) = \sum_{i \in G} H(X_i) - H(X_G) = D_{KL}\left(P(X_G) \middle\| \prod_{i \in G} P(X_i)\right) \tag{3}$$

, where H(X) is the entropy of the discrete random variable X, and G is a subset of X.

In simple terms, it is the sum of the entropies of each random variable in the subset minus the joint entropies of the subsets. At the same time, TC can be regarded as the KL divergence between the joint and edge distribution multiplicative (Steeg & Galstyan, 2014). Then, when TC is zero, KL divergence is zero, and the joint distribution is equal to the edge distribution multiplicative, meaning that the correlation within the data is zero, and the variables are independent.

Then the conditional TC is defined as follows:

$$TC(X|Y) = \sum_{i \in G} H(X_i|Y) - H(X|Y) \tag{4}$$

After this, we can measure how much Y explains the correlation of X:

$$TC(X;Y) = TC(X_G) - TC(X|Y) \tag{5}$$

If TC(X|Y) is equal to zero, TC(X; Y) will get the maximum value, which means that Y explains all the correlations of X. In order to maximize the TC value of the document to improve the representation of topics, the CorEx algorithm will be re-started multiple times and perform the run with the highest TC value.

CorEx topic model defines a topic by P(Y|X). That is to say, it is only defined as a discrete random variable that can affect X, and the value range is k possibility. At initialization, we randomly set α and the topic distribution P (Y|X) of the documents. Semi-supervised is achieved by fixing some values of the connectivity matrix. Normal α is between (0, 1) interval, and the anchoring word i to topic k can make $\alpha_{\mathrm{i,k}} = \delta_{\mathrm{i,k}}$, where δ is the strength of anchoring (Gallagher et al., 2017). Then through iteration, LDA constantly updates the topic assigned for each word, thus indirectly obtaining the topic distribution of the document and the topic word distribution. At the same time, CorEx topic model continuously updates the probability of each topic P(y_{k}), the topic distribution P($y_{\mathrm{k}}|x_{\mathrm{i}}$) of each word, the distribution matrix α of words to topic sub-sets, and the topic distribution P(y_{k}|x) of each document according to the formula.

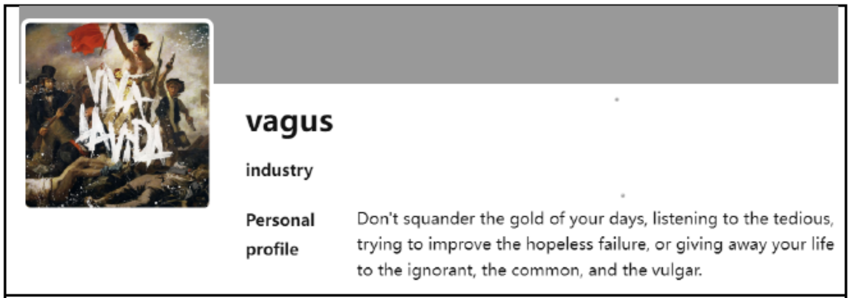
To conclude, we discuss the advantages of this approach over general probabilistic topic models. First, unlike LDA methods, no structural assumptions are required on the data, so CorEx topic model has fewer hyperparameters than LDA. In addition, CorEx can be generalized to hierarchical models and semi-supervised models without structural modification of the model, which can effectively meet the higher fine-grained requirements in IS subject research. Third, the trained topic model can be adapted to the domain through the unsupervised generation of anchor words. Most importantly, the words between topics in CorEx are disjointed, and there will be no repeated topics, which perfectly solves the problems of topic interpretability and separation.

Illustrative Case

To verify whether our proposed two-stage Corex topic modeling approach can be effectively applied to information systems research, we conducted an illustrative case study using 4,290,484 user-generated personal descriptions datasets on a representative online platform Zhihu.com. There are two reasons why we choose Zhihu as the research platform. First, we need enough texts to build a corpus, and Zhihu, one of China's largest online question-and-answer platforms, can meet this condition. Another reason is that we hope that the fields covered by the corpus are less studied to verify our two-stage model. Personal profile information on Zhihu is related to personal expression. This kind of information is less considered in traditional topic modeling because it is prone to problems of low interpretability and overlapping topics. Specifically, we use the LDA unsupervised topic modeling method to extract the topic words involved in the personal descriptions of Zhihu's four million users and use them to build a set of anchor words. We then embed these integrated anchor words into the Corex topic model for modeling. Finally, we summarize the topic modeling results to verify the topic separation and compare the interpretability of the two-stage Corex model with that of the Corex model.

Data Collection and Pre-processing

We collected personal disclosure text information (as shown in Figure 2) from Zhihu.com, the largest Chinese online question-and-answer community. To obtain as much user data as possible, we use a recursive crawl method. Specifically, we first found a user with many followers. After capturing this user's information, we obtained the information of all the users in his following list and fan list. We then crawl these users' their fan list and following list. And using this method, the data of 4,376,500 users had been captured by July 10, 2017. After data cleaning, such as encoding, sorting, replacing missing values, and deleting duplicate values, 4,290,484 clean records were obtained. Finally, in addition to name and profile picture, the user can provide some standardized information such as gender, residence, industry, career experience, educational experience, and personalized information: personal profile, as shown in Table 3. A profile comprises a short text description that allows users to describe themselves. For example, the text could disclose where they live or more private information such as their social accounts on other platforms, hobbies, personality, and physical appearance. Next, we pre-processed the text data. First, stop words are filtered out. Next, we delete all symbols except Chinese characters since we study the single-language text. Finally, we have a word segmentation operation.



vagus

industry

Personal profile
Don't squander the gold of your days, listening to the tedious, trying to improve the hopeless failure, or giving away your life to the ignorant, the common, and the vulgar.

Figure 2. User Interface on Zhihu.com

Field Name	Field Explanation	Example
Gender	Gender information disclosed by the user	Female

Place of Residence	Location information disclosed by the user	Nanchang, China
Career Experience	Employment information disclosed by the user	Programmer
Educational	Education information disclosed by the user	Nanchang University
Personal Profile	User-generated text describing oneself	This is a programmer in Nanchang

Table 3. Data Structure

Step 1-Latent Dirichlet Allocation Topic Model

There are two metrics to select the best number of topics: perplexity and coherence. However, the coherence eigenvalue could be a better proxy for the comprehensibility of topic headings compared to the perplexity. UMass is one of the methods to calculate coherence score, so we use UMass to calculate semantic consistency.

First, we find the optimal number of LDA topics by calculating the UMass Coherence score, the calculation formula is as follows (Mimno et al., 2011).

$$\mathrm{C}\left(\mathrm{k}; V^{(t)}\right) = \sum_{m-2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(V_m^{(k)}, V_l^{(k)}\right) + 1}{D(V_l^{(k)})} \tag{6}$$

This score represents the ratio of the frequency of one word and the frequency of another word contained in the words of all topic M under a certain topic k. Then the logarithm of this ratio and the normalization operation is taken. This is the conditional likelihood of co-occurrence between words because people's understanding of the topic model is more inclined to the frequency of co-occurrence of words belonging to the same topic in the corpus. UMass is a negative value; the closer it is to zero, the better the topic has been learned. We draw the coherence score curve as shown in figure 3, and it can be found that when the number of topics is 9, the learning effect of the LDA model is the best.

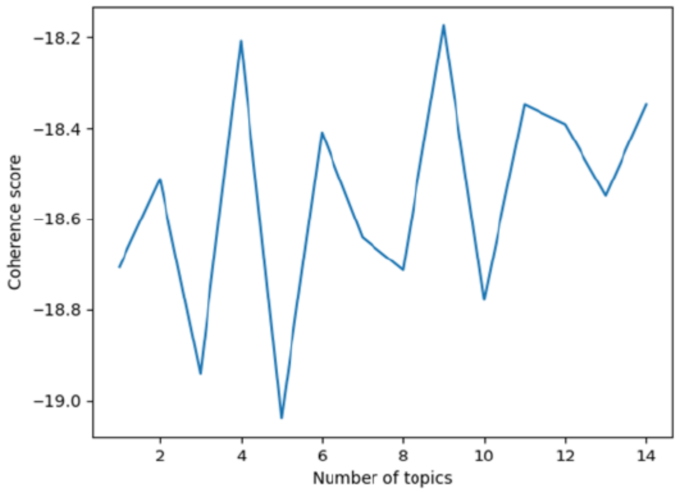


Figure 3. The Relationship between Coherence Score and the Topic Number

We listed the top 8 words contained in the 9 topics generated by the LDA in table 4.

Topic	word 1	word 2	word 3	word 4	word 5	word 6	word 7	word 8
Topic 1	student	major	college	senior year	design	hobby	computer science	engineering
Topic 2	study	undergraduate	jurisprudence	graduate student	master	accounting	medical science	doctor
Topic 3	software	machinery	development	car	architecture	design	Sales	technology
Topic 4	designer	product	game	manager	future	car	software	product
Topic 5	hobbyist	photography	movie	fitness	music	game	history	travel
Topic 6	college student	at school	major	engineering	startup business	study	take part in the postgraduate entrance exams	hobby
Topic 7	like	music	travel	photography	world	study	history	movie
Topic 8	life	strive	world	dream	work	freedom	music	design
Topic 9	finance	Internet	migrant worker	accounting	industry	practitioner	investment	product

Table 4. Result for Step 1 – LDA Topic Model

Then, based on these generated keywords and the domain knowledge about personal introduction in social media, we determined this paper's anchor words and corresponding categories. First, we remove the repetitive words that lead to topic coverage, such as "study," "major," and "design." Then, we look for words with category meaning from the topic words of LDA, such as "major," "hobby," "industry," and "study," and map other subject words into corresponding categories. Finally, a manual summary is made for the remaining topic words. For example, "interesting," "friendly," and other words are summarized into "personality." In addition, if a set of topic words is domain-related features, but the keywords do not appear in the LDA results, we can generate a set of anchor words based on the corpus. In this research, "account" is a set of anchor words that we generate manually. This kind of vocabulary comes from the corpus but is not shown in the LDA results. The reason is that this group of words does not own a high weight because they do not appear frequently in the corpus. Finally, we determined the anchor words of seven categories of topics and input them into the correlation explanations topic model. The anchor words are shown in Table 5.

	Anchored Word 1	Anchored Word 2	Anchored Word 3	Anchored Word 4
Industry	industry	practitioner	manager	doctor
Education	Study	undergraduate	college	master
Major	major	accounting	finance	software
Hobby	hobby	game	music	travel
Personality	personality	interesting	friendly	extroverted
Account	WeChat	microblog	public account	contact
Appearance	cute	good-looking	beautiful	handsome

Table 5. Anchor Words Generated by LDA

Step 2- Anchored Correlation Explanation Topic Model

We input the anchor words shown in Table 5 into the CorEx topic model according to the corresponding category, and the model reaches convergence after 20 iterations, as shown in figure 4.

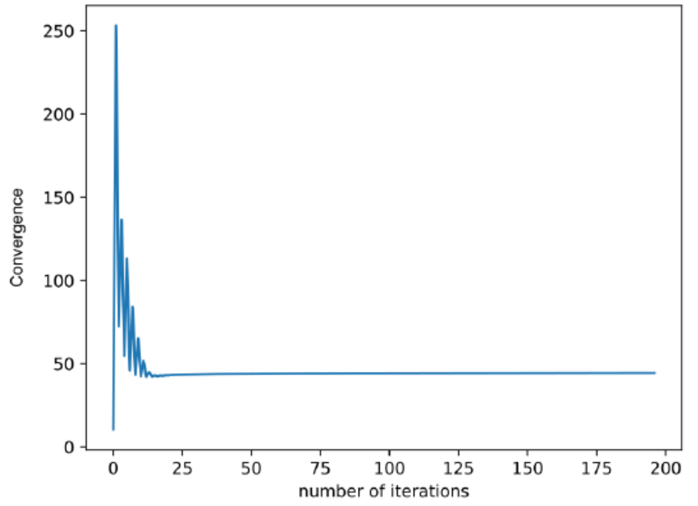


Figure 4. Iteration and Convergence

We output the final topic modeling results in Table 6, containing the top five topic words and the multivariate mutual information values with topics. These topic words are those with the highest multivariate mutual information with the topic, not the words with the highest within-topic probability as in LDA. For example, the topic words "Design," "engineering," "financial," "programming," and "Internet" can effectively summarize topic 1: Industry. We found that the two-step CorEx topic model generated the corresponding topics according to the seven categories we had set in advance. Moreover, there is no repetition of words between topics, which solves the problem of model interpretability and topic coincidence.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5
Topic 1: Industry	design	engineering	financial	programming	internet
	0.447	0.444	0.221	0.157	0.114
Topic 2: Education	study	college	senior high school	undergraduate	doctor
	0.606	0.534	0.148	0.146	0.120
Topic 3: Major	medical science	accounting	psychology	jurisprudence	computer science
	0.213	0.194	0.107	0.099	0.083
Topic 4: Hobby	hobby	photography	music	fitness	game
	0.720	0.197	0.170	0.158	0.138
Topic 5: Personality	strive	interesting	friendly	optimism	extrovert
	0.171	0.085	0.012	0.011	0.006
Topic 6: Account	WeChat	microblog	public account	private message	account number

	0.288	0.056	0.024	0.024	0.015
Topic 7: Appearance	cute	good-looking	beautiful	pretty	appearance
	0.057	0.016	0.007	0.007	0.006

Table 6. Result for Step 2

Evaluation Results

We compare the total correlation between the two-step CorEx topic model and CorEx topic model. Since CorEx topic model is not a generative but a judgmental model, traditional measures such as perplexity will no longer apply to the topic modeling approach in this paper. Perplexity only computes maximum likelihood and does not consider context (Chang et al., 2009), so we employ topical coherence scores to compute semantic coherence. Considering our model is quite different from the traditional LDA model, and the Corex model still lacks a mature Python package for corresponding consistency calculation, we need to improve the corresponding consistency calculation formula to generate the calculation code. In this paper, we choose the C_V metric to calculate coherence, which uses a sliding window and a window size of 110 to retrieve a co-occurrence count for a given word. And the reason we choose C_V is that its expected value is between 0 and 1, which makes it easier for us to generate the corresponding computational. At the same time, larger values of TC indicate better training of the CorEx topic model, as the model generates more topics of information about the document.

As shown in Table 7, the total correlation value of the CorEx topic model is 2.984, while the TC value of the two-step CorEx topic model is 8.331, indicating that the two-step CorEx topic model generates more topics about document information. And superior to the CorEx topic model. In addition, we found that the coherence score of the two-step CorEx topic model is greater than that of the CorEx topic model, which indicates that the topics generated by our proposed two-step CorEx topic model are more meaningful and semantically coherent and have better aspect interpretability. Moreover, we found that C_V values are close to 0.7 for both the two-step CorEx topic model and the CorEx topic model, indicating that both models are very nice and have strong interpretability and semantic coherence.

Topic	**CorEx topic model**		**Two-step CorEx topic model**	
	C_V (Coherence Score)	**T_C (Total Correlation)**	**C_V (Coherence Score)**	**T_C (Total Correlation)**
Average	0.681	2.089	0.686	8.331
Topic 1	0.694	0.436	0.677	0.638
Topic 2	0.703	0.384	0.711	2.164
Topic 3	0.682	0.356	0.658	1.494
Topic 4	0.708	0.299	0.663	0.923
Topic 5	0.674	0.277	0.706	2.109
Topic 6	0.651	0.187	0.673	0.686
Topic 7	0.658	0.151	0.713	0.318

Table 7. Comparative Analysis Result

Conclusions

We found that the LDA model effectively generates vocabularies related to prior domain knowledge. This operation is fast on a large data set containing 4,290,484 personal profiles. More importantly, these generated vocabularies can help us quickly understand the overview of the aspects contained in the text so that we can quickly organize the topic categories we want to generate and the corresponding vocabulary. We then fed these anchor words into the CorEx topic model, and after 20 iterations, the model reached convergence. We output the topic words corresponding to 7 topics and found that these words can explain the topics we expected to generate, as shown in Table 6.

Our two-step prior domain knowledge topic modeling also passed the validity test. First, the total correlation of the two-step CorEx topic model is greater than that of the CorEx topic model, indicating that the anchor words constructed by combining the topic words generated by LDA and the researchers' domain knowledge are better than anchor words with the highest multivariate mutual information with the label, the former generates more topics about document information. Additionally, we compare the coherence scores of the two-step CorEx topic model and the CorEx topic model. The result indicates that our proposed topic model has better aspect interpretability and semantic coherence.

As a result, the two-step method can extract domain knowledge-driven categories easily and quickly and make up for the shortcomings of traditional topic modeling methods. Specifically, this topic modeling method can alleviate the poor interpretability problems caused by traditional topic modeling methods, such as topic overlap and difficulty summarizing. More importantly, this domain knowledge-driven approach can effectively improve vocabulary distribution characteristics without compromising the quality of other topics to generate expected topics.

Theoretical and Practical Implications

This research has a certain contribution to theory and practice. The theoretical contributions are as follows. First, this paper outlines the methodological and application research on topic modeling in IS research. After that, we summarize the limitations of the methods used in current research, which will help IS researchers to understand different topic methods and their applications in a framework to have Helps to facilitate a broader discussion of topic model literature research and IS research methods. Second, we converted the method into an easy-to-understand graph structure by sorting out the logic of the method, and compared the concepts in LDA with those in CorEx topic model for a comparative explanation, which expanded the research theory on the topic model in the field of IS and enriched the existing theoretical research of LDA topic model. Third, we propose to use the LDA method to replace the original CorEx method of acquiring anchor words, and the effectiveness of this two-stage CorEx method has been verified, which provides a certain reference for the theoretical research of the CorEx method in IS research. Finally, the new method of integrating LDA topic model and CorEx topic model proposed in this research is of great significance for studying aspects of the text, which provides the possibility of higher fine-grained research for text analysis in IS research and a reference for further study of topic modeling methods in ISR. Specifically, this approach addresses the limitations of current IS research related to topic modeling. It can improve the interpretability and separability of topics and effectively control the lexical distribution characteristics, which is of great significance for extracting and classifying text content in specific fields in information system research.

The practical significance of this paper is as follows: first, we innovatively propose a new approach to solve researchers' main problems using topic models in current IS research; second, we provide a reference for how to apply this two-step Anchored correlation explanation topic model in IS research; finally, the implications of applying this approach could be widespread, as it would be possible to extract more specific aspects of textual data, such as reviews of online products or services or tweets in social media, which will be of great significance for various online platforms with user-generated text data. Research practitioners can benefit from this new approach. Using this method, they can generate expected topics according to user-generated text from different fields and understand text content at a higher level of granularity. For example, for tweets related to personal information, topics such as "hobby," "occupation," "personality," and "profession" can be generated and classified, which can effectively help researchers to explore the relationship between the content dimension of the text and other factors. At the same time, this method can help them generate domain knowledge anchor words quickly and accurately. At a time when the CorEx method is not widely used in information systems research, we illustrate the concept of the method to help researchers understand it, thus providing a broader opportunity for textual data-driven information systems research.

Limitations and Future Study

This study still has certain limitations. First, validation of the method relies on data from a single website. Given that the analysis object is text data, this provides a certain reference value for applying the two-step topic model in IS research. However, to increase the generalizability and validity of the results, the same technique should be applied to textual data from other types of websites and various services, such as online travel guide tripadvisor.com or online gig economy platform fiverr.com, which will provide deeper insights into IS research in the travel and outsourcing industries. In addition, the comparison of methods is still insufficient. Except for the CorEx topic model, this two-step topic model should be compared with other existing supervised topic models, such as Labeled LDA (LLDAModel), Supervised LDA (SLDAModel), and semi-supervised topic models such as Partially Labeled LDA (PLDAModel), etc. for effect comparison, to verify the effectiveness of the method further.

Acknowledgements

The work described in this paper was partially supported by InnoHK initiative, The Government of the HKSAR, and Laboratory for AI-Powered Financial Technologies.

References

Ahmad, S. N. & Laroche, M. (2015). How do expressed emotions affect the helpfulness of a product review? Evidence from reviews using latent semantic analysis. *International Journal of Electronic Commerce*, *20*(1), 76–111. https://doi.org/10.1080/10864415.2016.1061471

Andrzejewski, D., Zhu, X. & Craven, M. (2009). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *ACM International Conference Proceeding Series*, *382*. https://doi.org/10.1145/1553374.1553378

Banker, R. D., Frost, T. S. & Tripathi, M. K. (2022). The Determinants of InformationWeek 500 Selection and Its Implications: A Textual Analysis Approach. *Journal of Information Systems*, *36*(1), 81–109. https://doi.org/10.2308/ISYS-2020-070

Bao, J., Chen, Y., Yin, J., Chen, X. & Zhu, D. (2023). Exploring topics and trends in Chinese ATC incident reports using a domain-knowledge driven topic model. *Journal of Air Transport Management*, *108*, 102374. https://doi.org/10.1016/j.jairtraman.2023.102374

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Boschetti, A. (2015). *Automatic topic-modelling with Latent Dirichlet Allocation*. Intent HQ. https://engineering.intenthq.com/2015/02/automatic-topic-modelling-with-lda/

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 288–296. https://proceedings.neurips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

Debortoli, S., Müller, O., Junglas, I. & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, *39*(1), 110–135. https://doi.org/10.17705/1cais.03907

Denny, M. J. & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

Denzler, T., Enders, M. R. & Akello, P. (2021). Towards a semi-automated approach for systematic literature reviews. *27th Annual Americas Conference on Information Systems, AMCIS 2021*. https://scholar.archive.org/work/ybs7vmz3cvf2fkaegazkuzytaa/access/wayback/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1251&context=amcis2021

Dumais, S. T. (2004). Latent Semantic Analysis. In *Annual Review of Information Science and Technology* (Vol. 38, pp. 188–230). Information Today. https://doi.org/10.1002/aris.1440380105

Eickhoff, M. & Neuss, N. (2017). Topic modelling methodology: Its use in information systems and other managerial disciplines. *Proceedings of the 25th European Conference on Information Systems, ECIS 2017*, 1327–1347. https://core.ac.uk/download/pdf/301372354.pdf

Evangelopoulos, N., Zhang, X. & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70–86. https://doi.org/10.1057/ejis.2010.61

Gallagher, R. J., Reing, K., Kale, D. & Ver Steeg, G. (2017). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Transactions of the Association for Computational Linguistics*, *5*, 529–542. https://doi.org/10.1162/tacl_a_00078

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, 50–57. https://doi.org/10.1145/312624.312649

Io, H. N. & Lee, C. B. (2020). Social Media Comments about Hotel Robots. *Journal of China Tourism Research*, *16*(4), 606–625. https://doi.org/10.1080/19388160.2020.1769785

Jabr, W., Zhao, K., Cheng, Y. & Srivastava, S. (2018a). Dynamics of Online Review Text: What Changes and What Matters over Time. *Ssrn*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3208339

Jabr, W., Zhao, K., Cheng, Y. & Srivastava, S. (2018b). What are they saying? A methodology for extracting information from online reviews. *International Conference on Information Systems 2018, ICIS 2018*. https://www.researchgate.net/profile/Yichen-Cheng-3/publication/330600656_What_Are_They_Saying_A_Methodology_for_Extracting_Information_from_Online_Reviews/links/5c49f335a6fdccd6b5c597e2/What-Are-They-Saying-A-Methodology-for-Extracting-Information-from-O

Jagarlamudi, J., Iii, H. D. & Udupa, R. (2012). Incorporating lexical priors into topic models. *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, 204–213. https://aclanthology.org/E12-1021.pdf

Jeong, J. (2023). Identifying Always-the-Same-Rating Reviewers on Amazon.Com Using Big Data Analytics. *IEEE Transactions on Computational Social Systems*, 1–16. https://doi.org/10.1109/tcss.2023.3235727

Kar, A. K. & Dwivedi, Y. K. (2020). Theory building with big data-driven research – Moving away from the "What" towards the "Why." *International Journal of Information Management*, *54*. https://doi.org/10.1016/j.ijinfomgt.2020.102205

Kraskov, A., Stögbauer, H., Andrzejak, R. G. & Grassberger, P. (2005). Hierarchical clustering using mutual information. *Europhysics Letters*, *70*(2), 278–284. https://doi.org/10.1209/epl/i2004-10483-y

Lancichinetti, A., Irmak Sirer, M., Wang, J. X., Acuna, D., Körding, K. & Amaral, L. A. (2015). High-reproducibility and high-accuracy method for automated topic classification. *Physical Review X*, *5*(1). https://doi.org/10.1103/PhysRevX.5.011007

Lee, S., Song, J. & Kim, Y. (2010). An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, *51*(1), 1–10. https://doi.org/10.1080/08874417.2010.11645444

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H. & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 262–272. https://aclanthology.org/D11-1024.Pdf

Mukkamala, A. & Beck, R. (2018). The role of social media for collective behaviour development in response to natural disasters. *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018*. https://pure.itu.dk/portal/en/publications/the-role-of-social-media-for-collective-behaviour-development-in-

Nikolenko, S. I., Koltcov, S. & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393

Olorunsola, V. O., Saydam, M. B., Lasisi, T. T. & Ozturen, A. (2023). Exploring tourists' experiences when visiting Petra archaeological heritage site: voices from TripAdvisor. *Consumer Behavior in Tourism and Hospitality*. https://doi.org/10.1108/cbth-05-2021-0118

Palese, B. & Piccoli, G. (2020). Evaluating topic modeling interpretability using topic labeled gold-standard sets. *Communications of the Association for Information Systems*, *47*, 433–451.

https://doi.org/10.17705/1CAIS.04720

Rai, A. (2016). Editor's Comments: Synergies Between Big Data and Theory. *MIS Quarterly*, *40*(2), iii–xi. https://dl.acm.org/doi/abs/10.5555/3177617.3177618

Ramage, D., Dumais, S. & Liebling, D. (2010). Characterizing microblogs with topic models. *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, 130–137. https://doi.org/10.1609/icwsm.v4i1.14026

Steeg, G. Ver & Galstyan, A. (2014). Discovering structure in high-dimensional data through correlation explanation. *Advances in Neural Information Processing Systems*, *1*(January), 577–585. https://proceedings.neurips.cc/paper/2014/hash/4f6ffe13a5d75b2d6a3923922b3922e5-Abstract.html

Titov, I. & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, 111–120. https://doi.org/10.1145/1367497.1367513

Twitter. (2023). *Essential Twitter Statistics You Need to Know in 2023*. Socialshepherd. https://thesocialshepherd.com/blog/twitter-statistics

Wallach, H. M., Mimno, D. & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 1973–1981. https://proceedings.neurips.cc/paper/2009/hash/0d0871f0806eae32d30983b62252da50-Abstract.html

Wei, C. P., Chiang, R. H. L. & Wu, C. C. (2006). Accommodating individual preferences in the categorization of documents: A personalized clustering approach. *Journal of Management Information Systems*, *23*(2), 173–201. https://doi.org/10.2753/MIS0742-1222230208