Topic Modeling on Indonesian Online Shop Chat

Ahmad Fathan Hidayatullah
Department of Informatics, Universitas Islam Indonesia
Yogyakarta, Indonesia
fathan@uii.ac.id

Wisnu Kurniawan
Department of Informatics, Universitas Islam Indonesia
Yogyakarta, Indonesia
wisnuko13@gmail.com

Chanifah Indah Ratnasari
Department of Informatics, Universitas Islam Indonesia
Yogyakarta, Indonesia
chanifah.indah@uii.ac.id

ABSTRACT

This paper aims to discover topics from an Indonesian online shop chat. Moreover, we employed Latent Dirichlet Allocation to find out what kind of topics that are often discussed and conversation trends between buyers and customer service. Several tasks were performed, such as, collecting data, preprocessing, phrase aggregation, topic modeling, and topic analysis. We found several attracting findings during our experiments. In preprocessing task, product name extraction from URLs assisted to discover the intended product from the customer's conversation. On the other hand, the phrase aggregation task helped us to merge various terms which have same intended meaning, so that, we could obtain better topical model result and easier to determine the topic label.

CCS Concepts

• Computing methodologies~Discourse, dialogue and pragmatics • Computing methodologies~Topic modeling • Applied computing~Online shopping

Keywords

Topic modeling; Topic model; Latent Dirichlet Allocation; Online Shop; Bahasa Indonesia.

1. INTRODUCTION

Today, buy and sell transactions are not only happened in the real world, but also happened online through the internet. People can simply find, order, and purchase their needs just by shopping through online shops. Moreover, online shops are not only selling their goods through the website, but also through social media and instant messaging platform.

Customer service is one of the most important things in an online shop. Good customer service will give satisfaction to the customers. Therefore, there are some chat platforms used by the online shops to communicate and chat with their customers, such as Facebook messenger, WhatsApp, Line, and chat widget on their website. With these chat platforms, customers can easily make conversations with the customer service with a variety of chat application choices based on the chat apps used by customers.

The current chat platform can already find out about how much traffic conversations weekly, daily, to hourly periods. However, there is no further analysis of what conversations are often discussed in order to understand the needs of buyers. The information about the needs of buyers can provide appropriate policies to be applied at the right time. For example, in determining promotional policies for a product at a certain time, knowing the trends of products, and knowing the habits of buyers. It is also important for the owner, manager, or customer service to obtain the topics from the chat history. This information will be very helpful and provide benefits for the online shops.

According to the explanation above, topic modeling approach can be proposed to discover topics from the chat between customers and customer service in an Indonesian online shop using Latent Dirichlet Allocation (LDA) topic modeling. The revealed topics also could provide an illustration about the customer's needs and trends in order to increase sales and customer satisfaction.

The rest of this paper is structured as follows: Section 2 provides the related work. Section 3 describes the theoretical background. Our experiment is discussed in section 4. In section 5, we describe our result and discussion. Finally, we provide our conclusion of this research in section 6.

2. RELATED WORK

A lot of researchers have employed LDA-based topic modeling and its modification to discover what topics that often appear in a large text document [16]. In this section, we will discuss about some previous works that applied topic modeling in commerce domain. Ko, et al. [8] employed LDA-based topic modeling to generate product opportunities based on the customer's interest from online customer product reviews. Wang, et al. [15] also applied topic modeling to analyze customer preferences by extracting key topics of online product reviews using LDA topic modeling approach. Christidis, et al. [5] also analyzed customer preferences by utilizing LDA probabilistic topic model. According to the research, latent topic analysis could present illustration into the customer preferences and support in recommending items to users. Lee and Yoshihara [9] applied semi-supervised LDA topic modeling to reveal the customer's needs from customer purchase histories. Zhang, et al. [17] proposed dynamic topic modeling using Twitter dataset for monitoring temporal evolution of market competition. Anoop and Asharaf [1] used LDA to extract concepts and relationship the product illustration in e-commerce.

3. THEORETICAL BACKGROUND

3.1 Topic Modeling Using LDA

The concept of topic modeling considers that a document has the possibility to consist of several topics with their respective probabilities, where a topic is composed of a set of words [3]. Topic modeling is commonly used to perform tasks such as exploring corpus, classifying documents, and getting information from a corpus [14]. The one of the most popular and powerful

NLPIR 2019, June 28–30, 2019, Tokushima, Japan
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6279-5/19/06…$15.00
DOI: https://doi.org/10.1145/3342827.3342831

methods to perform topic modeling is Latent Dirichlet Allocation (LDA). LDA is a blended model which presumes that each document composes of assorted topics and those topics would generate some words [4]. LDA is a generative probabilistic model of a set of texts called corpus. There are three generative processes performed on each document in the corpus [2]:

- Select topics randomly from the distribution of topics for each document.
- Select words from the distribution of words related to the chosen topic.
- Repeat processes 1 and 2 for all existing documents.

3.2 Topic Coherence

In this research, we utilize topic coherence to obtain the appropriate number of topics from the dataset. Topic coherence captures semantic information from the topic produced and assesses the interpretation of the topic. Topic coherence has metrics that are consistent with human interpretation [11]. The higher the topic coherence score, the better the human interpretation. Topic coherence is calculated by making pairwise comparisons between words on a particular topic which results in a measure of the quality standard of a topic. Figure 1 shows the workflow for calculating topic coherence.

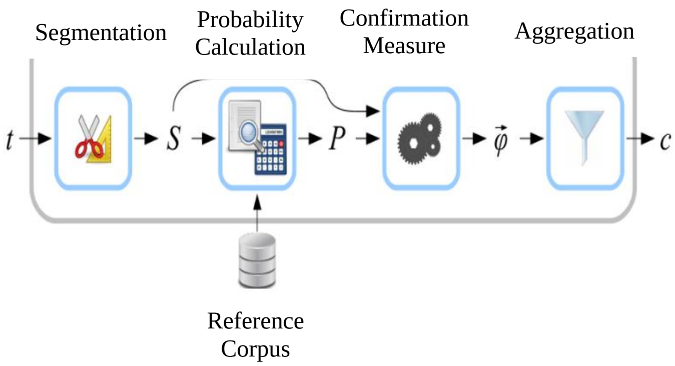


Figure 1. Topic Coherence Workflow [12]

Figure 1 provides the steps of topic coherence calculation that consists of the following tasks:

- Segmentation

This stage performs topic segmentation from a set of topics *(t)* to obtain a segmented topic *(S)*.

- Probability calculation

Calculate probability estimation from each topic segment based on the reference corpus to obtain calculated probabilities *(P)*.

- Confirmation measure

Measure the quality of topics according to a certain metric in each partition to obtain phi vector (φ) which is defined as a vector of the confirmed measures coming out from the confirmation module.

- Aggregation

Aggregate the quality numbers by calculating the average to produce the coherence value *(c)*.

In this study, we perform topic coherence calculation using Cv measure which is proven to have the best results with human interpretation. The Cv measure is a combination between the indirect cosine measure with the NPMI (Normalized Pointwise Mutual Information) and the Boolean sliding window [12].

4. EXPERIMENT

We conducted our experiment by following the pipeline as shown by the Figure 2. There are five phases in the pipeline, including data collection, preprocessing, phrase aggregation, topic modeling, and topic analysis.

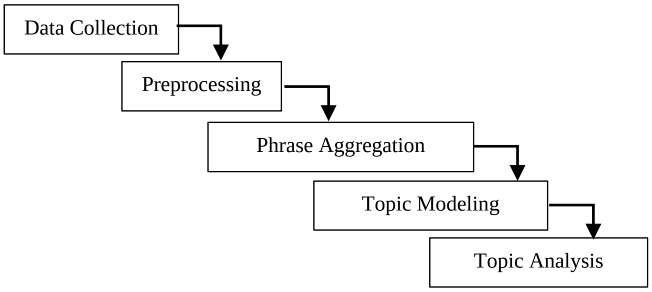


Figure 2. Experiment Pipeline

4.1 Data Collection

We collected the chat data from a fashion online shop in Indonesia, namely Berrybenka. All the conversations were written in Bahasa Indonesia, either formal or non-formal language. The data were collected from March until December 2017. This research has collected approximately 1,4 million conversations. Table 1 shows the details amount of raw chat for each month.

Table 1. Amount of Raw Data

Month	Amount of Raw Data
March	70,031
April	159,368
May	181,033
June	204,778
July	175,076
August	156,053
September	137,007
October	141,009
November	120,976
December	124,139
Total	1,469,470

4.2 Preprocessing

Our preprocessing tasks were divided into two types, common tasks and specific tasks. Common preprocessing tasks mean that those tasks are generally performed in text preprocessing, including case folding, stemming, stop word removal, abbreviated word normalization, and removing unused characters (symbols, punctuations, email, URLs, numbers). As for specific tasks, we employed some tasks that are only applied to overcome the problem for our dataset, such as, template message removal, product name extraction from a URL, and concatenating negation. All of the preprocessing steps performed in our experiments are described below:

- Removing template messages

This task removes the repeated sentences sent by customer services that appear more than 10 times in our dataset.

- Extracting product name from URL

This task aims to extract product name that usually found in the URL. Table 2 shows the example of extracting product names from the URL.

Table 2. Extracting Product Name from URL

Before	After
Barang nyya seperti ini ya kak https://berrybenka.com/clothing/tops/186801/gisela-blouse-in-peach	Barang nyya seperti ini ya kak **gisela_blouse_in_peach**
https://berrybenka.com/clothing/tops/244652/minidotie-zip-blouse-in-white	**minidotie_zip_blouse_in_white**

- Removing URLs

This task omits URLs that do not contain a product name.

- Removing email

This task removes email from the chat.

- Removing numbers

Numbers are removed from the chat.

- Case folding

Case folding transforms all letters in lowercase.

- Removing symbols and punctuation

Symbols and punctuations are removed in this task.

- Reducing repeated characters from word as shown in Table 3.

Table 3. Removing Repeated Characters

Before	After
haloooo	*halo* (hello)
Okeee	*Oke*

- Abbreviated word normalization

This task aims to normalize all abbreviated words into their standard form as shown in Table 4.

Table 4. Abbreviated Word Normalization

Non-standard Words	Standard Words
pesn	*pesan* (order)
trnsfr	*transfer* (transfer)

- Stemming

Stemming task transforms the word into its non-changing form by removing prefixes and suffixes.

- Removing word with less than 4 characters

This task removes meaningless words which contain less than 4 characters, for example, *'kak'* and *'nya'*.

- Concatenating negation

This task identifies the negation word in Indonesia, such as, *'tidak'* and *'tak'* which mean not. If there is a negation in a sentence, it will be concatenated with the next word [7]. Table 5 illustrates the example of concatenating negation.

Table 5. Concatenating Negation

Before	After
Barang itu sudah **tidak tersedia**	Barang itu sudah **tidak_tersedia**
Barang **tidak dikirim** jika uangnya belum ditransfer ya	Barang **tidak_dikirim** jika uangnya belum ditransfer ya

- Removing stop word

This task is important to reduce dimensionality by omitting some meaningless words.

4.3 Phrase Aggregation using N-gram

After all datasets has been preprocessed, we performed phrase identification task. In this research, we identify phrases as two or three words that minimum appear five times together. Therefore, in order to identify phrases from the sentence, we build a model using bigram and trigram model. After phrases were identified, we then performed phrase aggregation that aims to aggregate phrases which have the same intended meaning by building a base dictionary. Base dictionary will map phrases into their intended meaning. For example, in the first row in Table 6, the two phrases, *'link info'* and *'info link'* are aggregated into a phrase *'link_info'*.

Table 6. Phrase Aggregation

Phrases	Intended Meaning
link info, info link	*link_info* (information link)
status order, status pesan	*status_pesan* (order status)

4.4 Topic Modeling

We utilized Gensim library provided by Python to perform topic modeling. The steps to build topic model are described below:

- Build dictionary

A dictionary is created from bag of words in the dataset. Furthermore, the dictionary shows how many words and how many times the appearance of those words in the dataset.

- Build corpus

The dictionary then transformed into a TF-IDF (Term Frequency-Inverse Document Frequency) corpus by converting a list document into a document term matrix format.

- Build LDA model and coherence value calculation

In this research, we build LDA model using corpus and dictionary, and number of topics as the input parameter. The corpus and dictionary parameters are obtained from the previous tasks. Furthermore, we train the model by assigning 1 to 10 number of topics and calculating the coherence value for each number of topic to obtain the best number of topic.

5. RESULT AND DISCUSSION

5.1 Number of Topic

Before obtaining the topic modeling result, we have to specify the number of topics. In our study, the best number of topics in each month was obtained by calculating the coherence score with the range between 1 to 10 topics. The highest coherence score indicates the appropriate number of topics. Figure 3 shows the detail graphs of number of topic gaining against the coherence score, respectively, from March until December 2017.

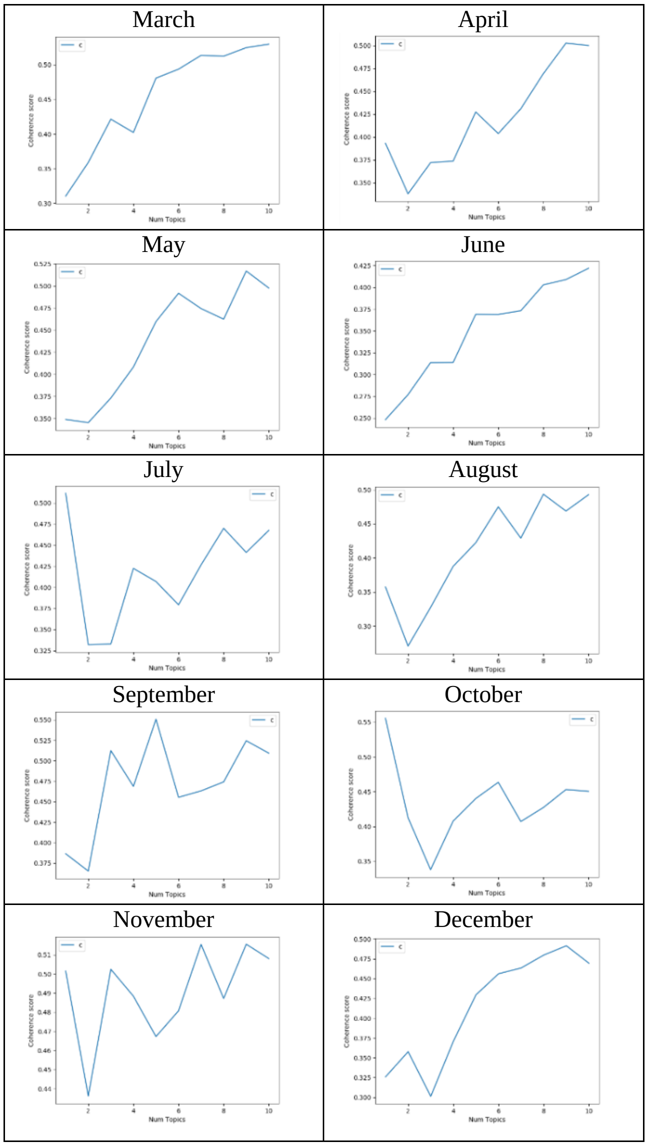


Figure 3. The Number of Topics Based on The Coherence Score

Overall, the coherence score in March, April, May, June, August, and December tended to increase. On the other hand, the topic coherence score of the rest months tended to fluctuate. In March, June, and August, the number of topics were achieving the maximum number with respectively 10 topics. Therefore, it can be inferred that there were various topics discussed in the conversation between the customer service and customers. On the contrary, the maximum coherence score in July and October were obtained when the number of topic is equal to one. This indicates that the conversations were only directed into a particular topic.

5.2 Topic Modeling Results

In this research, the topic modeling results are illustrated by utilizing the word cloud. Word cloud is a simple and communicative visualization that shows the most frequent words from text documents [10]. In word cloud, the larger the size of the words, the more often the words appear in a document. In addition, the size of the words also reveals the importance of the words in a collection of texts [6].

For example, the word promo in figure 4 have the largest size that indicates the word is frequently appeared in the conversation. In addition, we determined the topic label based on the most frequent words appeared in the word cloud. In particular, the topic label can be deduced by observing some interrelated words in the word cloud. According to figure 4, the word *'promo', 'diskon'* (discount), and *'order'* were quite dominating. Therefore, we can conclude that people were mostly talking about promotion, discount, and order in July and October.

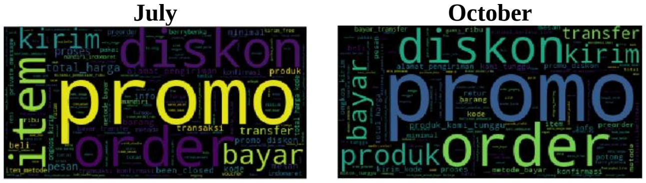
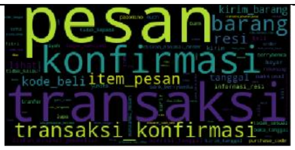
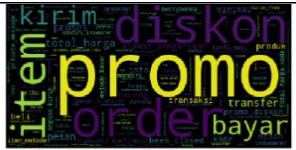
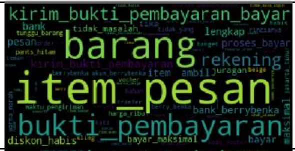
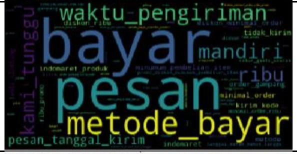
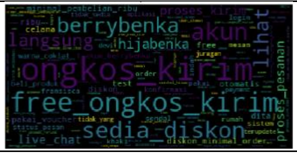


Figure 4. Word Cloud for July and October

In the next discussion, we will discuss about several examples of our topic modeling results. Table 7 shows five examples of the most interpretable topic results from our dataset.

Table 7. Word Cloud and Topic Label

No	Word Cloud	Topic Label
1	 pesan barang konfirmasi resi kode_beli item_pesan transaksi transaksi_konfirmasi	Order Confirmation Transaction
2	 kirim diskon item promo order bayar transaksi transfer	Promo and Discount
3	 kirim_bukti_pembayaran_bayar barang rekening item pesan bukti_pembayaran diskon_habis	Item order and Proof of Payment
4	 waktu_pengiriman bayar mandiri ribu pesan metode_bayar pesan_tanggal_kirim	Payment
5	 berrybenka akun langsung hijabenka ongkos_kirim free_ongkos_kirim sedia_diskon live_chat	Shipping Costs

For example, the most suitable topic label for word cloud number one is order confirmation transaction. This topic label was deduced from the most frequent words in the word cloud, including, *'pesan'* (order), *'transaksi'* (transaction), *'konfirmasi'* (confirmation), *'transaksi_konfirmasi'* (confirmation transaction) and *'barang'* (item).

In the word cloud number 2, the term promo is the most dominant word. Besides that, we also found the terms, such as, *'diskon'* (discount), order, and item. Therefore, it can be concluded that the most appropriate topic for word cloud number 2 is promo and discount. Word cloud number 3 illustrates about item order and proof of payment. This topic label was determined from the terms *'item_pesan'* (item order), *'barang'* (item), and *'bukti_pembayaran'* (proof of payment). The most suitable topic label for word cloud number 4 is payment. This is concluded from these terms, including, *'bayar'* (payment), *'pesan'* (order), and *'metode_bayar'* (payment method) that frequently appeared in the word cloud. The word cloud number 5 is related to shipping costs. This topic label is determined by the terms, including, *'ongkos_kirim'* (shipping costs) and *'free_ongkos_kirim'* (free shipping costs).

5.3 Comparison of Topic Model

In this section, we discuss about the comparison of our topic model results. In more detail, we compare our topic model result before and after performing phrase aggregation task. Figure 5 provides an example of topic model result visualization on the conversation data using PyLDAvis library in Python. PyLDAvis is a web-based interactive topic model visualization which intended to assist users interpret the LDA model result that has been fit to a corpus [13].

Particularly, we compared the topic model result by analyzing the top-10 most salient terms from the PyLDAvis visualization.

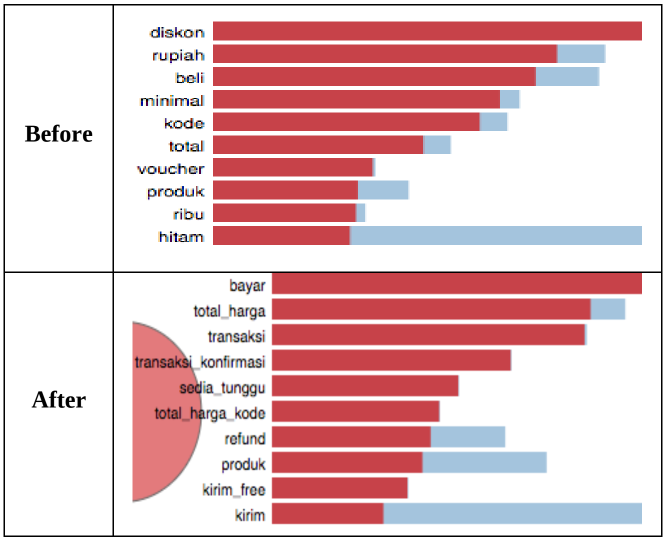


Figure 5. Comparison of Topic Model Results

Based on the result in Figure 5, the terms that appear before phrase aggregation task were too general such as, *'rupiah'* (Indonesian currency), *'beli'* (buy), *'minimal', 'kode'* (code), *'total', 'produk'* (product), and *'ribu'* (thousand). In addition, this is also difficult to find the relation between those words. Therefore, it causes difficulty in determining the topic label.

After phrase aggregation task were performed, the appearing words in the visualization were more specific and having more meaning. In addition, the topic result also more understandable and easier to specify the topic label than the previous results. From the result, we could see the relation between among the emerging words. For example, from the word *'bayar'* (pay), *'total_harga'* (total price), *'transaksi'* (transaction), and *'transaksi_konfirmasi'* (confirmation transaction), we can draw a conclusion that the most appropriate topic label for those terms is payment transaction.

6. CONCLUSION

In this paper, we applied topic modeling using Latent Dirichlet Allocation to discover topics from online shop conversation data. By applying topic modeling, we could find several insightful topics that illustrate necessary information from the customers. We also made several interesting discoveries through our experiments, particularly, in preprocessing and phrase aggregation task. In preprocessing task, product name extraction from URLs assisted to discover the intended product from the customer's conversation. As for phrase aggregation, it helped us to merge various terms which have same intended meaning, so that, we could obtain better topical model result and easier to determine the topic label.

7. REFERENCES

[1] Anoop, V.S. and Asharaf, S. 2017. A Topic Modeling Guided Approach for Semantic Knowledge Discovery in e-Commerce. *International Journal of Interactive Multimedia and Artificial Intelligence*. 4, 6 (2017), 40. DOI:https://doi.org/10.9781/ijimai.2017.03.014.

[2] Blei, D.M. et al. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res. 3*. 993–1022.

[3] Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM*. 55, 4 (Apr. 2012), 77. DOI:https://doi.org/10.1145/2133806.2133826.

[4] Chen, S. and Wang, Y. 2011. Latent Dirichlet Allocation. *Technical report, Department of Electrical and Computing engineering.* University of California, San Diego 4: 87-110.

[5] Christidis, K. et al. 2010. Exploring Customer Preferences with Probabilistic Topics Models. (2010), 13.

[6] Hidayatullah, A.F. et al. 2018. Twitter Topic Modeling on Football News. *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*. 467-471.

[7] Hidayatullah, A.F. and Sn, A. 2014. Analisis Sentimen Dan Klasifikasi Kategori Terhadap Tokoh Publik pada Twitter. *Seminar Nasional Informatika (SEMNASIF)*.

[8] Ko, N. et al. 2018. Identifying Product Opportunities Using Social Media Mining: Application of Topic Modeling and Chance Discovery Theory. *IEEE Access*. 6, (2018), 1680–1693. DOI:https://doi.org/10.1109/ACCESS.2017.2780046.

[9] Lee, T.Y. and Yoshihara, K. 2015. Getting to Why: Semi-Supervised Topic Modeling of Customer Purchase Histories. *SSRN Electronic Journal*. (2015). DOI:https://doi.org/10.2139/ssrn.2880805.

[10] Lohmann, S. et al. 2015. Concentri Cloud: Word Cloud Visualization for Multiple Text Documents. *2015 19th International Conference on Information Visualisation* (Jul. 2015), 114–120.

[11] Mimno, D. et al. 2011. Optimizing Semantic Coherence in Topic Models. *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262-272.

[12] Röder, M. et al. 2015. Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (Shanghai, China, 2015), 399–408.

[13] Sievert, C. and Shirley, K. 2014. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (Baltimore, Maryland, USA, 2014), 63–70.

[14] Wang, C. and Blei, D.M. 2011. Collaborative topic modeling for recommending scientific articles. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* (San Diego, California, USA, 2011), 448.

[15] Wang, W. et al. 2018. Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*. 29, (May 2018), 142–156. DOI:https://doi.org/10.1016/j.elerap.2018.04.003.

[16] Yang, T.-I. et al. 2011. Topic Modeling on Historical Newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 96-104.

[17] Zhang, H. et al. 2015. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15* (Sydney, NSW, Australia, 2015), 1425–1434.