Year: 2020

Assessing Customer Needs Based On Online Reviews: A Topic Modeling Approach

Jauhari, Thariq M ; Kim, Soomin ; Kovacs, Mate ; Serdült, Uwe ; Kryssanov, Viktor V

Originally published at:
Jauhari, Thariq M; Kim, Soomin; Kovacs, Mate; Serdült, Uwe; Kryssanov, Viktor V (2020). Assessing Customer Needs Based On Online Reviews: A Topic Modeling Approach. In: 5th International Workshop on Innovations in Information and Communication Science and Technology, Malang, Indonesien, 9 March 2020 - 9 March 2020, 57-62.

ASSESSING CUSTOMER NEEDS BASED ON ONLINE REVIEWS: A TOPIC MODELING APPROACH

Thariq M. Jauhari[1], Soomin Kim[1], Mate Kovacs[1], Uwe Serdült[1,2], Victor V. Kryssanov[1]

[1] *College of Information Science and Engineering, Ritsumeikan University, Japan, serdult@fc.ritsumei.ac.jp*
[2] *Center for Democracy Studies Aarau, University of Zurich, Switzerland*

ABSTRACT

The fashion industry is one of the most exposed to new online trends manifesting themselves on the internet. Whereas fashion consumers used to get inspired from their preferred brand or print magazine to buy clothes, today, they are rather influenced by social media and online reviews. Online shoppers look for clothes on their own, basing their choices on individual preferences and values. In other words, consumers have become more focused on "indirect experiences" and "exploration" rather than buying products from specific brands in the store. Furthermore, consumers want to know more about the products, and the fashion market demands greater transparency. From online reviews and ratings, consumers can gather a variety of helpful subjective information from each other. This research is conducted by looking at online product review data from Amazon, one of the leading online shopping websites worldwide, to reveal the hidden topics that are available within the review texts. To do this, topic modeling is applied to the data to explore customer preferences and consumption trends. The results show that the online reviews used in this study can be grouped into four general topics discussed online: Accessories, Outfit, Quality, and Appearance. With this information available, it would benefit and improve fashion businesses in account for product development.

Key words: Fashion, Customer Preference, Online reviews.

1. INTRODUCTION

Fashion clothing carries a wide range of ideological meaning these days. Fashion is a visual culture due to fashion trends that represents an individual's identity in a specific environment. In this regard, the fashion trends can be the social agenda which to express one's identity such as their attitude and lifestyle. Today, the manufacturing of fashion clothing has been influenced by technological advances. The fashion industry is a multi-billion-dollar industry with direct cultural, social, and economic implications. In this regard, various fashion companies produce fashion outfits to attract consumers. To attract the consumers and due to high competition between companies, companies design strategies. One of the most important strategies these days is about understanding customer preferences. Understanding and analysing customer requirements are related to product development and to the success of marketing strategies. Especially in the fashion industry, it is important to read the trends due to fast rapidly changing times.

According to this, it is significant to find the method that will be most appropriate to a business model that the company adopts. Analysing and understanding customer preferences and needs are beneficial to help businesses grow. To analyse consumer preferences and possibly forecast the trends of fashion clothing, it also important to understand the connection between the industrial revolution and the fashion industry. As the influence of fashion magazines and the fashion industry moving to online retail, fashion is attracting more attention. In the case of the E-commerce industry, companies are benefited through the huge amount of customer's reviews. These days, online stores have become large scale shopping channels for selling a wide range of products. Data shows that over 60% of consumers worldwide (Asia, Middle East, Latin America, and Africa) are eager to shop online (Nielson.com, 2015). As many people became online shoppers, online reviews play a vital role in providing influential information which affects consumer decisions in online shopping (Chan *et al.*, 2008; Duan *et al.*, 2008; Engler *et al.*, 2015). Online reviews contain valuable information about products. However, the collected information will be wasteful if the company does not utilize it in an appropriate way. Companies that use e-commerce options have the ability to use feedback from online shops to improve products, putting them ahead of companies that do not use online services. If the company do not equip appropriate technology, however, it is worthless to gather a huge amount of customer information. This paper proposes a study on customer need assessment based on Amazon customer reviews in the fashion segment. Customer reviews are the insight of what is going well in the businesses with the products being offered and also which sectors that need an improvement. Hence, it provides essential information to better adjust the businesses to fit the customers' needs more accurately. Our goal is to help online

IICST2020: 5th International Workshop on Innovations in Information and Communication Science and Technology, Malang, Indonesia

retailers to obtain requirement elicitation for their product innovation and improvement as well as to increase product transparency to their consumers.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 demonstrates the methodology. Section 4 shows the result and discussion. Lastly, in section 5, we offer concluding remarks.

2. RELATED WORK

Microscopic fashion is a kind of general mechanism in a variety of aspects that include modern lifestyles and in particular, individual preference (Aspers and Godart, 2013). In other words, it is the same as explaining that younger generations have a different style of clothes and type of music than the older generation. Fashion also reminds us of synchronism. To make fashion unique, we do not only choose clothes based on personal preference, but also tonify the particular appearance through hairstyles, body shape, and behaviour. Fashion is as well be seen as a definitive measure of socio-economic class. However, nowadays, fashion is often described as an industry selling life and dreams, not as a system selling clothes.

In this rapid-growing digital era, big data has brought revolutionary change in businesses, especially in fashion design. Any technological tools that could possibly help to give the best outcome for the businesses, for instance, data mining, analysis, and engineering, would be carried out to process information for better decision making and set strategies (Brown *et al.*, 2011).

Big data analytics is the process of examining large and varied data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions (Jain *et al.*, 2017). In the last decade, the aspect of fashion design has been changed. It has become possible only because of big data (McAfee and Brynjolfsson, 2012). It was not possible to investigate the choice of every user's opinion about how they feel about a fashion, what they think about it and how would they prefer to have it. At present, big data gives us the opportunity to review each user's opinion and to predict which fashion will they think perfect for them which creates the versatility in fashion design. Big data has allowed businesses to create targeted marketing campaigns. From top to bottom, companies use big data to ensure the quality of products, fix the target market and develop new innovative styles in order to keep pace with the incessant demand for creative new styles in fashion.

A research in text analysis worked on aspects evaluated on online reviews and how sentiment responds to different aspects based on two different sets of reviews (Jo and Oh, 2011). One data set is an electronic device reviews from Amazon, and the other data set is a restaurant reviews from Yelp. These two datasets have 22,000 total reviews and this study randomly selected about 5000 reviews from each section. By using sentiment classification and language models such as SLDA and ASUM, it discovered aspects and sentiment in a large amount of reviews.

3. METHODOLOGY AND DATA

3.1 Latent Dirichlet Allocation

In the presented study, topic modeling is used to extract customer preferences from online reviews. Topic modeling allows the user to discover and summarize latent semantic structures in large volumes of text. Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) is an extension of Latent Semantic Indexing, and it is one of the most popular methods of topic modeling. LDA is an unsupervised clustering technique capable of extracting topics of semantically related words. It is used on the Bag-of-Word representation of documents, and it assumes a generative process about how the documents of a particular corpus are created. In the generative process, documents are assumed to be initially empty. Then topics and the corresponding topic words are iteratively assigned to the documents until every document is created. Using LDA means inverting this assumed generative process, to obtain the hidden topics. For the inversion, the variational inference approach is implemented in this study, as it is introduced in the original paper (Blei *et al.*, 2003).

Since probabilities are calculated for all words in the corpus for each topic, all words appear in every topic, just with different probability values. Accordingly, topic distributions are assigned to every document in the corpus. As the user must decide the number of topics, several LDA models are built (between 5 and 99 topics with the step size of 2). The final model is chosen based on the models' *topic coherence*, introduced in the next Section.

3.2 Topic Coherence

Topic coherence is an approach to evaluate a single topic through assessing the similarity of the semantics amongst the high scoring words within the topic (Stevens *et al.*, 2012). This will help in differentiating the topics that are interpretable semantically from the ones that are used as the artefacts of statistical inference. There are various techniques for coherence metrics, however, in this study, we implemented the *Cv* measure as suggested by Röder *et al.* (2015). There are four parts of *Cv* calculation, 1) the data is segmented into word pairs, 2) for each pair of words or a single word, their probabilities is calculated, 3) confirmation measure is calculated to examine the support strength to another set of words, 4) overall coherence score is calculated. *Cv* has proven to be performing better than pointwise mutual information (PMI) and shows better correlation in regard to human topic ranking data (Bouma, 2009).

3.3 Data

Data for this study was obtained from the Computer Science Department at the University of California San Diego (He and McAuley, 2016). The Amazon review data set contains 5,789,920 reviews, posted from 1996 to 2014. In the data, the variable "reviewText" was used.

Prior to building the model, data was pre-processed to clean it from unnecessary words, punctuations, and special characters. The purpose is to get rid of unwanted information that may disturb the training process which may affect the result. Following the process is to transform each sentence into a single word format. Then, each word will be lemmatized - meaning that it is converted into its dictionary form - depending on the type of the words, either noun, adjective, verb, or adverb. These words are stored into bag-of-words model for training purposes. Words that occur in less than 5 documents and appear in more than 80 percent of the documents are removed. Hence, the result will only focus on the words that are meaningful and relevant to the generated topics.

4. RESULTS AND DISCUSSION

According to the *Cv* score, our best representative topics fall in the total number of 51 topics with its score 0.602, Figure 8, across the reviews. Following that, we visualise our best result using LDAvis - an interactive visualisation of topics in the form of a web system, Figure 9. This visualisation technique allows us to see the topics in a global view and observe how distinct they are from each other. In the LDAvis, the right part shows the frequencies of each word appearing in the documents as a bar chart. The blue coloured bars denote the word frequencies in the overall documents whilst the red coloured bars denote the word frequencies in the documents related to a particular topic. In addition to that, through comparing the width between the red bar and blue bar, users can instantly recognise whether the term is highly exclusive for the selected topic. LDAvis also allow to flexibly rank the words depending on the usefulness of topics for interpretation purposes (Chuang *et al.*, 2012). Words that are most related to the topic is displayed in the Figure 4 along with the specified category. From the results, we can learn that most customers are commenting on accessories, outfit, quality, and appearance when they are shopping cloths online. Visualisation of the topic is created using the pyLDAvis that is available in gensim that is shown in Figure 3.

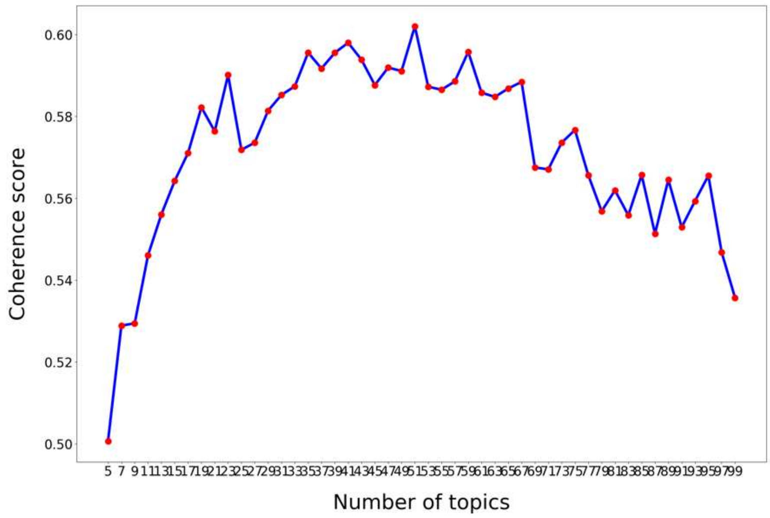

Fig. 2. Coherence measure to find the most representative number of topics

LDAvis applies two other measures for identifying terms usefulness to understand the topics, *distinctiveness* and *saliency* based on Chuang *et al.* (2012). Both will examine the weight of information conveyed from each term through Kullback-Liebler divergence calculation of the marginal distribution of topics (distinctiveness)—presented as bubble—and the topic distribution given the term, and further calculates the saliency—weighted by the total frequency of the term. Furthermore, the visualisation of topic shows the inter-topic differences by computing the inter-topic distances using Jensen-Shannon divergence (Sievert and Shirley, 2014). The default scaling is using the Principal Components for 2D visualization.

We analyse our results by picking up one term for each topic that is the most relevant for a specific topic. For instance, as presented in the Figure 3, Topic 51 is selected. We picked the word "hand" as it shows the highest relevancy compared to other terms with a relatively high lift. Then, categorisation of topics is based on the terms that occupied in the four quadrants, Figure 4. In the quadrant I, the terms involved are "bracelet", "bag", "shoe", "tie", "heel", "watch", and so forth. Since most of terms are describing additional items, we named the quadrant I into Accessories. In the quadrant II, the terms involved are "suit", "coat", "dress", "shirt", "jean", "short", and so forth. Most of these words are talking about clothing, hence we named this quadrant into Outfit. In the quadrant III, the terms involved are "review", "return", "price", "size", "wear", "find", and so forth. These terms are mostly describing about the quality, hence we named it Quality. In the last quadrant, the terms involved are "love", "cute", "product", "look", "good", "colour", and so forth. These terms are mostly consisting of adjectives on how the items appear, hence we named the quadrant IV into Appearance.

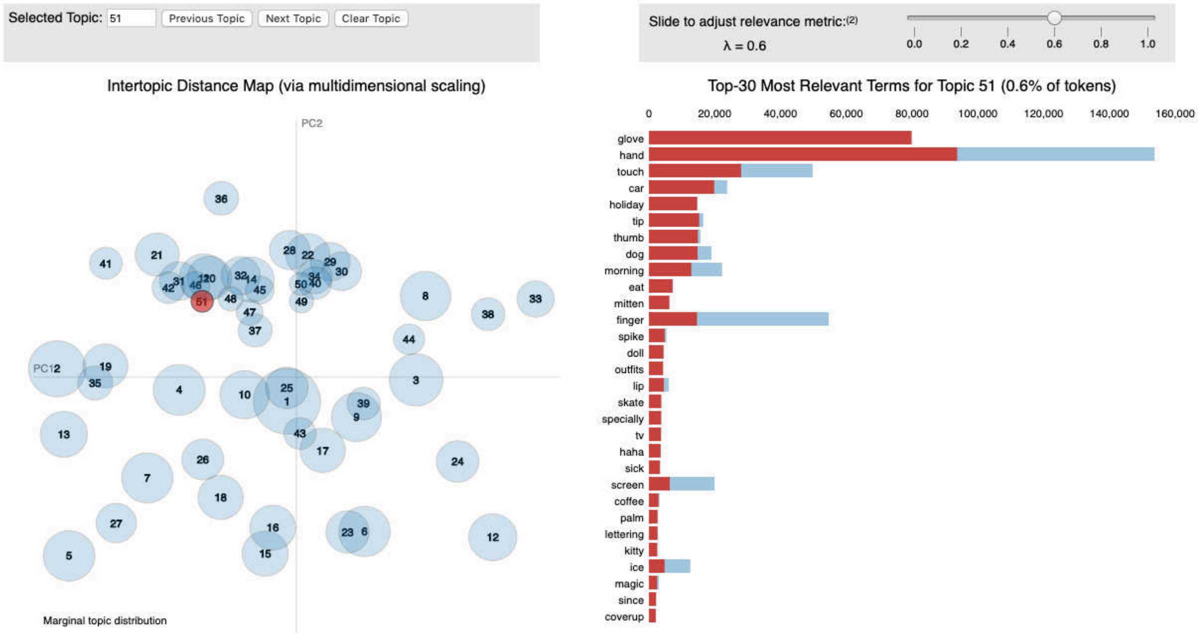

Fig. 3. Visualisation of 51 number of topics

As the result shows, the four identified categories can be seen as items that most consumers buy online—Accessories and Outfit—as well as it requirements—Quality and Appearance. Most customers would prefer items related to clothing such as shirts, coats, dresses, and jeans, or accessories such as bags, watches, and bracelets whilst also considering the quality of the goods, for instance the material it is made from and its appearance (whether or not it looks cute or lovely). Thus, it would benefit and improve fashion businesses in account for product development. As having a creative and unique product is quite essential in the fashion business, these topics could further improve the requirement elicitation. Based on these topics (results) obtained, companies could create a cognitive map as a strategic options development tool to structure the customer demands and feedbacks gained from the products purchased which further can result in better decision making for the companies, for instance, if to prioritise expansion of a certain product segmentation or to do more promotion for a wider consumer exposure. Furthermore, these topics could also help businesses to better describe their products, hence making the product to become more transparent to the consumers. Having the right information written in the product description is necessary not only to the targeted consumers but also in general, since online purchases are really depend on the cyberspace appearance that includes pictures and the quality of information provided (Lohse and Spiller, 1998; Kolesar and Galbraith, 2000). Also, these topics obtained could also be used to transform the way products being grouped on e-commerce websites.

On the other hand, Authors do expect the result to appear corresponding to fashion items and requirements from the consumers as shown in the Figure 4. However, we were also expecting to see in terms of how the online shopping experience within the fashion segment have been going so far, since it could also be one of the consideration when consumer decides to make an online purchase. Nevertheless, online reviews were intended to give an overall illustration on how satisfied a customer is.

Further improve could be done to achieve better results and a more accurate evaluation. For instance, on this study, the method was not tested in an industrial environment. Therefore, the usability of the approach is not evaluated. Correspondingly, domain experts would be needed to analyse the results and extract the non-trivial topic words that would help in requirement elicitation.

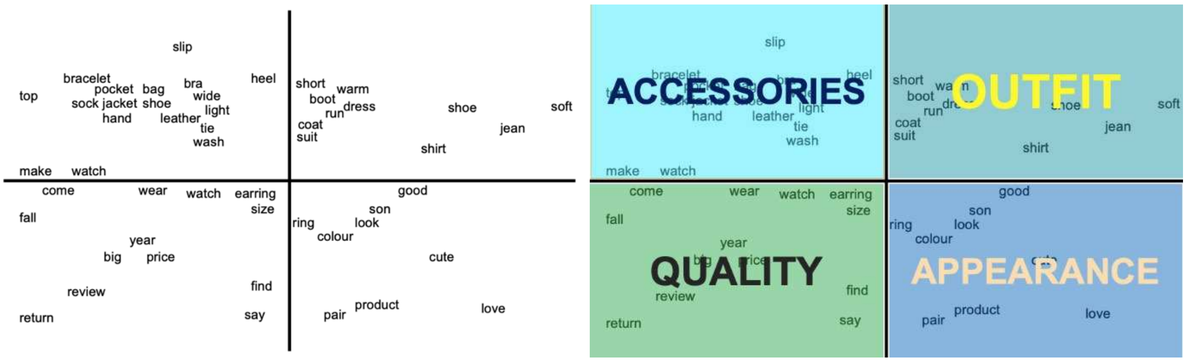

Fig. 4. Categorization of topics

5. CONCLUSIONS

In this paper, we gained information from extracting groups of words from a dataset which could be potentially used to analyse the consumer preferences and consumption tendency by implementing topic modeling with LDA for Amazon customer review data. The data used contained 5,789,920 reviews from 1996 to 2014. Our goal was to reveal hidden topics to explore customer preferences and customer needs. Results shows that the domain is best represented with the total number of 51 topics. Furthermore, we visualise our result using LDAvis. The analysis presents that there are four major categories which customer seeks when they are shopping online; Accessories, Outfit, Quality, and Appearance. Our research contributes on two major aspects. First, topics extracted gave information on what the customers desire or need when looking for fashion items in online stores. Second, through applying the topic modeling in customer reviews, companies could possibly measures the overall customer satisfaction rate based on the product purchased and/or the whole online shopping experience (i.e. delivery service, product description, customer service, and so forth). However, this study has not been tested in an industrial environment, hence the usability of this approach cannot be verified.

The present study can be improved in several respects. First of all, in this study we only focus on revealing the latent topics given the textual information. Future research can expand the current study by applying predictive, so we can investigate or forecast the fashion trends. Furthermore, as there is a huge usage of adjective within the reviews, sentiment analysis could be conducted on the reviews to see what product receives more positive or negative reviews.

REFERENCES

Aspers, P., and Godart, F. (2013). Sociology of fashion: Order and change. *Annual Review of Sociology*, 39, 171-192.

Berawi, M.A. (2018). The Fourth Industrial Revolution: Managing Technology Development for Competitiveness. *International Journal of Technology*, 9(1), 1.

Bischof, J.M., and Airoldi, E.M. (2012). Summarizing topical content with word frequency and exclusivity, In: *Proceedings of the 29th International Conference on Machine Learning*, Langford, J., and Pineau, J. (Eds.), 1-8. International Machine Learning Society: Edinburgh, Scotland, UK.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77.

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction, In: *Proceedings of the Biennial GSCL Conference 2009*, Chiarcos, C., de Castilho R.R., and Stede, M. (Eds.), 43-53. Gunter Narr Verlag: Tübingen.

Brown, B., Chul, M., and Manyika, J. (2011). Are you ready for the era of' big data'? *McKinsey Quarterly*, 4, 24-27, 30-35.

Bughin, J., Chui, M., and Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech- enabled business trends to watch. *McKinsey Quarterly*, 4, 26-43.

Chuang, J., Manning, C.D., and Heer, J. (2012). Termite: visualization techniques for assessing textual topic models, In: *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI 12*, Tortora, G., Levialdi, S., and Tucci, M. (Eds.), 74-77. ACM: New York NY.

Duan, W., Gu, B., and Whinston, A.B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, 45(10), 1407-1424.

Engler, T.H., Winter, P., and Schulz, M. (2015). Understanding online product ratings: a customer satisfaction model. *Journal of Retailing and Consumer Services*, 27, 113–120.

He, R., and McAuley, J. (2016). Ups and Downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, In: *WWW16: Proceedings of the 25th International Conference on World Wide Web*, Bourdeau, J., and Hendler, J.A. (Eds.), 507-517. International World Wide Web Conferences Steering Committee: Geneva, Switzerland.

Heng, Y., Gao, Z., Jiang, Y., and Chen, X. (2018). Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, 42, 161–168.

Jain, S., Bruniaux, J., Zeng, X., and Bruniaux, P. (2017). Big data in fashion industry. *IOP Conference Series: Materials Science and Engineering*, 254(15), 1-6.

Jo, Y., and Oh, A.H. (2011). Aspect and sentiment unification model for online review analysis, In: *Proceedings of the fourth ACM international conference on Web search and data mining*, King, I., Nejdl, W., and Li, H. (Eds.), 815-824. ACM: New York NY.

Kolesar, M.B., and Galbraith, R.W. (2000). A services marketing perspective on e-retailing: implications for e-retailers and directions for further research. *Internet Research*, 10(5), 424-38.

Lohse, G.L., and Spiller, P. (1998). Electronic shopping. *Communications of ACM*, 41(7), 81-89.

McAfee, A., and Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10), 62-68.

Nielsen.com. (2015). The Future of Grocery. Retrieved January 20, 2020, from https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/nielsen-global-e-commerce-new-retail-report-april-2015.pdf

Ohlhorst, F.J. (2012). What is Big Data? In: *Big Data Analytics*, 1-10. Wiley: Hoboken NJ.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures, In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM 15*, Cheng, X., Li, H., Gabrilovich, E., and Tang, J. (Eds.), 399-408. ACM: New York NY.

Sievert, C., and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics, In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, Chuang, J., Green, S., Hearst, M., Heer, J., and Koehn, P. (Eds.), 63-70. Association for Computational Linguistics: Baltimore MD.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring Topic Coherence over many models and many topics, In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Tsujii, J., Henderson, J., and Pasca, M. (Eds.), 952-961. Association for Computational Linguistics: Stroudsburg PA.