Hey Siri. Ok Google. Alexa:
A topic modeling of user reviews for smart speakers

Hanh Nguyen
Bocconi University
hn@hanhng.com

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Abstract

User reviews provide a significant source of information for companies to understand their market and audience. In order to discover broad trends in this source, researchers have typically used topic models such as Latent Dirichlet Allocation (LDA). However, while there are metrics to choose the "best" number of topics, it is not clear whether the resulting topics can also provide in-depth, actionable product analysis. Our paper examines this issue by analyzing user reviews from the Best Buy US website for smart speakers. Using coherence scores to choose topics, we test whether the results help us to understand user interests and concerns. We find that while coherence scores are a good starting point to identify a number of topics, it still requires manual adaptation based on domain knowledge to provide market insights. We show that the resulting dimensions capture brand performance and differences, and differentiate the market into two distinct groups with different properties.

1 Introduction

The Internet has provided a platform for people to express their opinions on a wide range of issues, including reviews for products they buy. Listening to what users say is critical to understanding the product usage, helpfulness, and opportunities for further product development to deliver better user experience. User reviews – despite some potentially inherent biases[1] – have quickly become an invaluable (and cheap) form of information for product managers and analysts (Dellarocas, 2006). However, the speed, amount, and varying format of user feedback also creates a need to effectively extract the most important insights.

Topic models, especially LDA (Blei et al., 2003), are one of the most widely used tools for these purposes. However, due to their stochastic nature, they can present a challenge for interpretability (McAuliffe and Blei, 2008; Chang et al., 2009). This is less problematic when the analysis is exploratory, but proves difficult if it is to result in actionable changes, for example product development. The main dimension of freedom in LDA is the number of topics: while there are metrics to assess the optimal number according to a criterion, it is unclear whether the resulting topics provide us with a useful discrimination for product and market analysis. The question is "*Can we derive market-relevant information from topic modeling of reviews?*"

We use smart speakers as a test case to study LDA topic models for both high-level and in-depth analyses. We are interested in to answer the following research questions:

- What are the main dimensions of concerns when people talk about smart speakers?
- Can the LDA topic mixtures be used to directly compare smart speakers by Amazon, Google, Apple, and Sonos?

Smart speakers are a type of wireless speaker that provides a voice interface for people to use spoken input to control household devices and appliances. While still relatively new, smart speakers are rapidly growing in popularity. As the Economist (2017) put it: "voice has the power to transform computing, by providing a natural means of interaction." We use a dataset of smart speaker reviews and coherence scores as a metric to choose the number of topics, and evaluate the resulting model both in terms of human judgement and in its ability to meaningfully discriminate brands in the market.

[1] People tend to over-report negative experiences, while some positive reviews are bought (Hovy, 2016).

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 76–83
Hong Kong, Nov 4, 2019.

	Raw data	After pre-processing
# reviews	53,273	
# words	1,724,842	529,035
# unique words	25,007	10,102

Table 1: Summary of dataset.

Contributions We show that LDA can be a valuable tool for user insights: 1) basic user concerns can be distinguished with LDA by using coherence scores (Röder et al., 2015) to determine the best number of topics, but an additional step is still needed for consolidation; 2) human judgement correlates strongly with the model findings; 3) the extracted topic mixture distributions accurately reflect the qualitative dimensions to compare products and distinguish brands.

2 Dataset

2.1 Data collection

From the Best Buy US website, we collect a dataset of 53,273 reviews for nine products from four brands: **Amazon** (Echo, Echo Dot, Echo Spot), **Google** (Home, Home Mini, Home Max), **Apple** (HomePod) and **Sonos** (One, Beam). Each review includes a review text and the brand associated with it. Our collection took place in November 2018. Due to their later market entries and significantly smaller market sizes, the number of available Apple and Sonos reviews is limited. Amazon, Google, Apple, and Sonos reviews account for 53.9%, 41.1%, 3.5% and 1.5% of the dataset, respectively.

2.2 Review text pre-processing

We pre-process the review text as follows: First, we convert all text to lowercase and tokenize it. We then remove punctuation and stop words. We build bigrams and remove any remaining words with 2 or fewer characters. Finally, we lemmatize the data. The statistics of the resulting bag-of-words representation are described in Table 1.

3 Methodology

3.1 Topic extraction

The main issue in LDA is choosing the optimal number of topics. To address this issue, we use the coherence score (Röder et al., 2015) of the resulting topics. This metric is more useful for interpretability than choosing the number of topics on held-out data likelihood, which is a proxy and can still result in semantically meaningless topics (Chang et al., 2009).

The question is: what is coherent? A set of topic descriptors are said to be coherent if they support each other and refer to the same topic or concept. For example, "music, subscription, streaming, spotify, pandora" are more coherent than "music, machine, nlp, yelp, love." While this difference is obvious to human observers, we need a way to quantify it algorithmically.

Coherence scores are a way to do this. Several versions exist, but the one used here has the highest correlation with human ratings (Röder et al., 2015). It takes the topic descriptors and combines four measures of them that capture different aspects of "coherence": 1) a segmentationn S^{one}_{set}, 2) a boolean sliding window $P_{sw(110)}$, 3) the indirect cosine measure with normalized pointwise mutual information (NPMI) $\tilde{m}_{\cos(nlr)}$, and 4) the arithmetic mean of the cosine similarities σ_a.

The input to the scoring function is a set W of the N top words describing a topic, derived from the fitted model. The first step is their segmentation S^{one}_{set}. It measures how strongly W^* supports W' by quantifying the similarity of W^* and W' in relation to all the words in W:

$$\left\{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\right\}$$

In order to so do, W' and W^* are represented as context vectors $\vec{v}(W')$ and $\vec{v}(W^*)$ by pairing them with all words in W:

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} \text{NPMI}(w_i, w_j)^\gamma \right\}_{j=1,\ldots,|W|}$$

The same applies for $\vec{v}(W^*)$. In addition:

$$\text{NPMI}(w_i, w_j)^\gamma = \left(\frac{\log \frac{P(w_i,w_j)+\varepsilon}{P(w_i)\cdot P(w_j)}}{-\log(P(w_i, w_j) + \varepsilon} \right)^\gamma$$

An increase of γ gives higher NPMI values more weight. ε is set to a small value to prevent logarithm of zero. We choose $\gamma = 1$ and $\varepsilon = 10^{-12}$.

Second, the probability $P_{sw(110)}$ captures proximity between word tokens. It is the boolean sliding window probability, i.e., the number of doc-

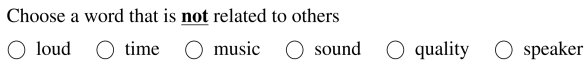

Figure 1: Example of word intrusion task in the survey

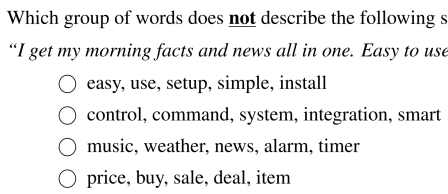

Figure 2: Example of topic intrusion task in the survey

uments in which the word occurs, divided by the number of sliding windows of size $s = 110$.

Third, given context vectors $\vec{u} = \vec{v}(W')$ and $\vec{w} = \vec{v}(W^*)$ for the word sets of a pair $S_i = (W', W^*)$, the similarity of W' and W^* is the cosine vector similarity between all context vectors.

$$s_{\cos}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2}$$

Finally, the cosine similarity measures are averaged, giving us a single coherence score for each model (each model has a different number of topics).

We fit LDA models, using Gensim library in Python, with the number of topics ranging from 2 to 20 to calculate the coherence score. For each model, we choose the top 20 words of each topic as inputs to calculate the model's coherence score. We move forward with the model with the highest coherence score (13 topics) for validation, and use the document-topic distributions and topic-word distributions from that model in the subsequent steps.

3.2 LDA validation

To evaluate the semantic interpretability of the resulting LDA model from the coherence score selection, we run a human judgment survey using word intrusion and topic intrusion. We used 125 human judges. Each of 125 human subjects responds to 10 questions (5 questions for word intrusion, and 5 questions for topic intrusion), which are randomly selected from a collection of 20 questions.

For the **word intrusion** task, each subject is asked to choose which word they think does not belong to the topic (Fig. 1). Each question is comprised of the 5 words with the highest probabilities in that topic, and one random word with low probability in that topic but high probability (top 5 most frequent words) in another topic. The word that does *not* belong to the topic is called the *true intruder word*. The hypothesis of word intrusion is that if the topics are interpretable, they are coherent, and subjects will consistently choose the true intruder words.

For topic k, let w_k be the true intruder word, $i_{k,s}$ be the intruder selected by the subject s. S is the number of subjects. The model precision for topic k is defined as the fraction of subjects agreeing with the model:

$$MP_k = \frac{\sum_s |(i_{k,s} = w_k)|}{S}$$

The model precision ranges from 0 to 1, with higher value indicating a better model.

For the **topic intrusion** task, each survey subject is shown a short review text and is asked to choose a group of words which they think do *not* describe the review (Fig. 2). Each group of words represents a topic. Each question is comprised of 3 topics with the highest probabilities LDA assigned to that review, and 1 random topic with low probability. The topic with low probability is called the *true intruder topic*. The hypothesis of topic intrusion is that if the association of topics to a document is interpretable, subjects will consistently choose the true intruder topic.

For review r, let j_r be the true intruder topic, $j_{r,s}$ be the intruding topic selected by subject s. θ_r is the probability that the review r belongs to each topic. The topic log odds for a review r are defined as the log ratio of a) the probability mass assigned to the true intruder to b) the probability mass assigned to the intruder selected by the subject:

$$TLO_r = \frac{\sum_s (\log \theta_{r,j_r} - \log \theta_{r,j_{r,s}})}{S}$$

The topic log odds have an upper bound of 0, which indicates the perfect match between judgments of the model and the subjects. This metric is preferred for the topic intrusion task rather than the model precision, which only takes into account right or wrong answers, because each topic has a probability of generating the review. Thus, the topic log odds serve as an error function (Lukasiewicz et al., 2018).

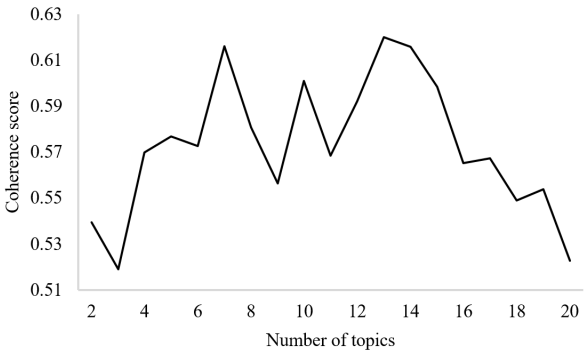

Figure 3: Coherence score for each model. Models with 7, 13, and 14 topics have highest coherence score.

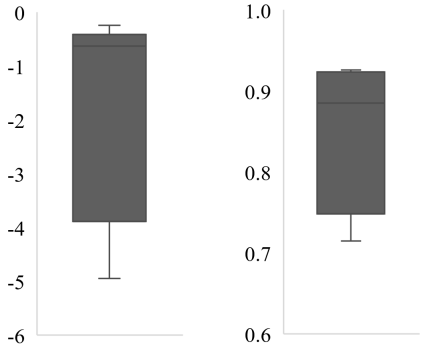

Figure 4: Model precision with word intrusion (left) and topic log odds with topic intrusion (right).

4 Results and discussion

4.1 Topic extraction

For each model, we compute topic coherence based on the top 20 words in each topic. The topic coherence plot (Fig. 3) shows three candidates for the best number of topics, 7, 13, and 14, all with a score of 0.62. We manually examine the top 20 words for each. The 7-topic model has some mixed and chained topics e.g., *"easy, use, great, setup, gift, christmas."* The 14-topic model does not provide any more meaningful topics compared to 13 topics. Thus, we choose the 13-topic model.

4.2 LDA validation and consolidation

We extract document-topic and topic-word distributions for 13 topics and evaluate them in a human judgment survey on word and topic intrusion. The mean of the word intrusion precision is 0.85 (standard deviation 0.086), and the mean of the topic log odds is -1.66 (standard deviation 1.58). Fig. 4 shows the box plots for the results of both tasks. Model precision and topic log odds are on different scales, see section 3.2. Model precision is sufficiently good, while topic log odds are acceptable, but with higher variance. They are on a par with the best models in (Chang et al., 2009; Arnold et al., 2016).

Reviews dominated by few topics show more agreement between model and human judges, reviews with many topics show a greater divergence. For example, for the review with the lowest level of agreement (lowest topic log odds): *"Once I managed to get all the apps synced with this speaker, I was blown away by the sound quality. Using Alexa's voice recognition is great, even from the other side of the room."*, LDA assigns fairly equal proportions to the top 3 topics (23%, 25%, and 32%). For the review with the highest level of agreement (highest topic log odds): *"I get my morning facts and news all in one easy to use system."*, LDA assigns 48% to a dominant topic, and 15% and 26% to the next two topics.

After running the intrusion tests with the 13-topic model, we manually merge some topics that were similar to each other. This process results in 8 dimensions (we call them "dimensions" to differentiate them from the 13-topic model of the previous steps). We use these 8 dimensions to measure brand performance.

As (Boyd-Graber et al., 2014) pointed out, different researchers might combine topics differently. Here, the merging step is based on our domain knowledge in the smart speaker market. We group topics with similar top words into one dimension. For topics that we cannot label, we group them to the most similar topics based on the top words. Doing so, we aim to make the topics maximally distinguished from each other, and to be able to label the topics appropriately.

Table 2 shows the respective top keywords. The following describes the resulting 8 dimensions.

1. **Price**: price and worthiness, especially as gifts. Example: *"Love my Echo Dot, great purchase! Made a great Christmas gift."* (Amazon)

2. **Integration**: ability to connect, and control devices/household appliances (e.g., lighting, thermostat) in a smart home. Bedroom and kitchen are the two rooms in which people put their smart speakers most often. Example: *"I use these in several rooms in my home to control lights and my AV system. They integrate with my Samsung Smart Things Hub*

Label	Top keywords
Price	price, buy, gift, christmas, worth, black_friday, money, sale, deal, item
Integration	light, control, command, system, integration, thermostat, room, ecosystem, connect
Sound quality	speaker, sound, quality, small, music, loud, great, room, bluetooth, volume
Accuracy	question, answer, time, response, quick, issue, problem, work, search, good
Skills	music, weather, news, alarm, timer, kitchen, morning, reminder, shopping_list
Fun	fun, family, kid, useful, helpful, great, friend, game, information, question
Ease of use	easy, use, set, setup, simple, install, recommend, connect, quick, work
Playing music	music, play, song, playlist, favorite, pandora, prime, stream, subscription, beam

Table 2: 8 merged dimensions and the keywords reveal how people use smart speakers and their perceptions.

and Harmony Hub." (Amazon)

3. **Sound quality**: ability to provide high-quality sound. Example: *"Can't believe this little device has such great sound quality"* (Apple). *"This is a great speaker! The sound is just WOW! And the speaker doesn't take up much space."* (Sonos)

4. **Accuracy**: ability to respond accurately to the users voice commands, provide answers to questions, and to issues they might encounter. Example: *"It is amazing how many simple questions stump Alexa. Too frequently the response I hear is "I don't understand" or "Sorry, I can't find the answer.""* (Amazon)

5. **Skills**: variety of applications that the smart speaker provides. They are referred to as "skills" in Amazon Alexa, and as "actions" in Google Assistant. I.e., music, weather forecast, news, alarms, setting kitchen timers, reminders, and shopping lists. Example: *"You can ask Alexa anything. Find information about the weather, sports, or the news. Also, ask her to play your favorite music. All you have to do is ask."* (Amazon)

6. **Fun**: pleasure to interact with smart speakers, especially with/for kids and family. Example: *"Lots of fun and lots of great information. It was fun to ask it all kinds of questions."* (Google)

7. **Ease of use**: ease of setup and connecting to an existing internet connection via the mobile app to use voice commands. Example: *"Fun and easy to operate. Connects to your Wi-Fi in a simple and quick manner."* (Amazon)

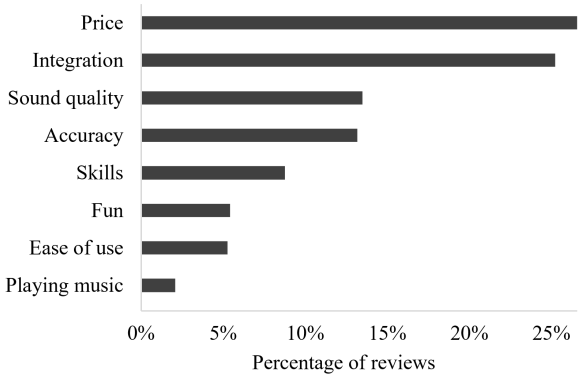

Figure 5: % of reviews based on dominant dimensions.

8. **Playing music**: ability to play music, connect with music services, like Amazon Music and Pandora. Example: *"Upload all of your music for free to Google Play Music and then tell the Mini what to play from your music and it will!"* (Google)

Since the LDA model can assign a review to multiple topics, it is more difficult to see the proportion of reviews for each. We define the *dominant dimension* for each review as the topic with the highest probability for the review. The most frequently mentioned dominant dimensions (Fig. 5) are *price* (27% of total reviews), *integration* (25%), *sound quality* (14%), and *accuracy* (13%).

4.3 Brand performance along dimensions

Brand performance measures how frequently each dimension was mentioned in user reviews.

As described in section 2.1, the amount of available data across companies is highly imbalanced. Thus, in order to compare the relative performance of brands along the 8 dimensions, we normalize the amount of data for each company. We define a relative dimension score for a brand b (Ama-

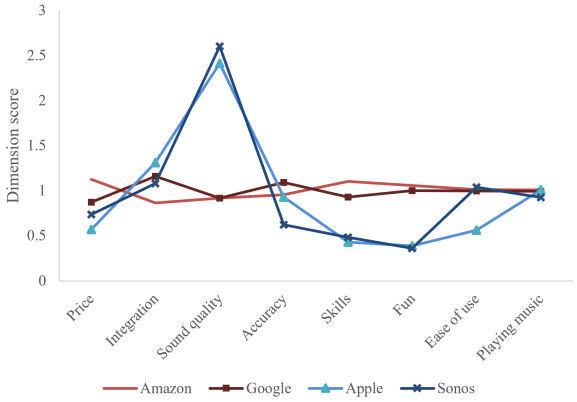

Figure 6: Company profiles along 8 dimensions form 2 groups with similar positioning.

zon, Google, Apple Sonos) along a dimension d_k ($k \in [1, 8]$) as the normalized topic probability:

$$DS_{b,d_k} = \frac{{}^{1}/_{N_b} \sum_{r=1}^{N_b} p_{r,d_k}}{{}^{1}/_{N} \sum_{r=1}^{N} p_{r,d_k}}$$

p_{r,d_k} is the probability that review r belongs to dimension d_k. N_b is the number of reviews for brand b. N is the total number of reviews for all brands.

The line plot in Fig. 6 reveals some interesting differences between the brands' relative strengths and weaknesses.

Amazon and Google speakers are similar to each other, with a balanced performance on all dimensions. On the other hand, Apple and Sonos speakers are also similar to each other, but with a focus on sound quality. This suggests a segmentation of the smart speaker market into two groups along those lines.

Apple and Sonos clearly outperform Amazon and Google speakers in terms of sound quality. Indeed, both Apple and Sonos speakers are high-end products, arguably the best sounding smart speakers on the market, using, e.g., adaptive audio (beamforming) to determine the position of a user and adjust its microphones accordingly. Sonos has digital amplifiers, a tweeter, a woofer, and a 6-microphone array, and an adaptive noise suppression algorithm.

Interestingly, Amazon and Google users mention using their speakers to listen to music as much as Apple and Sonos users do. This is most likely due to the fact that playing music is the most popular task on every smart speaker. However, it does suggest that only a few people are willing to pay extra for better sound quality, and that they do greatly appreciate sound quality and mention it often.

Amazon performs best in term of price, followed by Google. Users mention that prices are reasonable, and many people buy it as a gift for Christmas or during sales such as Black Friday. Amazon speakers do have the lowest prices among the 4 brands (Amazon: $49.99, Echo 2nd Gen: $99.99, Echo Spot: $129.99). Google's high-end speaker, the HomeMax ($399.00) is much less popular than its Home Mini ($49.00) and Home ($129.00). The main competition in terms of price and gift is between Amazon Echo Dot ($49.99) and Google Home Mini ($49).

For skills, Amazon/Google perform better than Apple/Sonos. Siri is strictly limited to Apple's ecosystem (e.g., users can only stream music from Apple Music, not from Spotify). This is potentially interesting for Sonos to distinguish themselves, as the speakers are Alexa-enabled (as of November 2018 when the reviews were collected), so users could exploit its skills just like Amazon users. One possible explanation could be that Sonos users focus more on music and sound quality, and that other skills become less important to them so they mention other skills less often.

5 Related work

Several topic models have been proposed, such as Latent Semantic Indexing (LSI) (Deerwester et al., 1990), Probabilistic LSI (pLSI) (Hofmann, 1999), and the most commonly used, Latent Dirichlet Allocation (Blei et al., 2003). LDA assumes that a document is comprised of mixtures over latent topics, and each topic is a distribution over words.

LDA has some limitations. The main limitations are the assumption that the number of topics is known and fixed, together with the validity of the assignments, and the interpretability of topics. LDA evaluation schemes can be categorized into *intrinsic evaluation* (holdout-log likelihood/ perplexity (Blei et al., 2003; Wallach et al., 2009), *topic coherence* (Newman et al., 2010; Röder et al., 2015), *human-in-the-loop* (word or topic intrusion (Chang et al., 2009; Lau et al., 2014)), and *extrinsic evaluation* (e.g., document clustering (Jagarlamudi et al., 2012), information retrieval (Wei and Croft, 2006)). Those work mainly focus on extracting meaningful high-level topic descriptors.

In this paper, we show that those techniques, when combined appropriately together, are useful in not only high-level topics but also in-depth insights from data. In order to do so, we address LDA limitations with topic coherence, human-in-the-loop, and incorporating human knowledge to merge topics for better quality (Boyd-Graber et al., 2014).

6 Conclusion

In this paper, we use the coherence score by Röder et al. (2015) as a guide to choose the optimal number of topics, and evaluate this choice with respect to human judgement and its ability to provide market insights. While coherence scores are judged meaningful (in word intrusion and topic intrusion) and provide a good starting point, they require and additional merging step based on domain knowledge to provide market insights. We merge the optimal choice of 13 topics into 8 dimensions for easier interpretation. We show that the topic mixture proportions are useful to give more insights about brand performance and market structure, separating the brands into two distinct camps with similar properties. Further research directions could assess the generalizability of the methodology on other datasets and tasks.

Acknowledgments

Hanh Nguyen is grateful to Raffaella Piccarreta for the inspiration that led to this work. The authors would like to thank Barbara Plank for useful feedback, as well as the reviewers for their helpful comments to make the paper clearer. Dirk Hovy is a member of the Bocconi Institute for Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit.

References

Corey W. Arnold, Andrea Oh, Shawn Chen, and William Speier. 2016. Evaluating topic model interpretability from a primary care physician perspective. *Computer Methods and Programs in Biomedicine*, 124:67–75.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and Their Applications*, 225255.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Ssociety for Information Science*, 41(6):391–407.

Chrysanthos Dellarocas. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10):1577–1593.

Economist. 2017. How voice technology is transforming computing - now we are talking. *The Economist*, 422(Jan):7.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Dirk Hovy. 2016. The enemy in your own camp: How well can we detect statistically-generated fake reviews – an adversarial study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 351–356, Berlin, Germany. Association for Computational Linguistics.

Jagadeesh Jagarlamudi, Hal Daume III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wojciech Lukasiewicz, Alexandru Todor, and Adrian Paschke. 2018. Human perception of enriched topic models. In *Business Information Systems*, pages 15–29. Springer International Publishing.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems*, pages 121–128. Curran Associates, Inc.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM Press.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM Press.

Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press.