LDA Topic Analysis for Product Reviews in Social Media Platform

Farrikh Alzami, Rama Aria Megantara, Dwi Puji Prabowo, Puri Sulistiyawati, Ricardus Anggi Pramanendri, Ika Novita Dewi, Ifan Rizqa, Ritzkal

1Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia
2Department of Informatics Engineering, Ibn Khaldun University, Bogor, Indonesia

E-mail: aria@dsn.dinus.ac.id

ABSTRAK

In recent years, social media has evolved into a prominent platform for the assessment of products and the collection of customer opinions. Consequently, in our quest to gain insights into the subjects that have been deliberated upon, we employed the technique of Latent Dirichlet Allocation (LDA) for the purpose of topic modeling. The significance of this topic modeling lies in its capacity to uncover valuable information pertaining to specific aspects of a product that resonate with or displease customers. To maintain the integrity of the temporal context, we refrained from employing stop words, a decision that can potentially eliminate valuable chronological details. This corpus of information holds considerable value for businesses, as it provides them with a window into consumer preferences. This, in turn, enables them to make informed decisions regarding product development and marketing strategies. With a calculated coherence score of 0.621520, our LDA-based topic modeling process yielded three optimal topics. Topic 0 revolved around discussions concerning price and the availability timeline of products. In Topic 1, the focal point was the difficulty of obtaining a product due to its absence in the market. Finally, Topic 2 delved into post-usage ownership and the aspects of a product that customers appreciate.

Introduction

Social media has transformed the landscape for businesses and consumers by becoming a pivotal platform for product reviews and customer feedback. Millions of people use social media channels to share their experiences, opinions, and thoughts on various products, services, and brands. These reviews provide a valuable source of quantitative and qualitative data for businesses (Adak et al., 2022). Thus, the social media has provided researchers and businesses with a wealth of textual data that can be analyzed to gain valuable insights into customer preferences, opinions, and sentiments (W. Liu et al., 2019; Sarker & Roy, 2020).

One popular method used in the analysis of social media comments and reviews is Latent Dirichlet Allocation (LDA) topic modeling (Kim et al., 2017; B. Liu et al., 2021, p. 202; Zhao et al., 2015). Latent Dirichlet Allocation is a widely used unsupervised probabilistic topic modeling technique that assumes each document is composed of a set of topics. By applying LDA to social media comments and reviews, researchers can uncover the underlying themes and topics discussed by users (Andy & Andy, 2021). This method helps in finding hidden relationships among text documents and discovering topics among them (Rüdiger et al., 2022). Using LDA topic analysis in social media research allows for a comprehensive understanding of discussions and reactions within different social media sites (Gupta et al., 2021).

LDA topic analysis using social media comments and reviews for product reviews has gained traction in recent years. This approach helps researchers extract valuable information from large volumes of social media data, enabling them to identify the main topics and themes that users discuss in product reviews. One successful application of LDA topic modeling in social media research is the extraction of product features from online reviews (Park et al., 2020; Shi et al., 2021)

Moreover, the use of LDA topic analysis in social media comments and reviews for product reviews has several strengths (Khanbhai et al., 2021). Firstly, LDA topic analysis allows for the extraction of key topics and themes from a large collection of text data. This enables businesses to gain a comprehensive understanding of the opinions and sentiments expressed by customers regarding their product reviews, providing valuable insights into customer preferences, opinions, and sentiments.
products or services. Secondly, LDA topic analysis provides insight into the relationships between different reviews and customer sentiment.

Before delving into LDA, we must know the limitation of LDA methods, such as: 1) Regarding content: It doesn't account for the order of words within a document. This means it may fail to capture the context, interpret sarcasm, or identify multi-word expressions that convey specific meaning (Shen et al., 2022); 2) Determining the number of topics can be tricky: The optimal number of topics isn't known beforehand and finding the right one can require several trial-and-error iterations (Song et al., 2020); 3) Text preprocessing requirement: LDA models require significant preprocessing of the text (such as stemming, stop-word removal, tokenization) before they can correctly identify topics (Dorr et al., 2022); 4) Results interpretation: The latent topics generated by LDA may sometimes be difficult to interpret or may not make intuitive sense (S. Park et al., 2021); 5) Model randomness: As an unsupervised method, LDA relies on randomness. This means that the model may produce slightly different results each time it's run with the same parameters and dataset (Wang & Chen, 2022); 6) Failure to capture polysemy: LDA has trouble capturing words that have multiple meanings, because it assumes that each word belongs to a single topic (Garg et al., 2021).

From the research gaps that provided in previous paragraph, we using the Text preprocessing requirement as research contribution in matter of stopwords. Thus, By analyzing the topics and sentiments expressed in social media comments and reviews, researchers (especially company who interested in particular product) can gain insights into the specific features that customers like or dislike about a particular product. This information can be valuable for businesses as it helps them understand customer preferences and inform product development and marketing strategies (Chumwatana, 2018; Micu et al., 2021).

Thus, in this research, we conduct experiment by using data from indonesian youtuber which review the product. The data itself is come from the comments from the viewers which using indonesian language. We believe that the comments from the viewers, have several interesting topics that can be covered using Latent Dirichlet Allocation Model. From the experiment, we successfully obtained 3 optimum topic with coherence point by 0.621520.

Method

Here, we are using Latent Dirichlet Allocation (LDA) which is a probabilistic topic modeling technique used to discover hidden topics within a collection of documents. LDA assumes that documents are mixtures of topics and that topics are mixtures of words. It is widely used in natural language processing and text mining for tasks like document categorization, information retrieval, and content recommendation. Here's a overview of how LDA works:

First. Initialization: Determine the number of topics (K) you want to extract from the document collection. This is a hyperparameter you need to specify beforehand.

Second. Iteration: LDA uses an iterative process to update the document-topic and topic-word distributions. It repeats the following steps until convergence: a. E-step (Expectation): For each word in each document, compute the conditional probability that the word belongs to each of the K topics, given the current topic-word and document-topic distributions. This step calculates the probability of assigning each word to each topic. b. M-step (Maximization): Update the document-topic and topic-word distributions based on the probabilities calculated in the E-step.

Third. Convergence: LDA repeats the E-step and M-step until a convergence criterion is met, such as the number of iterations or when the model parameters no longer change significantly.

Fourth. Output: Once the LDA model has converged, it provides two key outputs: a. The learned document-topic distribution for each document, which indicates the mixture of topics in each document. b. The learned topic-word distribution for each topic, which shows the likelihood of words appearing in each topic.

Fifth. Topic interpretation: After training, LDA can be used to interpret the discovered topics. You can inspect the top words associated with each topic (those with the highest probabilities) to understand the content and meaning of the topics.
Then, this research using quantitative methods, where our object are are youtube channel from Gadgetin as we know is trustworthy youtuber and have many subscriber. The dataset itself is obtained by scraping the viewers comment in the Youtube about smartphone about infinix note 30 indonesia (link: watch?v=PCs-8KTlhKg). The detailed description can be seen as follows:

From figure 1, we can see the detailed procedure as follows:

**First.** We put the url link into https://communalytic.org/ which have abilities to extract the viewers comments;

**Second.** After the comments have been extracted, we proceed to preprocessing data using python library and NLTK library. In here, we are: a. Tokenizing the data, b. using ekphrasis library to sanitize the data (remove username, email, number, url); and c. using lowercase, remove emoji, remove excessive tab and spaces and remove punctuation.

**Third.** In here, we did not use the stopwords, due to it remove the important words. For example as seen at table 1, the stop words remove time feature. Even though topic modeling using Noun, but we want to know, if the time have meaning in our research. Thus. We consider not using stop words in this research.

<table>
<thead>
<tr>
<th>Original text</th>
<th>Tokenize and cleaning</th>
<th>Stop words</th>
</tr>
</thead>
<tbody>
<tr>
<td>komen dulu nonton belakangan</td>
<td>[komen, dulu, nonton, belakangan]</td>
<td>[komen, nonton]</td>
</tr>
<tr>
<td>2 hari yang lalu, dapat di 2.250</td>
<td>[hari, lalu, dapat, di]</td>
<td>[]</td>
</tr>
</tbody>
</table>

**Fourth.** To obtain better result, we are using stemming the data to reduce words to their root or base form, removing suffixes and prefixes. This process helps in reduce noise in text data caused by different word forms. In here we are using Sastrawi library to stemming the data.

**Fifth.** After the data has been stemmed. The next part is move to Topic Modeling using LDA.

In topic modeling itself, we are using steps as figure below:

**Figure 2.** delve down in topic modeling

detailed steps in figure 2 can be explained as follows:

**First.** We are using gensim library due to it have vast abilities in topic modeling.
Second). In gensim library, we are using Phrases model. The tokens (from previous methods) are processed using phrases model. In here, we are using two times phrases as follows: a. Bigram: we set the occurrence a phrase into 5. The threshold is set to 100 due to we will created the phrase where the score is above 100. b. Trigram: the result from bigram, we put into this phrases. In here we set the threshold to 100 due to we will created the bigram phrase where the score is above 100.

Third). Augmentation: This process allows us to augment original tokenized text data with the identified multi-word expressions (bigrams and trigrams) created by the models. These multi-word expressions may capture important combinations of words that can be valuable.

Fourth). Creating dictionary from the augmentation data
Fifth). Creating corpus from dictionary by utilized doc2bow where doc2bow is used to converts each document (a list of tokens) into a bag-of-words representation (BoW). In a BoW representation, each word in the document is represented as a tuple (word_id, word_frequency), where word_id is the unique integer ID of the word in the dictionary, and word_frequency is the count of that word in the document.
Sixth). Create topic using Latent Dirichlet Allocation
Seventh). Check optimal topic by using coherence score. Here, Coherence score, in the context of topic modeling and natural language processing (NLP), is a metric used to evaluate the quality and interpretability of the topics generated by a topic model. Coherence measures how well the words within a topic are related and whether they form meaningful, coherent themes. It helps assess whether the topics produced by a model make sense and are useful for understanding the underlying content of a collection of documents.

Results

By using range 2 to 10 topics, we are found that the optimal number of topics is 3 with the coherence score 0.621520 which can be seen at figure 3:

![Figure 3. coherence score](image)

About topics, we can see the results obtained from the topic modeling using LDA as table 2:

<table>
<thead>
<tr>
<th>Topic</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025*”berapa_lama” + 0.025*”dude” + 0.019*”ada” + 0.017*”baru” + 0.017*”juga” + 0.016*”harga” + 0.014*”review” + 0.013*”handphone” + 0.013*”infinix” + 0.012*”nonton”</td>
<td></td>
</tr>
</tbody>
</table>
From table 2, we can see that topic 0 is discussed about price and time it will be available to purchase. In topic 1 it discussed about the product is hard to obtain due to it not available in market. In topic 2, it discussed about ownership.

To better understand which comments belong to which topics, we assigned the topics into the original dataset which can be seen at table 3 (only sample):

<table>
<thead>
<tr>
<th>Document_No</th>
<th>Dominant_Topic</th>
<th>Topic_Perc_Contrib</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>0.922</td>
<td>['speakernya', 'geter', 'dudeet', 'sampai', 'body', 'belakang', 'kayak', 'speaker', 'hajat']</td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td>0.6044</td>
<td>['jelas', 'infinix', 'lah', 'pakai', 'nanya']</td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>0.9245</td>
<td>['sudah', 'tau', 'pro', 'reguler', 'jadi', 'kayak', 'tidak', 'sewaaw', 'itu']</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0.6373</td>
<td>['android', 'web', 'sistem', 'nonaktif', 'dude', 'biar', 'tidak', 'dapat', 'pemberitahuan', 'terus']</td>
</tr>
<tr>
<td>38</td>
<td>2</td>
<td>0.9188</td>
<td>['iya', 'buat', 'game', 'oke', 'sektor', 'laen', 'gimna', 'dude']</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0.9562</td>
<td>['di', 'atur', 'aplikasi', 'pilih', 'aplikasi', 'mau', 'di', 'nonaktif', 'pemberitahuan', 'sudah', 'selesai', 'infinix', 'note', 'pro', 'saya', 'aman']</td>
</tr>
</tbody>
</table>

**Conclusion**

By Using Latent Dirichlet Allocation for social media comments of product review, we and especially product owner can know what topics are discussed in their product. Thus, it make awareness to the product owner to act according the topics discovered. As discussed above, we are aiming in Text preprocessing requirement research gaps. It shows that in topic modeling, we do not need to remove stop words due to it remove the time domain. Although Latent Dirichlet Allocation can be used for topic modeling, it still lacks the aspect of sentiments. Such as: what topic have the positive sentiment and which one have negative sentiment. Thus, in the future research, we will conduct the aspect-based sentiment analysis for product reviews.
Bibliography


