Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849 e-ISSN: 2548-1398
Vol. 7, No. 12, December 2022

# OPTIMIZING CONTENT MARKETING USING AUTOMATIC KEYWORD EXTRACTION TO GET TOPIC PREDICTION

**Savitri Indriyani**
Swiss German University, Indonesia
Email: savitri@gmail.com

## Abstract

Digital news with a variety topics is abundant on the internet. The problem is to classify news based on its appropriate category to facilitate user to find relevant news rapidly. The manual categorization of text documents requires a lot of financial and human resources to do the process. In order to get so, topic modeling usually used to classify documents. In the used topic models (LSA, LDA) each word in the corpus of vocabulary is connected with one or more topics with a probability, as estimated by the model. Many (LDA, LSA) models were built with different values of coherence and pick the one that produces the highest coherence value. Based on the result, we summarized some points, three models above can answer the question in Research Question, those models can be applied in the future to company's automation prosess of determining topic automatically. LDA using BOW and LSA using BOW would be priority option to be applied.

**Keywords**: Topic Modelling, LDA, LSA, Bag of Words, TF-IDF.

## Introduction

The importance of digital marketing has increased from time to time as part of a marketing strategy that is now increasingly being practiced by any organization, including startup company. Therefore, a digital marketing strategy cannot be successful without having a quality content marketing.

Many companies are interested to increase the use of the content marketing tool in their marketing policy, as they notice the limitations of the traditional marketing communication strategy, as well as the huge opportunities brought by digital marketing (Chen et al., 2017).

According to the statistics provided by the Content Marketing Institute, the importance of content marketing is growing, as 70% of B2B marketers are creating more content than they did one year ago (Vinodhini & Chandrasekaran, 2012) and (Liu, 2012). This is explained by the fact that “60% of B2B decision makers say branded content helps them make better purchase decisions, while 61% of consumers are more likely to buy from

| | |
|---|---|
| **How to cite:** | Indriyani, Savitri (2022) Optimizing Content Marketing Using Automatic Keyword Extraction To Get Topic Prediction, *Syntax Literate : Jurnal Ilmiah Indonesia (7)12,* http://dx.doi.org/10.36418/syntax-literate.v7i12.10782 |
| **E-ISSN:** | 2548-1398 |
| **Published by:** | Ridwan Institute |

companies that offer custom content" .Content itself must be relevant to your audience and create a powerful brand image, this become important because company need to win customer's trust and royalty (Oghaz et al., 2020). The importance of valuable content is that company can build interest that transforms into lasting relationships. Content marketing can be transformed in many ways, for examples Infographics, webpages, podcast, etc. Company can choose the media to publish the content such as Microblogs like Twitter, social platforms like Facebook, or forums like LinkedIn Discussions or facebook forums (Payak et al., 2020).

As a company that focuses on developing technology in agriculture, Biops Agrotekno Indonesia offers the concept of precision farming (precision farming) (Thomas et al., 2016), which is a measurable agricultural concept that is able to adjust the supply of water and nutrients according to plant needs that can be done automatically and can be monitored in real time using applications on smartphones (Röder et al., 2015).

One of the marketing strategies they used to increase engagement with customers is content marketing strategy, by posting campaign content using trending agriculture topic at that time to their webpages and social media (Shi et al., 2017).

## Methods

The experiment was performed on a dataset consisting news article from selected e-newspaper. News article datasets originating from *kompas.com* and *detik.com* which is some of the digital news site in Indonesian language which are sought after news seekers (Loza et al., 2014). Data taken from the website are news published in December 2021 and early 2022.

The data are collected from selected article e-newspaper using its full content. Data collected using web scraping method, and in this experiment Python was selected as a device for executing web scraping. The dataset only consist the content of the articles. In the text processing, the steps consist of these following steps : tokenizing, remove punctuation, stopwords, and stemming (Onan et al., 2016).

## Results and Discussion

### A. Experiment and Data Analysis

Corpus, dictionary, and a number of topics are needed to train the (LDA and LSA) model, where each word in the corpus of vocabulary is then connected with one or more topics with a probability which estimated by the model. LDA and LSA model was built with various topics where each topic is a mixture of keywords and each keyword contributes a certain weight to the topic (Keneshloo et al., 2016).

The topic modeling experiment was the phase which carried out to form the best topic model by conducting experiments on input parameters.

### B. Experiment and Evaluation

Before obtaining the topic modeling result, we have to specify the number of topics. In this experiments, the best number of topics in each method was obtained by calculating the coherence score with the range between 5 to 10 topics. The highest coherence score indicates the appropriate number of topics (Onan et al., 2016).

The size of each words presented the importance of the words in a collection of texts. In addition, we determined the topic label based on the word presented.

## 1. Topic Modelling using LDA and Bag of Words as Word Vectorizer

In our experiment, we tested coherence score by applying a different number of topics t, t = 5,6,7,8,9,10 topics. Coherent score evaluation of topic modelling performance using LDA and Bag of Words as Word Vectorizer as shown by Figure 15 and the result of coherence score for each number of topics as shown by the Table 1.

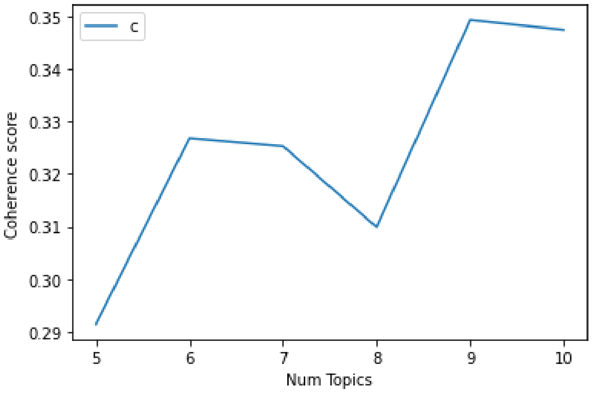

**Figure 1. Coherent score of Topic Modelling using LDA and Bag of Words**

**Table 1.**
**Output of Coherent score of Topic Modelling using LDA and Bag of Words**

| Num Topics | Coherent Value |
|---|---|
| 5 | 0.291444 |
| 6 | 0.326842 |
| 7 | 0.325318 |
| 8 | 0.309921 |
| 9 | 0.34935 |
| 10 | 0.347481 |

From the result of Table 4, num topics = 9 giving the highest score of Coherent score among other num topics (Lee & Kim, 2008). Bon the result of coherent score, Topic Modelling using LDA and Bag of Words simulated using number of topic = 9, which obtain following topics as shown by Table 5. For each word value embedded to keyword, we summarize the absolute value to Column Score. Result on the Table 2 sorted descending by the score (Pilato & Vassallo, 2014).

**Table 2.**
**Topics generated by LDA Topic modelling (BOW)**

| Opic | Core | Keyword + Score |
|---|---|---|
| Topic: 6 | 0.140 | 0.030*"makmur" + 0.029*"program" + 0.012*"erick" + 0.012*"indonesia" + 0.010*"tembakau" + 0.010*"tingkat" + 0.010*"bumn" + 0.010*"thohir" + 0.009*"tanam" + 0.008*"pupuk" |
| Topic: 3 | 0.132 | 0.023*"korban" + 0.018*"warga" + 0.017*"motor" + 0.014*"desa" + 0.013*"orang" + 0.010*"laku" + 0.010*"polisi" + 0.009*"rumah" + 0.009*"bawa" + 0.009*"kabupaten" |
| Topic: 1 | 0.130 | 0.019*"bri" + 0.016*"umkm" + 0.016*"usaha" + 0.014*"tingkat" + 0.013*"erick" + 0.012*"persen" + 0.011*"jeruk" + 0.011*"dukung" + 0.010*"thohir" + 0.008*"tanam" |
| Topic: 5 | 0.127 | 0.033*"pupuk" + 0.017*"indonesia" + 0.014*"teknologi" + 0.012*"program" + 0.011*"lahan" + 0.010*"subsidi" + 0.008*"karya" + 0.008*"update" + 0.007*"kembang" + 0.007*"hektar" |
| Topic: 8 | 0.122 | 0.023*"hama" + 0.014*"tanam" + 0.013*"manfaat" + 0.012*"indonesia" + 0.012*"motor" + 0.012*"alami" + 0.011*"update" + 0.009*"buah" + 0.008*"curi" + 0.008*"rumah" |
| Topic: 7 | 0.115 | 0.015*"harga" + 0.014*"pupuk" + 0.013*"pangan" + 0.013*"komoditas" + 0.012*"tingkat" + 0.012*"indonesia" + 0.011*"sektor" + 0.009*"program" + 0.008*"lahan" + 0.008*"sawit" |
| Topic: 0 | 0.103 | 0.014*"tembakau" + 0.013*"indonesia" + 0.012*"tanah" + 0.010*"lahan" + 0.010*"air" + 0.010*"resap" + 0.009*"tingkat" + 0.009*"program" + 0.008*"sumur" + 0.008*"perintah" |
| Topic: 2 | 0.100 | 0.017*"harga" + 0.016*"persen" + 0.010*"tingkat" + 0.010*"anj" + 0.009*"naik" + 0.008*"sawit" + 0.008*"gunung" + |

| | | |
|---|---|---|
| | | 0.008*"rumah" + 0.007*"panen" + 0.007*"curi" |
| Topic: 4 | 0.098 | 0.015*"jokowi" + 0.011*"update" + 0.010*"bangun" + 0.010*"ekspor" + 0.009*"negara" + 0.009*"panen" + 0.009*"impor" + 0.009*"menteri" + 0.008*"bendung" + 0.008*"kabupaten" |

2. **Topic Modelling using LDA and TF-IDF as Word Vectorizer**

In our experiment, we tested coherence score by applying a different number of topics t, t = 5,6,7,8,9,10 topics. Coherent score evaluation of topic modelling performance using LDA and TF-IDF as Word Vectorizer as shown by Figure 16 and the result of coherence score for each number of topics as shown by the Table 2.

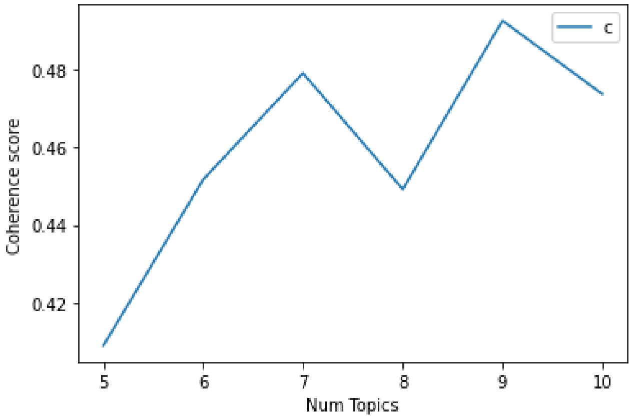

**Figure 2. Coherent score of Topic Modelling using LDA and TF-IDF**

**Table 3.**
**Output of Coherent score of Topic Modelling using LDA and TF-IDF**

| Num Topics | Coherent Value |
|---|---|
| 5 | 0. 408928 |
| 6 | 0. 451697 |
| 7 | 0. 479131 |
| 8 | 0. 449178 |
| 9 | 0. 492648 |
| 10 | 0. 473736 |

Based on the result of coherent score, Topic Modelling using LDA and TF-IDF simulated using number of topic = 9, which obtain following topics as shown by

Table 7. For each word value embedded to keyword, we summarize the absolute value to Column Score. Result on the Table 7 sorted descending by the score (Beliga et al., 2015) dan (Albalawi et al., 2020).

**Table 4.**
**Topics generated by LDA Topic modelling (TF-IDF)**

| Topic | Score | Keyword + Score |
|---|---|---|
| Topic: 8 | 0.052 | 0.009*"makmur" + 0.007*"program" + 0.006*"erick" + 0.006*"thohir" + 0.006*"ekspor" + 0.005*"sektor" + 0.004*"kompetisi" + 0.003*"bumn" + 0.003*"kopi" + 0.003*"rawan" |
| Topic: 0 | 0.049 | 0.006*"jokowi" + 0.006*"bawang" + 0.006*"impor" + 0.005*"keluh" + 0.005*"telepon" + 0.005*"korban" + 0.004*"curi" + 0.004*"masuk" + 0.004*"tanggung" + 0.004*"putih" |
| Topic: 6 | 0.048 | 0.006*"tembakau" + 0.006*"anj" + 0.006*"hama" + 0.005*"resap" + 0.005*"lahan" + 0.004*"inovasi" + 0.004*"kumbang" + 0.004*"sumur" + 0.004*"industri" + 0.004*"air" |
| Topic: 1 | 0.047 | 0.007*"mosaik" + 0.005*"romawi" + 0.005*"vila" + 0.005*"pupuk" + 0.005*"temu" + 0.004*"jagung" + 0.004*"hiu" + 0.004*"subsidi" + 0.004*"subsektor" + 0.004*"sensor" |
| Topic: 4 | 0.045 | 0.007*"motor" + 0.005*"mitra" + 0.005*"tembakau" + 0.004*"mata" + 0.004*"gawai" + 0.004*"yogyakarta" + 0.004*"curi" + 0.004*"probolinggo" + 0.004*"hama" + 0.004*"polisi" |
| Topic: 3 | 0.043 | 0.006*"pupuk" + 0.005*"makmur" + 0.005*"padi" + 0.005*"bendung" + 0.004*"program" + 0.004*"produktivitas" + 0.004*"persen" + 0.004*"tingkat" + 0.003*"trenggalek" + 0.003*"harap" |
| Topic: 2 | 0.042 | 0.006*"umkm" + 0.005*"pupuk" + 0.005*"bri" + 0.005*"organik" + 0.004*"persen" + 0.004*"anj" + 0.004*"bangun" + 0.003*"sektor" + 0.003*"ekspor" + 0.003*"padi" |
| Topic: 7 | 0.040 | 0.008*"cabai" + 0.005*"pidekso" + 0.004*"beras" + 0.004*"panen" + 0.004*"waduk" + 0.003*"harga" + 0.003*"digital" + 0.003*"impor" + 0.003*"tanah" + |

| | | |
|---|---|---|
| | | 0.003*"wonogiri" |
| Topic: 5 | 0.036 | 0.005*"pati" + 0.005*"nasi" + 0.005*"masak" + 0.003*"estate" + 0.003*"gula" + 0.003*"alat" + 0.003*"korban" + 0.003*"rempahrempah" + 0.003*"serangga" + 0.003*"rempah" |

### 3. Topic Modelling using LSA and Bag of Words as Word Vectorizer

In our experiment, we tested coherence score by applying a different number of topics t, t = 5,6,7,8,9,10 topics. Coherent score evaluation of topic modelling performance using LSA and Bag of Words as Word Vectorizer as shown by Figure 17 and the result of coherence score for each number of topics as shown by the Table 5.

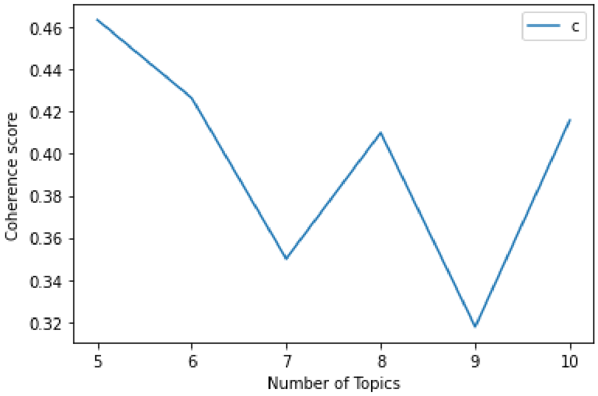

**Figure 5 Coherent score of Topic Modelling using LSA and Bag of Words**

Based on the result of coherent score, Topic Modelling using LSA and Bag of Words simulated using number of topic = 5, which obtain following topics as shown by Table 9. For each word value embedded to keyword, we summarize the absolute value to Column Score. Result on the Table 9 sorted descending by the score.

### 4. Topic Modelling using LSA and TF-IDF as Word Vectorizer

In our experiment, we tested coherence score by applying a different number of topics t, t = 5,6,7,8,9,10 topics. Coherent score evaluation of topic modelling performance using LSA and TF-IDF as Word Vectorizer as shown by Figure 18 and the result of coherence score for each number of topics as shown by the Table 6.

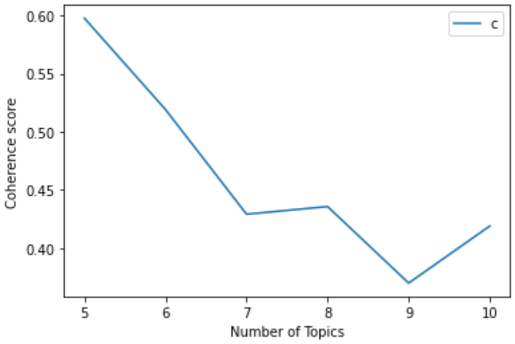

**Figure 6. Coherent score of Topic Modelling using LSA and TF-IDF**

Based on the result of coherent score, Topic Modelling using LSA and TF-IDF simulated using number of topic = 5, which obtain following topics as shown by Table 11. For each word value embedded to keyword, we summarize the absolute value to Column Score. Result on the Table 11 sorted descending by the score.

5. **Analysis of Topic Modelling Result From Expert Analysis Team**

The results from the modeling using four methods: (a) LDA and Bag of Words as Word Vectorizer; (b) LDA and TF-IDF as Word Vectorizer; (c) LSA and Bag of Words as Word Vectorizer; (d) LSA and TF-IDF as Word Vectorizer; have been analyzed by three people from BIOPS Agrotekno. The people doing the analysis are the ones who are doing the SEO project of the company. As the hashtag used in the research (#pertanian, #petani, and #teknologi pertanian) are similar with some of the keywords that is usually used in the company's SEO project, the point of view could be used as the expert reference. There are two parameters that is used to validate the results. There are: (a) the relevance of keywords to the major topics; (b) the coherence between words in one topic group.

Based on the result, There are some of the keywords that are considered as irrelevant. Some of the irrelevant words are: "korban", "motor", "karya", "curi", "hiu" etc. These words are not closely related to the groups of words the expert team expected. At first, they thought that it could be that at the period of time where the articles are taken, these words are actually related to the main topics. However, when they analyzed the LSA and Bag of Words as Word Vectorizer, the relevance rate wass high. The keywords resulted from this method are more relevant compared to the other methods. Therefore, they picked this method as the best method to give the most relevant results. Nevertheless, there are still some keywords that is irrelevant resulted by this method, yet the amount is fewer compared to others. (Note: this point of view is only based on the scope used in the research. Change in parameters may also change the result and the validation).

The Relevance from their opinion between the topic and the extracted keywords as shown by Table 12, 13, 14, and 15. It represented by the color, Green is Relevant (correlation High), Yellow is Not too relevant (correlation Medium), and Red is Irrelevant (correlation Low).

They also analyze the coherence between words in one topic. From four methods, they could say that again LSA and Bag of Words as Word Vectorizer method resulted the most coherence between words. For example, the Topic 1: "pangan" + "lahan" + "estate" + "harga" + "food" + "persen" + "ekspor" + "komoditas" + "program" + "indonesia". Even only based on these set of words, they get a picture that the topic is around the food estate program by Indonesian government to increase the food security and export.

Based on the validation they already done, they prefer the **LSA and Bag of Words as Word Vectorizer** method as the best compared to the others. The conclusion is based on at least two factors: Relevance and Coherence, combined with their knowledge in the sectors. They believe that this modeling could help us increasing the efficiency of our SEO projects, especially in the keywords searching. In the future, before it can be used further, another validation with other parameters (periods of time, topics, media sources, etc).

## Conclusion

Based on the results that got on in this research, we applied topic modeling using Latent Dirichlet Allocation and Latent Semantic Analysis to discover topics from selected article from e-newspaper, which are Kompas.com and Detik.com. By applying topic modeling, we could find several insightful topics that illustrate necessary information from the articles.

Based on the result, we summarized some points, three models above can answer the question in Research Question, those models can be applied in the future to company’s automation prosess of determining topic automatically. LDA using BOW and LSA using BOW would be priority option to be applied.

## BIBLIOGRAPHY

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, *39*(1), 1–20.

Chen, Y., Rabbani, R. M., Gupta, A., & Zaki, M. J. (2017). Comparative text analytics via topic modeling in banking. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8.

Keneshloo, Y., Wang, S., Han, E.-H., & Ramakrishnan, N. (2016). Predicting the popularity of news articles. *Proceedings of the 2016 SIAM International Conference on Data Mining*, 441–449.

Lee, S., & Kim, H. (2008). News keyword extraction for topic tracking. *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, *2*, 554–559.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Loza, V., Lahiri, S., Mihalcea, R., & Lai, P.-H. (2014). Building a Dataset for Summarization and Keyword Extraction from Emails. *LREC*, 2441–2446.

Oghaz, T. A., Mutlu, E. Ç., Jasser, J., Yousefi, N., & Garibay, I. (2020). Probabilistic model of narratives over topical trends in social media: A discrete time model. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 281–290.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232–247.

Payak, A., Rai, S., Shrivastava, K., & Gulwani, R. (2020). Automatic text summarization and keyword extraction using natural language processing. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 98–103.

Pilato, G., & Vassallo, G. (2014). TSVD as a statistical estimator in the latent semantic analysis paradigm. *IEEE Transactions on Emerging Topics in Computing*, *3*(2), 185–192.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search* *and Data Mining*, 399–408.

Shi, L.-L., Liu, L., Wu, Y., Jiang, L., & Hardy, J. (2017). Event detection and user interest discovering in social media data streams. *IEEE Access*, *5*, 20953–20964.

Thomas, J. R., Bharti, S. K., & Babu, K. S. (2016). Automatic keyword extraction for text summarization in e-newspapers. *Proceedings of the International Conference on Informatics and Analytics*, 1–8.

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6), 282–292.