Topic Modeling of Indonesian Children's Literature Using Latent Semantic Analysis

Winda Monika[1], Vita Amelia[2], Qory Islami Aris[3], Arbi Haza Nasution[4]

{windamonika@unilak.ac.id[1], vitaamelia@unilak.ac.id[2], qoryislamiaris@unilak.ac.id[3], arbi@eng.uir.ac.id[4]}

[1,2,3]Library Science, Universitas Lancang Kuning, Indonesia
[4]Informatics Engineering, Universitas Islam Riau, Indonesia

Abstract. The Covid-19 pandemic resulted a significant reduction in the number of publications of literary works, especially works of children's literature, which once dominated the first ranking of the most popular subjects at the Indonesian Publishers Association. So that efforts are needed to raise the enthusiasm of Indonesian children's literature. In addition, data on children's literature is scattered across various portals that compile metadata/bibliography and synopsis summaries of children's literature stories from various sources or online portals. A portal that collects data on all Indonesian children's literature that is grouped automatically according to the most dominant topics reflecting the context of each literary work is urgently needed to facilitate search, develop teaching materials, and inspire the creation of new literary works. This study uses qualitative methods and experiments with several stages of research including metadata analysis of models of children's literature, secondary data collection, topic modeling of children's literature with Latent Semantic Analysis (LSA). The results of the study showed that the topic clustering was most often raised by children's authors published on the Mizan.

Keywords: Topic Modeling, Indonesian children's literature, Latent Semantic Analysis.

1 Introduction

Since the emergence of COVID-19 pandemic, smartphone and any kind of digital devices have been introduced and massively used even by early school children. This kind of digital exposure has changed the way children experience in learning, playing, and entertaining which all the interaction are done in digital environment. As a media, digital devices could bring benefits if it is used in appropriate way, yet several studies mentioned that it could bring side effects to children as well. The roles of parents and teachers are important to guide the children determining which allowed apps, content, and portal only to be accessed.

Digital material content is valuable resources that could be provided as a source of information delivered to children. The creation of this digital material content requires high quality of children literature or literary works as references. Children's literature is a medium or means of character education for children/students [1] which should be used in developing teaching materials for students starting from the Early Childhood Education level to High School [2]. The creation of many works of children's literature shows an improvement in the

ICEEES 2023, October 30-31, Pekanbaru, Indonesia

DOI 10.4108/eai.30-10-2023.2343063

quality of literary literacy among Indonesian children. Furthermore, many children's literary works can be accessed in electronic and audio formats. This has become one of the strategic issues to reach a larger market.

Children's literary works have begun to receive appreciation and great attention from Indonesian society since the emergence of “Kecil-Kecil Punya Karya”(KKPK) by publisher Mizan which published works by a 7-year-old child in December 2003[3]. This section is a fresh start in the world of children literature in Indonesia, where it provides room for children to express their intellectual, creation, and imagination to be embodied into a literary work. This move deserves to be appreciated and acknowledged in society. However, the enthusiasm of connoisseurs of children's literature has decreased dramatically during the Covid-19 pandemic. Based on data from the Indonesian Publishers Association (IKAPI), the publication of children's books with subjects, namely Children's Books 1001, used to be ranked first with an increase from 22.31% (2013) to 22.64%. During the Covid-19 pandemic, there was a decline in book sales for all subjects where as many as 58.2% of publishers experienced a decline in sales exceeding 50% from normal times. Therefore, efforts are needed to increase productivity and production in children’s literary works. A portal that collects data on all Indonesian children's literature that is grouped automatically according to the most dominant topic reflecting the context of each literary work is urgently needed to facilitate search, make it easier for users to explore the meaning contained in Indonesian children's literature, provide recommendations to stakeholders for develop teaching materials for students from a wide selection of existing Indonesian children's literature and inspire the creation of new literary works.

In the current Big Data phenomenon, the explosion of information makes it difficult to understand events or concepts in a document. For humans it is easy to read and understand the text of a document, but not with machines that are only given text without any context/subject to the text. Therefore, the method of understanding the context of the text is called topic modeling. This study aims to model the Indonesian children’s literary works topic using latent semantic analysis.

2 Literature Study

2.1 Topic Modeling

Topic modeling is a data mining approach used to analyze the hidden themes and thematic structures from large amounts of text such as documents or corpus using mathematical algorithms [4, 5]. Topic modeling can identify themes related to letter patterns in big data. When data containing words and text are found, the topic model maps words in terms of word similarities and occurrences, with the aim of obtaining closeness between words in the corpus [6, 7].

2.2 Latent Semantic Analysis (LSA)

LSA is a theory and method for extracting meaning from text based on statistical computations from document collections [8]. LSA is part of the most widely used topic modeling [9]. Several studies have used several different algorithms in text extraction, categorization, and analysis, where the use of LSA shows that the results of the categorization

of LSA obtain high ratings [10, 11]. The initial stage, the words per paragraph are made into a matrix or called document-term matrix (DTM). Furthermore, mathematical methods are used to reduce the matrix known as singular value decomposition (SVD). The SDV estimates the vector for each word and evaluates the proximity of the words in multidimensional vectors [12].

Pre-processing. Pre-processing is the stage of preparing the text in the dataset by separating sentences into individual words based on spaces (tokenizing), removing punctuation and symbols (filtering), converting all words to lowercase (case folding), changing words affixes become basic words (stemming) and remove common words (stopwords) [13]. The result of the initial processing is a bag-of-words model for each children's literature text.

Word Weighting. At this stage, the process of changing data in bag-of-words from text to numeric is carried out using the TF and IDF weighting methods [14]. There are several variations of TF weighting and several variations of IDF weighting. Several experiments will be carried out to determine which combination of weighting schemes is the most appropriate.

$$q(w) = fd(w) + \log(|D|/fD(w)) \quad (1)$$

fd(w) is the repeated term of the word w in history d, FD(w) is the number of files containing the words w and | D | number of domes in collection D [15].

Singular Value Decomposition (SVD). is a method for identifying patterns of relationships between words and concepts contained in a document, or patterns of relationships between documents and these concepts. SVD will describe a matrix as shown in Figure 1, A ∈ ℜm × n where m is the number of different terms and n is the number of sentences determined from within the text. Usually terms (terms) differ greater than the number of sentences (m ≥ n).

$$\begin{bmatrix} \begin{bmatrix} a_{1,1} \\ a_{2,1} \\ \dots \\ a_{m,1} \end{bmatrix} \begin{bmatrix} a_{1,2} \\ a_{2,2} \\ \dots \\ a_{m,2} \end{bmatrix} \dots \begin{bmatrix} a_{1,n} \\ a_{2,n} \\ \dots \\ a_{m,n} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} u_{1,1} \\ \dots \\ u_{m,1} \end{bmatrix} \dots \begin{bmatrix} u_{1,k} \\ \dots \\ u_{m,k} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \sigma_1 & \dots & 0 \\ & \dots & \\ 0 & \dots & \sigma_k \end{bmatrix} \begin{bmatrix} \begin{bmatrix} v_{1,1} \\ \dots \\ v_{1,k} \end{bmatrix} \dots \begin{bmatrix} v_{n,1} \\ \dots \\ v_{n,k} \end{bmatrix} \end{bmatrix}$$

$$A_k \qquad U_k \qquad \Sigma_k \qquad V_k^T$$

Figure 1. SVD Approach of the Matrix

SVD of matrix A ∈ ℜm × n is defined as follows (Mashechkin et al., 2011).

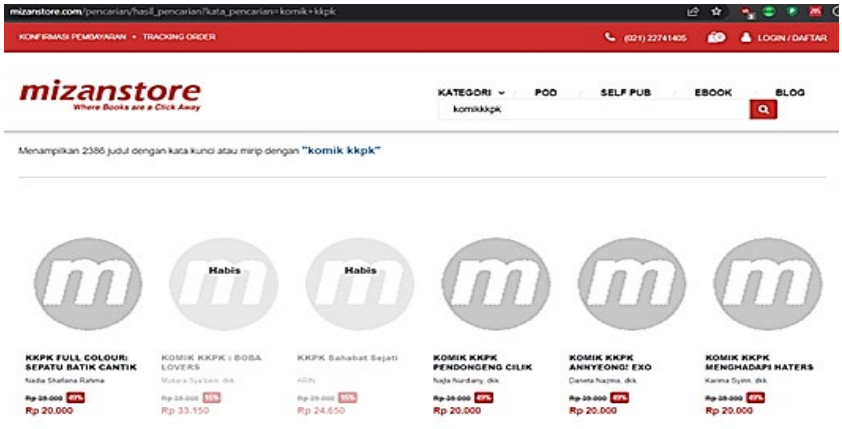
$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \quad (2)$$

U ∈ ℜm × n is a matrix formed from columns of orthonormal matrices called left singular vectors (ui); Σ = diag(σ1, σ2, …, σn) ∈ ℜm × n is a diagonal matrix such that σ1 ≥ σ2 ≥ … ≥ σr > 0 = σr +1 = … = σn, r = rank(A) ≤ min(m, n) is the power of matrix A and σi is the singular value of matrix A; and V ∈ ℜm × n are orthonormal matrices called right singular vectors (vi).

3 Methodology

3.1 Experiment Design

This study use data from various portals that collect metadata/bibliography and synopsis summaries of children's literature stories. There are approximately 2386 titles in the Mizan collection (KKPK), 194,477 titles for children's stories on Wikipedia/DBpedia. To specify the searching, the keyword used is “karya anak”. As depicted in Fig 2, metadata (e.g., title and author) is gathered as a dataset to be analyzed further. Web crawling techniques was used by using python script.



	Title	Author
0	KKPK FULL COLOUR: SEPATU BATIK CANTIK	Nadia Shafiana Rahma
1	KOMIK KKPK : BOBA LOVERS	Mutiara Sya'bani, dkk.
2	KKPK Sahabat Sejati	ARIN
3	KOMIK KKPK PENDONGENG CILIK	Najla Nurdiany, dkk.
4	KOMIK KKPK ANNYEONG! EXO	Daneta Nazma, dkk.
5	KOMIK KKPK MENGHADAPI HATERS	Karima Syirin, dkk.
6	KOMIK KKPK PERPUSTAKAAN MISTERIUS	Azra Naila Hanun, dkk.
7	KOMIK KKPK : RAMADHAN ADVENTURE	Nabila Nurkhalisah
8	KKPK KAKAKKU, PAHLAWANKU	Fayanna Ailisha Davianny
9	KKPK.PERTEMUAN TERAKHIR	FRISKA SIVA MAULIDA
10	Komik KKPK: Dragon War	FAIKAR
11	KKPK.PENYAMARAN ARINA	SYADRINA RACHMA PARAMITHA & DANESH
12	KOMIK KKPK RAHASIA RUMAH KAKEK	Mutiara Sya'bani, dkk.
13	KOMIK KKPK STOP PLAGIAT!	Kirana Riris, dkk.
14	KOMIK KKPK BEKAL ISTIMEWA BUNDA	M. NauvalAssyafiqi, dkk.
15	KOMIK KKPK RUMAH PETUALANGAN	Ica Annysa, dkk
16	KOMIK KKPK : MACARON DAY	
17	KKPK Smart Princess	KINTAN AYU SIVA S
18	KOMIK DETEKTIF KKPK: PETUALANGAN CHERRY	ERAMAYAWATI
19	KOMIK KKPK RATU UPDATE	Abdurrahmansyah
20	KOMIK DETEKTIF KKPK: BUKAN SEMBARANG GAMERS	Nafisa Zunilova
21	KOMIK KKPK KEJUTAN RUMAH KOSONG	Najwa Futhana R., dkk.
22	KOMIK KKPK PULAU HARTA KARUN	Hasya Nabilla, dkk.
23	KOMIK DETEKTIF KKPK: MISTERI HANTU FESTIVAL	RASYIQA ANNISA THOHIRA
24	KOMIK KKPK : NO PLASTIC BAG	
25	KOMIK KKPK YOUTUBER CHALLENGE	

Figure 2. Dataset

4 Results

The data obtained is then pre-processed in the form of data cleaning by removing unnecessary characters such as punctuation marks, changing titles to lowercase letters, and removing common words that are not related to the topic (stopwords).

```
# Load the regular expression library
import re
# Remove punctuation
df['Title_processed'] = \
df['Title'].map(lambda x: re.sub('[,\.!?:]', '', x))

# Convert the titles to lowercase
df['Title_processed'] = \
df['Title_processed'].map(lambda x: x.lower())
df['Title_processed'] =
df['Title_processed'].str.replace('-new','')
df['Title_processed'] =
df['Title_processed'].str.replace('kkpk','')

# Print out the first rows of papers
df['Title_processed'].head()

# Stopword removal
stopword_en = stopwords.words('english')
stopword_list = stopwords.words('indonesian')
new_stopwords = ['kkpk', 'kkpkmy', 'next', 'komik',
'kkpknext', 'g', 'nomik', 'reg', 'new', 'full', 'class',
'colour', 'luks', 'si', 'vol', 'paket']
stopword_list.extend(stopword_en)
stopword_list.extend(new_stopwords)
df['Title_without_stopwords'] =
df['Title_processed'].apply(lambda x: ' '.join([word for
word in x.split() if word not in (stopword_list)]))
reindexed_data = df['Title_without_stopwords']
stopword_list
for row in reindexed_data:
    if row.find('new')>0:
        print(row)
```

At this stage the Vectorizer is carried out, namely the process of changing or transforming text into a vector based on the frequency (count) of each word that appears throughout the text. Vectorizer data will be displayed in the form of a bar chart.

```
count_vectorizer = CountVectorizer()
words, word_values = get_top_n_words(n_top_words=15,

count_vectorizer=count_vectorizer,

text_data=reindexed_data)

fig, ax = plt.subplots(figsize=(16,8))
ax.bar(range(len(words)), word_values);
ax.set_xticks(range(len(words)));
ax.set_xticklabels(words, rotation='vertical');
ax.set_title('Top words in headlines dataset (excluding
stop words)');
ax.set_xlabel('Word');
ax.set_ylabel('Number of occurences');
plt.show()"
```

The results of the vectorizer as depicted in Fig 3. show that some of the words that appear most often from the dataset include "mystery", "mysterious", "secret", "school", "home", "adventure", and so on.

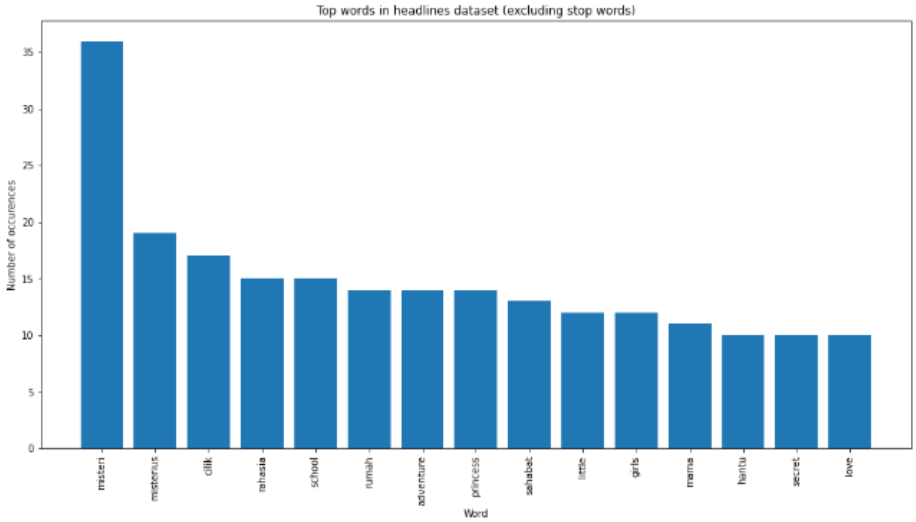


Figure 2. Chart of Vectonizer of Indonesian Children's Literary Works

The results of the vectorizer then look for patterns of relationships between words in the document by implementing SVD.

```
small_count_vectorizer =
CountVectorizer(max_features=40000)
small_text_sample = reindexed_data.sample(n=800,
random_state=0).values

print ('Headline before vectorization: {}'.
format(small_text_sample[123]))

small_document_term_matrix =
small_count_vectorizer.fit_transform(small_text_sample)

top_n_words_lsa = get_top_n_words(10, lsa_keys,
small_document_term_matrix, small_count_vectorizer)

for i in range(len(top_n_words_lsa)):
    print ("Topic {}: ". format(i+1), top_n_words_lsa[i])


top_n_words_lsa = get_top_n_words(10, lsa_keys,
small_document_term_matrix, small_count_vectorizer)
for i in range(len(top_n_words_lsa)):
    print ("Topic {}: ". format(i+1), top_n_words_lsa[i])
```

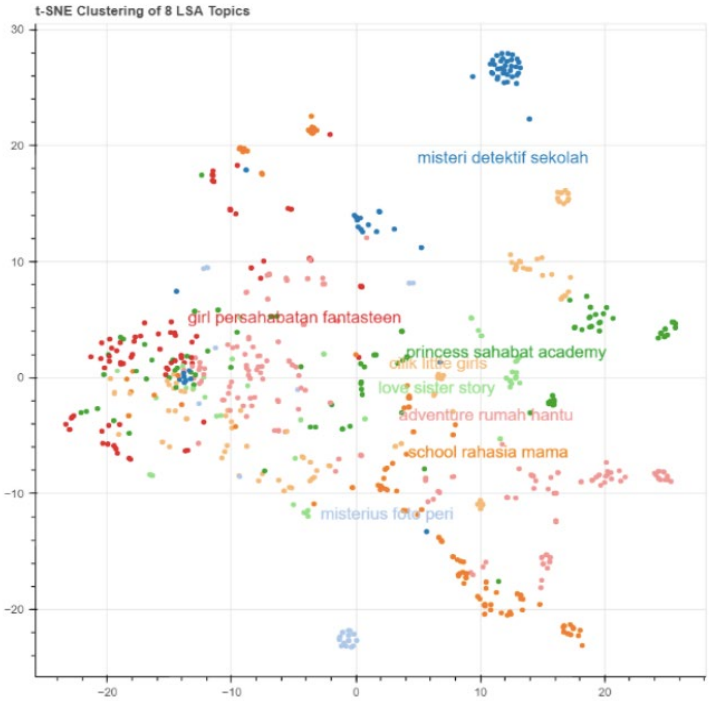


Figure 3. Clustering LSA Results

LSA clustering displays the 8 most/biggest topics from the dataset in the form of:
Topic 1: misteri detektif sekolah tersembunyi peri malam festival kucing malas tua
Topic 2: misterius perpustakaan foto congklak hati kupu kecil suara palsu buku
Topic 3: school rahasia best day ghost rpl beautiful sepatu days friend
Topic 4: cilik little persahabatan girl seri diary story cooking star pesta
Topic 5: princess academy republishes anak deluxe teman jeritan smart lovely kepo
Topic 6: mama secret love vs happy sister mom fun false korea
Topic 7: bunda pop gara girls ajaib party lets semangat horor nenek
Topic 8: adventure rumah sahabat hantu petualangan world fantasteen magic edisi cookies

5 Conclusion

Topic modeling using latent semantic analysis groups the titles of children's writers' work on the Mizan Kecil-Kecil Punya Karya (KKPK) portal into 8 topics. This topic describes the themes most often used in writing by children's writers. Topic modeling using LSA can be visualized in the form of clustering, making it easier to find and understand topics from big data. The experimental results show that the LSA algorithm can map topics from the KKPK title well, but the evaluation process cannot be fully automated, where it still involves humans to determine the accuracy of the modeled topics. Future research can compare LSA results with other topic modeling techniques in text summarization to find the best algorithm.

References

1. Sukirman, S.: Karya Sastra Media Pendidikan Karakter bagi Peserta Didik. J. Konsepsi. 10, 17–27 (2021).
2. Putri, R.A.: Nilai Moralitas Sebagai Pengembangan Karakter Anak Dalam Seri Dongeng 3D Nusantara: Malin Kundang. J. Edukasi Khatulistiwa Pembelajaran Bhs. dan Sastra Indones. 4, 63–71.
3. Akbar, R., Rifai, I., Lee, J.: Ways with Words: Exploring Children Author's Voices in Indonesia's Children Book Series (KKPK). In: 1st UMGESHIC International Seminar on Health, Social Science and Humanities (UMGESHIC-ISHSSH 2020). pp. 156–165 (2021).
4. Valdez, D., Pickett, A.C., Goodson, P.: Topic Modeling: Latent Semantic Analysis for the Social Sciences. Soc. Sci. Q. 99, 1665–1679 (2018). https://doi.org/10.1111/ssqu.12528.
5. Li, Q., Li, S., Zhang, S., Hu, J., Hu, J.: A review of text corpus-based tourism big data mining. Appl. Sci. 9, 3300 (2019).
6. Murakami, A., Thompson, P., Hunston, S., Vajn, D.: 'What is this corpus about?': using topic modelling to explore a specialised corpus. Corpora. 12, 243–277 (2017).
7. Chauhan, U., Shah, A.: Topic modeling using latent Dirichlet allocation: A survey. ACM Comput. Surv. 54, 1–35 (2021).
8. Evangelopoulos, N.E.: Latent semantic analysis, (2013). https://doi.org/10.1002/wcs.1254.
9. Bellaouar, S., Bellaouar, M.M., Ghada, I.E.: Topic modeling: Comparison of LSA and LDA on scientific publications. In: 2021 4th international conference on data storage

and data engineering. pp. 59–64 (2021).
10. Zhang, W., Kong, S., Zhu, Y., Wang, X.: Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach. Cluster Comput. 22, 12619–12632 (2019).
11. O'callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Syst. Appl. 42, 5645–5657 (2015).
12. Kowsher, M., Hossen, I., Tahabilder, A., Prottasha, N.J., Habib, K., Azmi, Z.R.M.: Support Directional Shifting Vector: A Direction Based Machine Learning Classifier. Emerg. Sci. J. 5, 700–713 (2021).
13. Malley, B., Ramazzotti, D., Wu, J.T.Y.: Data pre-processing. (2019).
14. Yan, D., Li, K., Gu, S., Yang, L.: Network-based bag-of-words model for text classification. IEEE Access. 8, 82641–82652 (2020).
15. Kasture, N.R., Yargal, N., Singh, N.N., Kulkarni, N., Mathur, V.: A Survey on Methods of Abstractive Text Summarization. Int. J. Res. Emerg. Sci. Technol. 7, 728–734 (2014).