Computational Linguistics Using Latent Dirichlet Allocation for Topic Modeling on Wattpad Review

1st Rachma Awantina[1], 2nd Wahyu Wibowo[2]
{rachma905@gmail.com[1], wahyu_w@statistika.its.ac.id[2]}

[1,2]Department of Business Statistics, Institut Teknologi Sepuluh Nopember,
Building TC, 2nd Floor, ITS Campus, Sukolilo, Surabaya, 60111

Abstract. The advancement of the digital era helps human activities because all needs are available in one hand, including reading or writing. The development of digital novel applications makes it easier for readers to access novels through gadgets and self-publishing platforms for writers. Wattpad as the most popular digital novel application on the Google Play Store with a total of more than one hundred million downloads offers a variety of interesting features that can be downloaded for free. The purpose of this research is to find out the results of topic modeling on user reviews of the Wattpad application. The data is taken from the Google Play Store by web scraping technique. The topic modeling uses Latent Dirichlet Allocation (LDA), an unsupervised learning method that is effective in finding different topics in a collection of documents where the document is the observed object, but the word distribution is a hidden structure. Topic modeling generates several words that represent the main topic so that people can understand the latest information about Wattpad.

Keywords: Latent Dirichlet Allocation, Topic Model, User Review, Wattpad

1. Introduction

Literary work is the result of artistic work that uses human objects and everyday life with language as the channeling medium. Literary works are born because of something that makes the soul of the author have a certain sense of events in this world, both personal experience and the reality of life that occurs in society. One of the literary works whose fans of all ages is the novel. The novel presents the story of life in an attractive language style. Novels have a storyline that attracts readers as if they entered the world of the characters in the novel so that they can feel their emotions [1]. Advances in technology make it easier for someone to access novels anytime and anywhere. The impact of these advances is the creation of a free digital novel application that pampers novel readers because they can enjoy their favorite reading on their respective devices, so there is no need to spend money to buy printed novels. Likewise for novel writers who do not need to find a publisher because the digital novel application is a place for self-publishing with the freedom to express their creative ideas. The phenomenon of online-based self-publishing has led to the existence of applications that can publish their work independently [2]. The author doesn't have to finish the novel's script and can slowly complete it while publishing parts that have already been written to the reader.

The advantages of digital novel applications include the availability of a comment column which is a medium of interaction between readers and writers and a place to give suggestions to

ICoSTA 2022, November 01-02, Medan, Indonesia

DOI 10.4108/eai.1-11-2022.2326169

writers, thus helping them to correct errors in their novels. In addition, readers can support the author by pressing the vote button. The most popular digital novel application with more than one hundred million downloads and the most reviews on the Google Play Store is Wattpad [3]. In addition, Wattpad can distribute information directly to users without the need for a distributor, so it is very fast to do. Another digital novel application that is quite popular on the Google Play Store is Dreame. Dreame was launched in August 2018. This application is almost similar to Wattpad, the difference is that there are some additional features such as coins to read so that access to Dreame stories becomes paid [4]. Thus, it raises the urgency of why Wattpad review data needs to be researched using the text mining method, namely topic modeling.

Topic modeling is a text document consisting of words, topics that can be written in many documents expressed by a combination of interrelated words, as well as techniques to conclude hidden topics from a text document. Topic modeling represents each document as a complex combination of several topics and each topic as a complex combination of several words, then also as a text mining tool to classify documents based on the results of topic conclusions [5]. Topic modeling is an algorithm that aims to find topics in unstructured text documents by analyzing how the topics are related to each other. The topic modeling method in this study uses Latent Dirichlet Allocation (LDA). The way LDA works in general is to enter a document set and some specified parameters. Then, produce a model consisting of weights that can be normalized to probabilities. These probabilities appear in two types, namely the probability that a particular document produces a topic in a position and the probability that a particular topic produces a word from a collection of vocabulary [6]. This study uses a topic modeling approach using Latent Dirichlet Allocation (LDA) to determine the topics of conversation related to the Wattpad review. In the remainder of the paper, Section 2 reviews the related work, whereas Section 3 explains the material and method used in this research. Section 4 summarizes the results, and Section 5 concludes this paper with suggestions for future research.

2. Literature Review

2.1 Wattpad

Wattpad is a digital novel application that was launched in December 2006, the result of a collaboration between Allen Lau and Ivan Yuen whose base is in Toronto, Canada. Although it has been around for a long time, this application only became popular in Indonesia around 2016. Wattpad has 15 million users with more than 400 million stories. Wattpad's vision is to entertain and connect the world with a story. Wattpad can also be used as a place to add friends because there is a follow feature. Wattpad is home to more than 65 million people who spend more than 15 billion minutes per month reading stories. Wattpad claims that 90% of its user activity is accessed via mobile and supports more than 50 languages [7]. This proves that in fact many people have their own interests in writing and reading. So with Wattpad, it makes it easier for them to channel that feeling. There are various genres of reading that can be enjoyed such as horror, romance, fantasy, and others. As for publishers, Wattpad makes it easy to find works as well as find out which novels are worthy of publication. Some works that are phenomenal and occupy the best seller position will usually be adapted into series or films.

2.2 Google Play Store

Google Play Store is Google's official application store for devices using the Android operating system. The number of applications on the Google Play Store makes this application store very interesting to be used as an object of research, especially in the field of text mining. One example is topic modeling which is used to find hidden topics from text documents automatically. However, the current obstacle is that Google itself does not provide an Application Programming Interface (API) so that data on the Google Play Store can be integrated with software that is being developed by software developers or as research data. Google provides APIs only for android developers whose applications are registered in the Google Play Store. The API is also very limited, developers can only manipulate certain data from their own applications [8].

2.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a simple algorithm for topic modeling that identifies hidden information in large documents. In machine learning, LDA belongs to the group of unsupervised learning or does not require labeled data. The Latent Dirichlet Allocation algorithm takes topics based on the distribution of multinomial words. Then, the topic is used to generate the word itself based on the multinomial distribution of the topic and these two steps are repeated for all the words in the document [9].

According to Blei [10], Latent Dirichlet Allocation is a generative probabilistic model of a collection of writings called corpus. The basic idea of LDA is that each document is represented as a random mixture of hidden topics, where each topic has a character that is determined based on the distribution of words in it. The LDA algorithm begins by determining the number of topics to be created. After that, the results of the topic modeling will be validated with a coherence score, which is a measure of the semantic similarity between the words that make up the topic. The higher the coherence score means that the topic modeling is said to be good in representing the document.

3. Material and Method

3.1 Data Collection

This study uses Wattpad reviews in English as a data source because Wattpad is the most popular digital novel application on the Google Play Store. The data was taken based on the time period from June 5, 2022 to August 8, 2022 and obtained as many as 5,000 reviews.

3.2 Data Pre-Processing

Data pre-processing is used to extract information from data sources through the identification and exploration of interesting patterns [11]. Based on the irregularity of the text data structure, the data pre-processing process is carried out through Google Colab using the NLTK package which requires several stages so that the text becomes a structured form, namely as follows.

The action taken in the first stage is case folding, which is changing all text characters into lowercase letters and tokenizing is the process of parsing text sentences into words and removing punctuation, spaces, and numbers in the text [12]. The action taken in the second stage is to eliminate stopwords and stemming the words with affixes. Stopwords are vocabulary that is not a unique feature, for example "by", "at", "because", and so on from a data. While stemming is the process of mapping and parsing a word into a basic word form which aims to eliminate affixes in each word.

3.3 Topic Modeling

The concept of topic modeling according to Blei [10] consists of entities, namely "words", "documents", and "corpora". “Word” is considered as the basic unit of discrete data in a document, defined as an item of vocabulary that is indexed for each unique word in the document. “Document” is an array of N words. A corpus is a collection of M documents and corpora is the plural form of corpus. While "topic" is the distribution of some fixed vocabulary. In simple terms, each document in the corpus contains its own proportion of topics discussed according to the words contained in it. The basic idea of topic modeling is that a topic consists of certain words that compose the topic, where in one document it is possible to consist of several topics with their respective probabilities. But in human understanding, documents are observable objects. Meanwhile, topics, distribution of topics per document, and the classification of each word on topics per document are hidden structures, so topic modeling aims to find topics in the hidden structure.This study uses Latent Dirichlet Allocation (LDA) for topic modeling. LDA is a generative model that may be used to model how documents are formed given a set of subjects and their words. LDA begins by identifying the words in each document, then develops a topic mixture for the document based on a predetermined set of subjects, with topic selection primarily based on the document's multinomial distribution, and finally, word selection based on the document’s multinomial distribution. LDA is an unsupervised method that uses related indicators to detect the semantic relationship between words in a group. In this study, the text data is converted into proper topic model structures.

The topic model algorithm executed for a particular number of topics depends on the dataset. In line with previous research, the researchers use the Gensim (RARE Technologies Ltd) [13] coherence model to determine the most appropriate number of topics based on the data. Researchers choose the number of topics that generate the highest coherence value. The higher the coherence score, the easier it is to comprehend the topic's word distribution on which subjects belong to it [14]. Each model's coherent topics are calculated using the coherence score by importing Coherence Model from Gensim, a model library in Google Colab. The number of topics selected ranges from 1 to 10, and the coherence score for each method with k topics is calculated. At the same time, researchers also display the line graph for easy reference of the coherence score versus the number of topics using Microsoft Excel.

Five topics are extracted for further analysis and discussion as this number of topics has the highest coherence score compared to others. The LDA is utilized from the LDA model library in Google Colab to extract keywords form the 5-topic set. After extracting the keywords using LDA, the researchers manually validate and label the topics by referring to the high-frequency keywords. Then, the Intertopic Distance Map is displayed by using the pyLDAvis package to visualize the distance between the topics selected and the frequency of the terms mentioned in each topic.

4. Experimental Results and Analysis

A total of 5000 reviews taken from Google Play Store Site from 5 June 2022 to 8 August 2022 in English. The process of scraping data using the google-play-scraper library in Google Colab. The extracted data is saved in CSV (Comma Separated Values) format consisting of username, time, score, and review. With all these data available, only the review text is concentrated on studying topic modeling with LDA (Latent Dirichlet Allocation) method.

4.1 Topic Modeling using Latent Dirichlet Allocation

1) Determination of the Number of Topics

The number of topics is determined by looking at the highest coherence score which shows the best topic modeling accuracy. Based on experiments to form 10 topics, 5 topics were obtained that were considered good in representing Wattpad reviews.

Table 1. Coherence Score of Each Topic

Number of Topics	Coherence Score
1	0.285268
2	0.284387
3	0.371061
4	0.374428
5	0.453319
6	0.413537
7	0.382189
8	0.431792
9	0.409204
10	0.445741

Based on Table 1, it can be seen that the relationship between the number of topics and the coherence score through the graph is as follows.

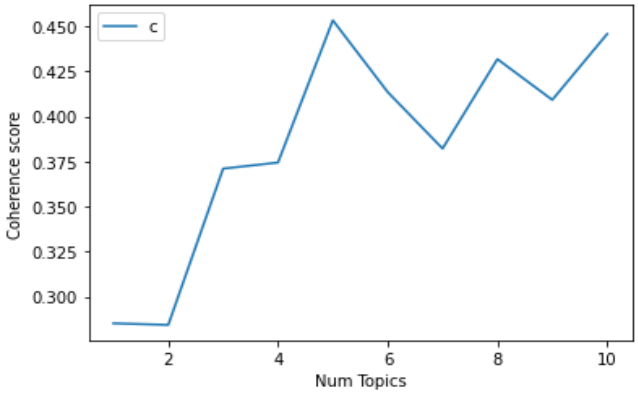

Fig. 1. Coherence Score for LDA

Figure 1 shows that the coherence score experienced an up and down condition, but the highest coherence score was obtained on a total of 5 topics, which was 0.453319. Thus, the number of topics becomes a reference for topic modeling with Latent Dirichlet Allocation.

2) LDA Topic Modeling Results

After obtaining the best number of topics based on the highest coherence score, the LDA model is then generated, each composed of 10 words with a certain weight.

Table 2. Topic Model

Topic	Model
1	0.048*"nice" + 0.036*"amazing" + 0.030*"please_fix" + 0.013*"read" + 0.012*"please_fix_please_fix" + 0.012*"long_time" + 0.010*"super" + 0.010*"stories" + 0.010*"open" + 0.009*"wattpad"
2	0.047*"best" + 0.021*"reading" + 0.016*"ever" + 0.015*"stories" + 0.014*"books" + 0.012*"read" + 0.011*"love" + 0.010*"pretty_good" + 0.010*"great" + 0.009*"old_wattpad"
3	0.125*"good" + 0.094*"love" + 0.016*"cool" + 0.015*"ads" + 0.015*"much" + 0.015*"ads_bottom" + 0.015*"many_ads" + 0.013*"easy_use" + 0.012*"stories" + 0.010*"wattpad"
4	0.041*"great" + 0.039*"really" + 0.026*"awesome" + 0.017*"stories" + 0.016*"love" + 0.014*"like" + 0.011*"read" + 0.010*"amazing" + 0.010*"excellent" + 0.009*"many"
5	0.046*"like" + 0.016*"reading_writing" + 0.014*"enjoy_reading" + 0.014*"wattpad" + 0.014*"perfect" + 0.009*"stories" + 0.009*"amazing" + 0.009*"read" + 0.008*"paid_stories" + 0.008*"reading"

Based on Table 2, it can be seen that the most important keywords for each topic are indicated by the highest weight. The keywords in topic 1 are “nice” at 0.048, topic 2 is “best” at 0.047, topic 3 is “good” at 0.125, topic 4 is “great” at 0.041, and topic 5 is “like” at 0.046.

3) Giving Topic Name

The pyLDAvis visualization illustrates the related words generated by topics, so they can be used for naming topics [15]. PyLDAvis provides two visualization panels. The left side panel shows the topic as a whole and can see the relationship or correlation between a topic with

another topic by looking at the Intertopic Distance Map. While the right side panel shows the frequency distribution of the 30 most relevant words for a particular topic.

a) Topic 1

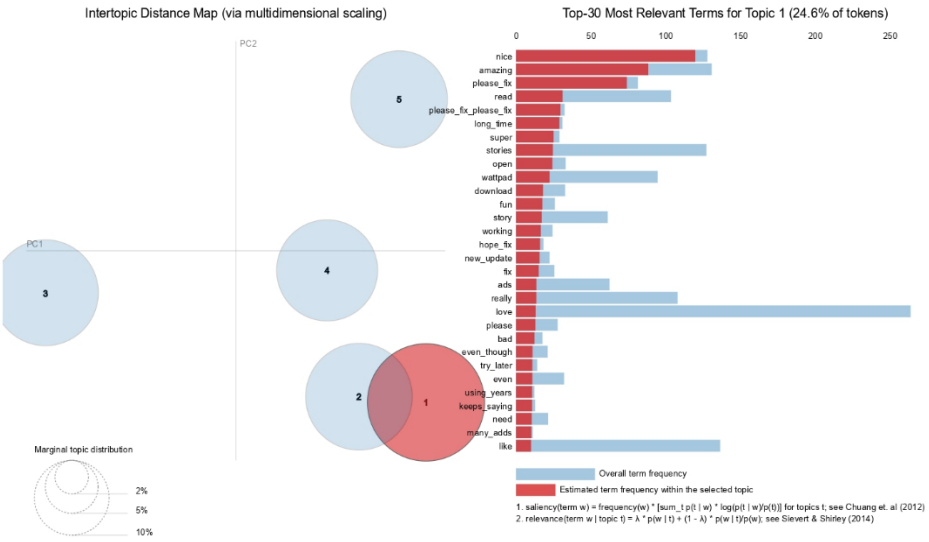

Fig. 2. PyLDAvis for Topic 1

The visualization in Figure 2 consists of several words such as "please fix", "download", "new update", so it will form a topic about user complaints related to problems in the Wattpad application.

b) Topic 2

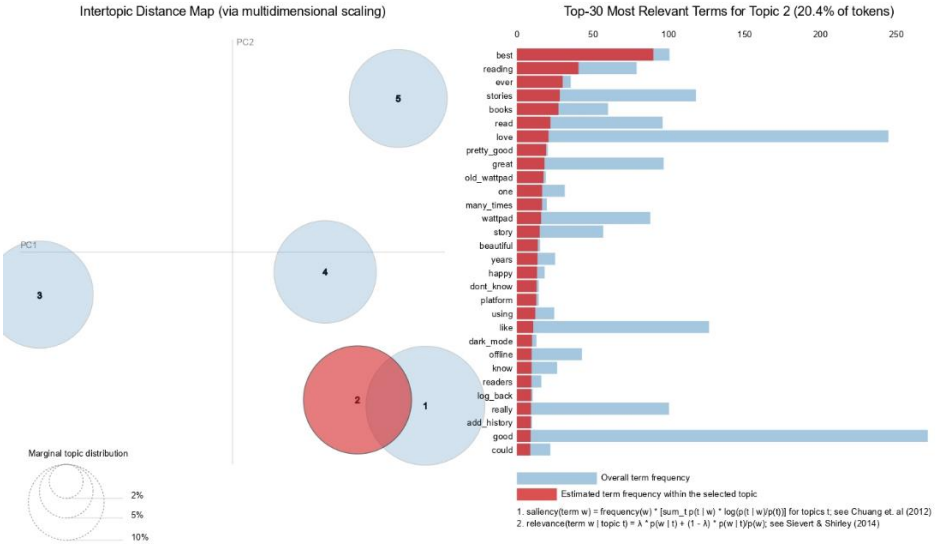

Fig. 3. PyLDAvis for Topic 2

The visualization in Figure 3 consists of several words such as “best”, “stories”, “books”, so it will form a topic about quality stories on Wattpad.

c) Topic 3

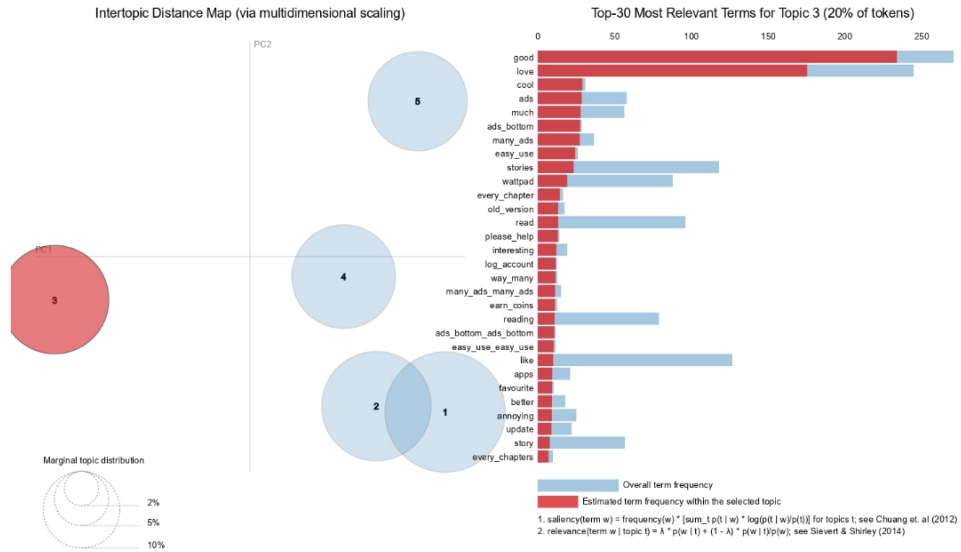

Fig. 4. PyLDAvis for Topic 3

The visualization in Figure 4 consists of several words such as “ads”, “ads bottom”, “many ads”, so it will form a topic the appearance of ads when accessing Wattpad.

d) Topic 4

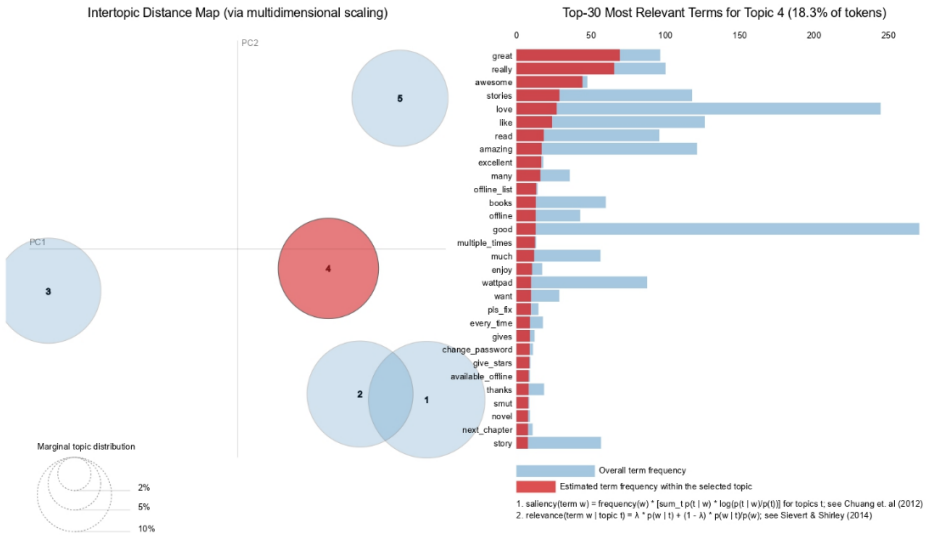

Fig. 5. PyLDAvis for Topic 4

The visualization in Figure 5 consists of several words such as “great”, “awesome”, “amazing”, so it will form a topic about the experience felt as a Wattpad user.

e) Topic 5

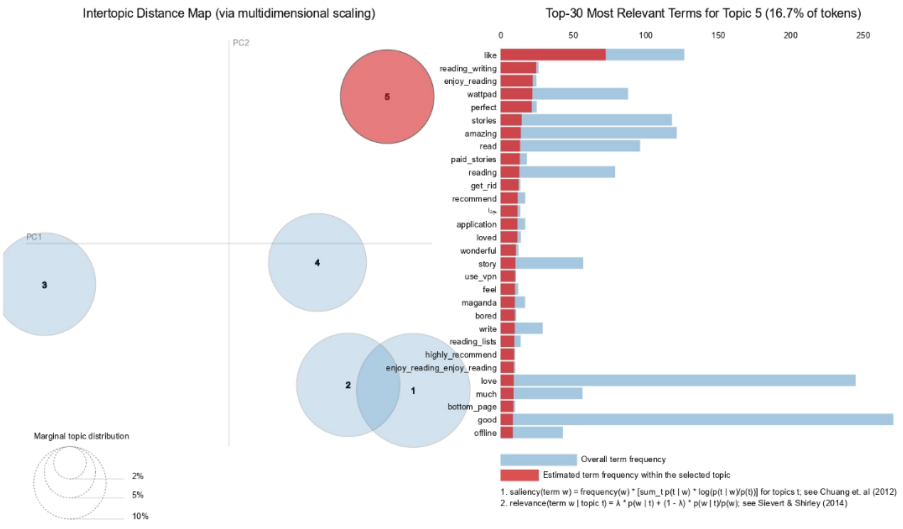

Fig. 6. PyLDAvis for Topic 5

The visualization in Figure 6 consists of several words such as “reading”, “writing”, “paid stories”, so it will form a topic about the opinions of writers and readers about paid stories.

4.2 Insights into Paid Stories on Wattpad

This research focuses on topic 5 about paid stories on Wattpad. Paid stories are one of the facilities provided by Wattpad for writers to earn money from their work [16]. To read paid stories, users need coins which can be purchased on the app. Wattpad uses a freemium business model, where we can download and use basic features for free. However, to be able to enjoy the full features, application users will be charged. There are 2 types of packages in the Wattpad application, namely the basic package which is free and the premium package which is paid. Paid packages are divided into 2, namely Premium and Premium+. Wattpad Premium users will get more features than basic users. These features include reading without being hampered by ads, unlimited offline stories, getting more coins when purchasing coins, and being able to change the theme of the application. This is certainly advantageous compared to basic users, where they can only save 25 offline stories and the appearance of ads while reading. As for Wattpad Premium+ users, they can get all the features in the Premium package as well as access to read selected paid stories based on the package selected. The cost that users must pay to be able to enjoy the Premium package is Rp65.000,- per month, Rp390.000,- per 6 months, or Rp650.000,- per year. Meanwhile, Premium+ users will be charged Rp129.000,- per month, Rp890.000,- per 6 months, or Rp1.309.898,- per year.

5. Conclusion and Recommendations

This study aims to identify topics related to user reviews of the Wattpad application using Latent Dirichlet Allocation (LDA). Based on the results of the topic modeling, 5 topics were found, namely "user complaints related to problems in the Wattpad application", "quality stories on Wattpad", "ads appear when accessing Wattpad", "experiences felt while being a Wattpad user", and “authors and readers' opinions on paid stories”. After the process was complete, the researcher chose to dig up information on the topic 5 related to paid stories to support authors and interesting packages to make it easier for readers that Wattpad users need to know. Future studies could include other digital novel application platforms, such as Dreame. In addition, Wattpad needs to conduct a survey about the price that users are willing to pay to buy the available packages. This is because the tendency of a person from each country is certainly different, so Wattpad needs to adjust how the culture of that country is.

Acknowledgment. The authors acknowledge and gratitude the financial support for this work is from Institut Teknologi Sepuluh Nopember through Research Grant with the contract number 1613/PKS/ITS/2022.

References

[1] Nasution, W., 2016. Kajian Sosiologi Sastra Novel Dua Ibu Karya Arswendo Atmowiloto: Suatu Tinjauan Sastra. *Metamorfosa,* pp. 14-27.

[2] Hane, 2012. *Spotlight on the Self-Publishing Market.* [Online] Available at: http://www.infotoday.com/it/sep12/index.shtml [Accessed 2022].

[3] Farhanah, N. & Yanti, P. G., 2021. *Resepsi Pembaca Novel Digital dalam Aplikasi Wattpad (Studi Kasus Novel Aksa Karya Marionette).* s.l., SENASBASA.

[4] Alea, 2021. *Platform Menulis yang Menghasilkan di Era Pandemi.* s.l.:Elementa Media.
[5] Nugroho, D. D. A. & Alamsyah, A., 2018. Analisis Konten Pelanggan Airbnb pada Network Sosial Media Twitter. *Proceedings of Management*, Volume 5.
[6] Bird, C., Menzies, T. & Zimmermann, T., 2015. *The Art and Science of Analyzing Software Data.* USA: Elsevier.
[7] Ulfa, S. A., 2018. Peranan Aplikasi Wattpad dalam Mengasah Kemampuan Menulis. pp. 1-10.
[8] Ilmawan, L. B., 2018. Membangun Web Crawler Berbasis Web Service untuk Data Crawling pada Website Google Play Store. *Jurnal Ilmiah*, Volume 10, pp. 215-224.
[9] Alfanzar, A. I., Khalid & Rozas, I. S., 2020. Topic Modelling Skripsi Menggunakan Metode Latent Dirichlet Allocation. *Sistem Informasi*, Volume 7.
[10] Blei, D. M., 2012. Probabilistic Topic Models. 55(4), pp. 77-84.
[11] Sanger, J. & Feldman, R., 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.
[12] Weiss, S. M., Indurkhya, N., Zhang, T. & Damerau, F. J., 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information.* New York: Springer.
[13] Röder, M., Both, A. & Hinneburg, A., 2015. *Exploring the Space of Topic Coherence Measures.* s.l., ACM.
[14] Alash, H. M. & Al-Sultany, G. A., 2020. Improve topic modeling algorithms based on Twitter hashtags. *Journal of Physics: Conference Series*, Volume 1660.
[15] Tatsat, H., Puri, S. & Lookabaugh, B., 2020. *Machine Learning and Data Science Blueprints for Finance.* 1 ed. USA: O'Reilly.
[16] Devi, S., 2021. *Pengaruh Perceived Value terhadap Purchase Intention Wattpad Premium Melalui Satisfaction Sebagai Variabel Mediasi; Telaah Pada Pengguna Aplikasi Wattpad.* Thesis. Tangerang: Universitas Multimedia Nusantara.