Analysis of User Comments Based on Topic Modeling using LDA on OVO E-Wallet

Albertus Dwiyoga Widiantoro[1], Bernardinus Harnadi[2]
Department of Information Systems
Soegijapranata Catholic University, Semarang, Indonesia
[1]yoga@unika.ac.id, [2]bharnadi@unika.ac.id

Abstract—Fintech OVO in Indonesia is an important part of cashless payment services. Users take advantage of the commenting service on the Playstore to convey messages to OVO managers. Hundreds of comments always appear every day, and this if not responded to will be a problem. The topic method of the Latent Dirichlet Allocation (LDA) model will be used to analyze the occurrence of user topics. Based on the 6-topic LDA model, we found that the trending topic was in topic 1, with a topic probability value of 0.235. Topic 1 mentions transaction difficulties with premium services with high OVO usage While the ease of transactions has the lowest total probability. The results of this topic can be used as a reference for OVO service providers to focus their performance on improving OVO applications. The impact of this research on service providers is to find out the topics discussed by OVO application users.

Keywords— OVO, LDA, Topic Model, tren, review

I. INTRODUCTION

OVO is one of the FinTech e-wallet services used by people in Indonesia. Android-based OVO user complaints submitted to the OVO Playstore. Research using OVO fintech comment data from Playstore extracts meaningful information, and generates features from data extracted using NLTK [1] for sentiment analysis.

Fintech can be a cross-disciplinary subject that combines finance, technology management and innovation management [2]. Fintech improves financial services processes by using technological solutions in line with existing business models. Fintech can be implemented on mobile devices including wireless, digital assistants, radio frequency devices, and NFC communication-based devices [3]. Fintech is integrated with mobile payments to complete financial transactions [4]. Fintech is a risky but valuable outcome of financial innovation [5], because it generates value for investors [6].

Through an online survey of US food delivery app customers, this study analyzes that app users direct them to use and recommend technology-based services [7]. This indicates that user comments are very important in the development of the company, it is also important to manage information dissemination effectively and strategically [8].

In addition, it was found that user engagement plays a mediating role between users' multiple attitudes and eWOM. The research results help the mobile sensor computing industry to develop effective strategies and build strong consumer-product relationships [9]. eWOM is very effective in brand promotion messages, and consumer experience [10].

A topic model (TM) is a statistical model that tries to find hidden topics in a collection of documents. Topics provide a summary of the corpus that cannot be obtained using document decomposition and similarity metrics for manually searching and understanding documents.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v9i2.5343

TM is used in text mining to find hidden semantic structures or also called latent variables. latent variables are hidden variables in the observed variables, namely documents [11]. TM assumes that the subject is a probability distribution over the vocabulary. TM is a technique to take unstructured text and automatically extract common topics. It's a great way to scan a large collection of text [12].

The key feature that differentiates the topic model from other grouping methods is the notion of mixed membership. However, in most cases it is more realistic to assume that the data actually belongs to more than one group or category. The TM assumes that the topic is a probability distribution over the vocabulary. The vocabulary probability is added to 1 or each topic, but mostly words with lower weight are truncated in the output. Likewise, we can represent individual documents as probability distributions over topics.

LDA requires documents to be rendered as word bags (for the Gensim library). This representation ignores the order of words in the document, but stores information about how many times each word occurs. A good topic model should be able to distinguish the two meanings depending on the context. Since there is no document tagging or human annotations, TM is an example of an unsupervised machine learning technique.

For the Gensim library, the default print behavior is to print a linear combination of the top words, arranged in descending order of possible words occurring in the topic. Then the words that appear on the left are the most significant words of the topic. The get_term_topics and get_document_topics functions are also used to further evaluate the results. get_term_topics returns the probability that a given word belongs to a given topic. Customer engagement continues to grow online and in real life. Online customer engagement generates big data in structured and unstructured formats at high speed [13].

LDA combined with TF-IDF and Doc2Vec increases the variety of feature sets for document classification. The experimental results show that the proposed one is strong against parameter changes [14]. A text representation model that combines Word2Vec and LDA word insertion techniques, that improves accuracy and LDA offers a solution to the high-dimensional and high-sparity problems caused by the BoW model [15].

II. METHOD

Text mining on Fintech OVO user comments is the application of data mining concepts in looking for text patterns from user comments to find useful information for developers or owners of Fintech OVO. The initial stage of text mining for Fintech OVO users is text pre-processing, which aims to prepare user comment text into structured data that can be processed at a later stage. The following are the stages in text preprocessing:

1. the process of removing unnecessary punctuation marks.
2. The stage of converting the entire text to lowercase so that all words are equal.
3. Tokenization, which is the stage of solving Fintech OVO user comments based on the constituent words.
4. Normalization of non-standard words used in Fintech OVO user comments are changed to correct words so that they are as expected.
5. Word deletion.
6. choose the right features, the word dimension reduction stage.

There are several stages in finding topics in the comment data of Fintech OVO users. The steps to get the topic are about the Gibbs sampling algorithm flow on the LDA, the following are the steps to get the LDA model:

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v9i2.5343

1. Fintech OVO data scraping using Python.
2. Doing text preprocessing which aims to reduce the dimensions of words in sentence combinations.
3. Changing Fintech OVO data into a corpus in the form of a term document matrix.
4. Perform the steps to get the LDA model:
   a. Determines the number of topics by looking at the highest possible log value.
   b. Take initialization for topic z randomly starting from {1, ..., K}
   c. At this stage the LDA model produces a probability distribution of the topic of OVO Fintech user comments for each document. Results The topic distribution has a Dirichlet distribution with alpha and beta parameters. The Dirichlet distribution was chosen as the previous distribution on the grounds that the probability of occurrence of a topic in OVO user comments describes the probability of occurrence of a vector in each document. This also means that each user comment topic that appears in each document is a probability vector.
   d. This stage produces a probability distribution of words that are on the user's comment topic specified in each document. The distribution of words on the topic of OVO user comments has a Dirichlet distribution with parameters. The Dirichlet distribution was chosen as the initial distribution because the probability of word occurrence from user comments describes the probability vector on the topic specified in each document. It can also be interpreted that the probability vector is in each user's comment on the specified topic in each document.
   e. The next step generates a probability distribution of the topic of OVO Fintech user comments for the topics that have been determined in the document based on their appearance. This distribution process can be referred to as a multinomial distribution. The multinomial distribution was chosen because it describes the probability of occurrence of Fintech OVO user comment topics between topic 1 to topic k for each user comment document.
   f. Returns the probability distribution for the words in the corpus (w) in the selected user's comment topic. This probability distribution is also called a Multinomial distribution. The results of this alternative Multinomial distribution illustrate the possibility of a word appearing among the words in the corpus of comments by Fintech users.
   g. This stage produces a probability distribution, then a joint posterior distribution is obtained which is an LDA probability model from the comments of Fintech OVO users.
5. Determine the topic in the formed LDA Model.
   Analyzing the topics formed is based on the highest probability of OVO Fintech user comment topics.

The purpose of this study is to identify the trend of OVO users' comments based on the topic of the model. The results of this study are useful to provide insight for companies to find out user comments so that companies can find out what topics are formed within 1 year.

III. RESULTS AND DISCUSSION

A. Processing using Google Collab

There are 2 ways of processing data using Google Collab, namely the data is placed on Google Drive and the processing is on Google Drive, the second way is that the data on Google Drive is entered into the Collab virtual space. Processing data directly using the Google Drive script as below.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v9i2.5343

```
from google.colab import drive
drive.mount('/content/drive')
path = '/content/drive/My Drive/dataset/'
df = pd.read_csv(path+'data/ovodataset1.csv', sep=';')
```

For reading data, the second way is to install PyDrive, after that authenticate the user and then mount the data. The data files on Google Drive are shared with anyone with link access and the role is changed to viewer, as shown in Figure 1.

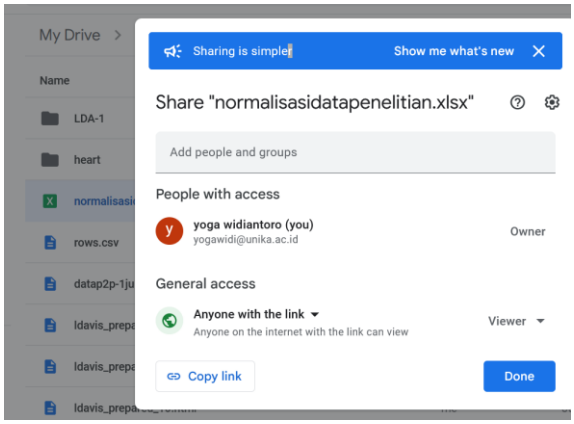

Figure 1. File Sharing in Google Drive

Links that have been formed like:

```
#https://drive.google.com/file/d/14T2ZVq9RyssFSyLbuIPEaNrfU7urgBXz/view?us
```

Take the share ID used to download data with the command

```
“
! gdown --id 14T2ZVq9RyssFSyLbuIPEaNrfU7urgBXz”.

!pip install -U -q PyDrive
import os
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials

auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials= GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)
from google.colab import drive
drive.mount ('/content/drive')

#https://drive.google.com/file/d/14T2ZVq9RyssFSyLbuIPEaNrfU7urgBXz/view?usp=sharing
! gdown --id 14T2ZVq9RyssFSyLbuIPEaNrfU7urgBXz
```

Data Pre-processing

Data that has been read will do some processing to eliminate data duplication, delete words that are not needed, delete symbols, delete punctuation marks.

Data that has been read will do some processing to eliminate data duplication, delete words that are not needed, delete symbols, delete punctuation marks. Using NLTK you can remove tab, new line, back slice, remove non-ASCII (emoticon, Chinese word), remove mention, link, hashtag, remove incomplete URL, remove number, remove punctuation, remove leading & trailing whitespace, remove multiple whitespace into single whitespace, remove single char.

NLTK stopwords are used to remove words that are not used in the analysis of the model topic, usually words that are commonly used in daily conversation, this process is determined by the researchers themselves.

```
from nltk.corpus import stopwords
  list_stopwords = stopwords.words('indonesian')
  Stopword dapat ditambahkan dengan menggunakan perintah list_stopwords.extend(["yg", "dg", 'yah'])tambahan ini dimasukkan dalam perintah berikut ini
  def stopwords_removal(words):
    return [word for word in words if word not in list_stopwords]
  df['textdata_tokens_WSW'] = df['textdata_tokens'].apply(stopwords_removal)
  print(df['textdata_tokens_WSW'].head())
```


ISSN 2442-7888 (online) DOI 10.24167/sisforma.v9i2.5343

Normalization stage replaces certain words with other words that are more appropriate. This process creates a normalization table as shown in table 1 below:

Table 1. Normalization of Words

before	after
gak	tidak
ngga	tidak
tdk	tidak
sdh	sudah
bermutuh	bermutu

Table 1 will be read by the system and then used for the data normalization process.

```
normalizad_word=
pd.read_excel('/content/normalisasidatapenelitian.xlsx
');
normalizad_word_dict = {}
for index, row in normalizad_word.iterrows():
  if row[0] not in normalizad_word_dict:
    normalizad_word_dict[row[0]] = row[1]
def normalized_term(document):
  return [normalizad_word_dict[term] if term in
normalizad_word_dict else term for term in document]
df['textdata_normalized'] =
df['textdata_tokens_WSW'].apply(normalized_term)
```

The next process is to do stemming using Literature. Python Sastrawi is a simple library that can convert words with Indonesian affixes into their basic form. This process makes it easy to match each word.

A. LDA Implementation Using Gensim

Determine the topic of OVO comments to be extracted, and determine the number of words per topic that is considered appropriate, so that each topic does not have the same topic meaning in this process, total_topics =6 and number_words =8 are determined.

The LDA setting in this research is

```
lda_model = Lda(doc_term_matrix,
num_topics=total_topics, id2word =
dictionary, random_state=100, passes=10,
alpha=0.01, eta=0.10,
per_word_topics=True,
chunksize=100).lda_model.show_topics(nu
m_topics=total_topics,
num_words=number_words)
```

Determination of the number of topics of 6 and the number of words in the topic of 8 is determined by repeated experiments so that in each topic there are no intersections between topics. With no intersection between topics, topic interpretation becomes easier.

B. Feature Interpretation

The results of the interpretation of the topics that have been built are top-up for credit services, lots of bonuses, failed premium processes, good bank transfers, difficulty updating, conditions cannot be used, satisfaction, verification difficulties.

C. Determining the Number of Topics

The initial stage of TM LDA will determine the right number of topics for Fintech OVO data. If the wrong number of topics is selected, this will sub-optimal performance resulting in incorrect OVO topics.

Doing word analysis on the LDA model must be careful because it is possible to find the fact that not all words on the topic can be interpreted correctly, the results of the LDA model do not produce good convergence with the resulting topics do not lead to the same discussion.

The results of the LDA process on Fintech OVO reveal 6 topics and consist of 8 user comments that have the highest probability of appearing in each topic of Fintech OVO. The words of OVO users' comments are arranged on each topic based on their similarities, compiling a specific topic. The topic of comments by Fintech OVO users is said to be convergent if the distribution of the words that make up the topic leads to the discussion of the same topic.

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v9i2.5343

Topic interpretation of Fintech OVO user comments of the LDA model is presented in Table 2.

D. Trend Topic Analysis

Detection of trending topic comments by Fintech OVO users can be detected by LDA well, where LDA can capture events well with a narrow topic coverage. To determine trending topic data on OVO Fintech user comments, it is done by looking at the highest probability value of each compiled topic. The probability value of the OVO Fintech topic from topic 1 to topic 6 is presented in Table 3.

From Table 3 it can be seen, the highest Fintech OVO topic is in topic 1, which is with a value of 0.235, on the interpretation of transaction difficulties on premium services. And the lowest review is the ease of transactions.

Table 3. Feature Probability

No	Interpretation	Total probability
1	Transaction difficulties on premium services	0.235
2	The balance hasn't arrived at the time of topup	0.216
3	Fun app	0.201
4	Slow in responding to reports	0.198
5	Good service but busy app	0.192
6	Easy transaction	0.19

ISSN 2442-7888 (online) DOI 10.24167/sisforma.v9i2.5343

Table 2. OVO features

No. topic	1	2	3	4	5	6
Topic interpretation	The balance hasn't arrived at the time of topup	Easy transaction	Fun app	Good service but busy app	Slow in responding to reports	Transaction difficulties on premium services
keyword	0.046* saldo	0.031* bayar	0.034* banget	0.039* sering	0.057* makin	0.047* transfer
	0.045* masuk	0.029* mudah	0.032* buka	0.033* bagus	0.049* pulsa	0.044* uang
	0.036* sudah	0.025* bantu	0.026* suka	0.031* baik	0.019* moga	0.027* gagal
	0.023* belum	0.024* lebih	0.026* sudah	0.021* download	0.018* hp	0.026* susah
	0.021* akun	0.021* transaksi	0.022* bintang	0.020* sistem	0.016* layan	0.024* upgrade
	0.016* login	0.021* guna	0.022* update	0.019* keluar	0.014* beli	0.023* transaksi
	0.015* top-up	0.020* cepat	0.020* sekarang	0.015* tarik	0.013* lapor	0.022* bank
	0.014* kirim	0.019* barang	0.019* baru	0.014* sibuk	0.012* lambat	0.022* premium

IV. CONCLUSION

Data processing of Fintech OVO users' comments utilizing the LDA algorithm produces 6 topics as be seen in table 3. They are sorted by the highest to lowest total probability of topics. It was found that the trending topic was on topic 1 with a topic probability value of 0.235. The term that appears in topic 1 means that the difficulty of transactions on premium services with high OVO usage, this indicates that there is a need to improve services in the process of moving to premium services. While the ease of transactions has the lowest total probability. The results of this topic can be used as a reference for OVO service providers to focus their performance on improving OVO applications. The impact

of this research on service providers is to find out the topics discussed by OVO application users.

ACKNOWLEDGMENT

We appreciation to the Directorate General of Higher Education, Ministry of Education and Culture of the Republic of Indonesia for funding this research and appreciate to the Information Systems Department, Soegijapranata Catholic University, Indonesia.

REFERENCES

[1] A. D. Widiantoro, A. Wibowo, and B. Harnadi, "User Sentiment Analysis in the Fintech OVO Review Based on the Lexicon Method," *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, 2021, doi: 10.1109/ICIC54025.2021.9632909.

[2] E. Z. Milian, M. D. M. Spinola, and M. M. De Carvalho, "Fintechs : A Literature Review and Research Agenda," *Electron. Commer. Res. Appl.*, p. 100833, 2019, doi: 10.1016/j.elerap.2019.100833.

[3] W. A. Alkhowaiter, "Digital payment and banking adoption research in Gulf countries: A systematic literature review," *Int. J. Inf. Manage.*, vol. 53, no. February, p. 102102, 2020, doi: 10.1016/j.ijinfomgt.2020.102102.

[4] S. Chandra, S. C. Srivastava, and Y.-L. Theng, "Evaluating the Role of Trust in Consumer Adoption of Mobile Payment Systems: An Empirical Analysis," *Commun. Assoc. Inf. Syst.*, vol. 27, no. 1, 2010, doi: 10.17705/1cais.02729.

[5] A. V Thakor, "Incentives to innovate and financial crises $," *J. financ. econ.*, vol. 103, no. 1, pp. 130–148, 2012, doi: 10.1016/j.jfineco.2011.03.026.

[6] M. A. Chen, Q. Wu, and B. Yang, "How Valuable Is FinTech Innovation?," *Rev. Financ. Stud.*, vol. 32, no. 5, pp. 2062–2106, 2019, doi: 10.1093/rfs/hhy130.

[7] D. Belanche, M. Flavián, and A. Pérez-Rueda, "Mobile apps use and WOM in the food delivery sector: The role of planned behavior, perceived security and customer lifestyle compatibility," *Sustain.*, vol. 12, no. 10, 2020, doi: 10.3390/su12104275.

[8] Y. K. Na and S. Kang, "Sustainable diffusion of fashion information on mobile friends-based social network service," *Sustain.*, vol. 10, no. 5, pp. 1–23, 2018, doi: 10.3390/su10051474.

[9] Y. Zhao, Y. Liu, I. K. W. Lai, H. Zhang, and Y. Zhang, "The impacts of attitudes and engagement on electronicword of mouth (eWOM) of mobile sensor computing applications," *Sensors (Switzerland)*, vol. 16, no. 3, 2016, doi: 10.3390/s16030391.

[10] S. Tabassum, M. G. Khwaja, and U. Zaman, "Can narrative advertisement and eWOM influence generation z purchase intentions?," *Inf.*, vol. 11, no. 12, pp. 1–16, 2020, doi: 10.3390/info11120545.

[11] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010, doi: 10.1109/MSP.2010.938079.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.

[13] S. Bagherzadeh, S. Shokouhyar, H. Jahani, and M. Sigala, "A generalizable sentiment analysis method for creating a hotel dictionary: using big data on TripAdvisor hotel reviews," *J. Hosp. Tour. Technol.*, vol. 12, no. 2, pp. 210–238, 2021, doi: 10.1108/JHTT-02-2020-0034.

[14] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," *Inf. Sci. (Ny).*, vol. 477, pp. 15–29, 2019, doi: 10.1016/j.ins.2018.10.006.

[15] D.-D. Science, "A Method of Short Text Representation Based on the Feature Probability Embedded Vector," 2021.