Topic Modelling Twitter Data with Latent Dirichlet Allocation Method

Edi Surya Negara
Department of Informatics,
Data Science Interdisciplinary
Research Center
Universitas Bina Darma
Palembang, Indonesia
e.s.negara@binadarma.ac.id

Dendi Triadi
Department of Informatics,
Data Science Interdisciplinary
Research Center
Universitas Bina Darma
Palembang, Indonesia
dendi.triadi@binadarma.ac.id

Ria Andryani
Department of Informatics,
Data Science Interdisciplinary
Research Center
Universitas Bina Darma
Palembang, Indonesia
ria.andryani@binadarma.ac.id

Abstract— **Twitter is a popular social media for every user to issue thoughts and emotional forms which are tweets, tweets that only have 140 characters with limitations to write in text. Twitter is one of the social media places to get information that is always up to date, tweets are categorized into big data because tweets are information that can be used as a source of data for research. Latent Dirichlet Allocation (LDA) as an algorithm that can process large text data (big data). In this study using the LDA method as an algorithm to produce topic modeling, each topic similarity, and visualization of topic clusters from the tweet data generated as many as 4 topics (Economic, Military, Sports, Technology) in Indonesian, where each topic has a number different tweets. The LDA method used in the processing of tweet data is successfully carried out and works optimally, in each topic extraction, topic modeling, generating index words that are in each topic cluster and computer visualization in the topic.LDA output shows optimal performance in the process of word indexing in Sport topics with 1260 tweets with an accuracy of 98% better than the LSI method in Topic Modeling.**

Keywords*— *Twitter, Topic Modelling, Latent Dirichlet Allocation

I. Introduction

Social media is an intracation media that continues to grow following the development of web technology. Social media is defined as an online information technology tool that allows every user to communicate easily through the internet in sharing information such as text messages, audio, video, images and so on [1]. In addition, social media is defined as an intermediary media that has a social orientation that can record user conversations and provides facilities for sharing information.

Twitter is present as a means of communication to exchange information about various events in the real world, short messages on Twitter generally reflect various events experienced by users in real-time [2]. Twitter is one of the media that connects each user with communication using tweets in the form of text, where each tweet is a liaison between users to share information with each other. Text is one of the media that is able to communicate in various ways, and text with a large size is also categorized into Big Data, where the growth of Big Data is currently growing, trillions of bytes are created every day from various sources. Big data is a term for a set of data that is so large or complex that it cannot be handled anymore with conventional computer technology systems [3].Utilization of big data using various methods and algorithms such as community detection has been widely used in various fields such as measuring the strength of Indonesian research with a social network analysis approach [16] [17].

Topic modeling is a technique developed to produce document representation in the form of keywords from documents. These keywords will be used in the indexing and document search process to be found again according to user needs. Latent Semantic Analysis (LSA) appears as the first technique that can produce document representation in the form of a collection of words. LSA is the method most widely known by attaching the Bag-of-Words feature as a document representation [4]. Probabilistic Latent Semantic Analysis (PLSA) developed by Hoffman in 1999 is an LSA that uses probabilistic values as a determinant of the topic weight of each existing document. As a new variant of LSA, the GLSA technique was proposed by Islam and Hoque in 2012 which changed the existence of terms in the term matrix to n-grams. Meanwhile, the method named Multidimensional Latent Semantic Analysis (MDLSA) proposed by Zhang, Ho, Wu, and Ye in 2013 is a method that reviews the relationship between terms and spatial distribution. The technique that involves syntactic aspects directly is SELSA (Syntactically Enhanced Latent Semantic Analysis) proposed by Kanejiya, Kumar, and Prasad in 2003. To model generative probabilistic on a collection of text data (corpus) a method called Latent Dirichlet Allocation (LDA) was used. LDA is a Bayesian Hierarchy model, in which a set of text data is modeled as a mixed model of various topics [5]. Topic modeling is a method for finding the main theme that includes a large and unstructured collection of documents that can arrange a dataset in accordance with the themes found in it. Topic modeling algorithms can be applied to large numbers of documents and can be adapted to various types of data, among applications, topic modeling has been used to find patterns in genetic data, images, and social networks. Topic modeling algorithm is a statistical method that analyzes words from the original text to find themes in the dataset, how the themes are connected to each other, and how they change over time [6].

In summary, the development of topic modeling algorithms began with the emergence of the Latent Semantic Indexing (LSI) algorithm which attempts to overcome the problems faced in the tf-idf scheme about reducing dimensions that are too large. Then the LSI method was developed using the maximum likelihood or Bayesian method which applies the concept of probability, known as the LSI Probabilistic (PLSI) method. Although the PLSI method can be useful for modeling probabilistic topics, it does not fully produce probabilistic models at the document level. To overcome this problem, then came the Latent Dirichlet Allocation method [7]. The LDA method can indeed work well at the document level which may have many topics in it, but the LDA cannot work optimally on topic modeling for datasets that contain concise text, such as the Twitter dataset. For that, we need the right method for modeling topics in a concise dataset, then LDA is developed into twitter-LDA for modeling topics in the twitter dataset [8].

With the Latent Dirichlet Allocation (LDA) method, the writer wants to group (clustering) in topic modeling (Topic Modeling) using Twitter data (tweet) which results in the grouping of Indonesian tweet data and analysis in grouping each topic from Twitter social media.

II. Literature Review

A. Twitter

Twitter is one of the most popular social media in the world since it was first published in 2006. Indonesia has used Twitter since it was first published and is among the most productive Twitter users. There are around 29 million Twitter users in Indonesia. In 2014, Indonesia ranked fifth as the country with the most tweets. Various information can be obtained from social media Twitter which is now increasingly being used. Users use it for various needs, for example for public, government and business needs. Tweets shared by Twitter users cover a variety of specific topics. Users can share opinions about the shared tweet. Topics represent the contents of many tweets that discuss the same context. From these tweets can be found the main topics that are being discussed by many users at that time by conducting topic analysis. Twitter is present as a means of communication to exchange information about various events in the real world, short messages on Twitter generally reflect various events experienced by users in real-time [2].

B. Topic Modelling

The concept of topic modeling consists of entities namely "words", "documents", and "corpora". "Word" is considered as the basic unit of discrete data in a document, defined as items of vocabulary that are indexed for each unique word in the document. "Document" is an arrangement of N words. A corpus is a collection of M documents and corpora is the plural form of the corpus. While "topic" is the distribution of some fixed vocabulary. Simply put, each document in the corpus contains its own proportions of the topics discussed according to the words contained in them [7]. In the topic modeling concept, there are several definitions of the topic model including:

a. The topic model is a type of statistical model for finding abstract "topics" that occur in document collections.
b. The topic model is a set of algorithms that uncover the thematic structures hidden in a collection of documents. This algorithm helps us develop new ways to search, search and summarize large text archives.
c. The topic model provides a simple way to analyze large volumes of text without labels. "Topic" consists of a group of words that often occur together.

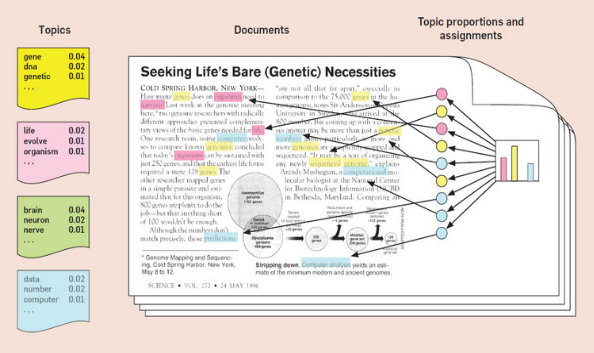


Fig. 1 Probabilistic topic model[6]

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the most popular topic modeling and analysis topic at the moment. LDA emerged as one of the chosen methods for analyzing very large documents. LDA can be used to summarize, cluster, connect or process very large data because LDA generates a list of topics that are weighted for each document [9]. The distribution used to obtain the distribution of per-document topics is called the Dirichlet distribution, then in the generative process for LDA, the results from Dirichlet are used to allocate words in the document to different topics. In LDA, documents are observable objects, while topics, distribution of per-document topics, classification of each word on topics per document are hidden structures, hence this algorithm is called Latent Dirichlet Allocation [6].

Procedure:

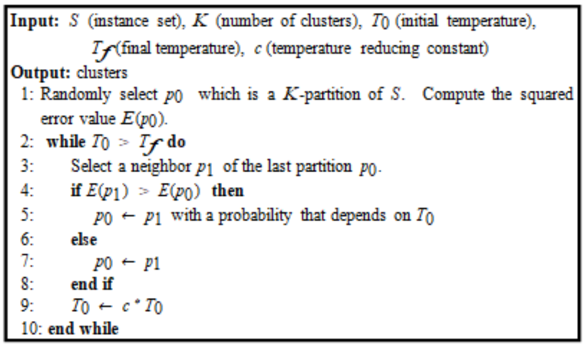
```
Input:  Number of document M
        Number of topics t
        β Vocabulary matrix

Output: Topic probability distribution for each word in document.
  Steps:
  1. Choose the topic distribution α
  2. Assign each word W in a document d to one of the t topics.
  3. For each word W in a document d
     • For each topic calculate P(Topic t | Document d)
     • Calculate P(word W | Topic t)
  4. The selection word W for a topic t is depends on the distribution of β Vocabulary words
```

In general, LDA works with the input of individual documents and several parameters, to produce outcomes in the form of a model consisting of weights that can be normalized according to probability. This probability refers to two types, namely type (a) the probability that a certain specific document produces a specific topic and type (b) the probability that a specific topic produces specific words from a vocabulary collection. Probability of type (a), documents that have been labeled with a list of topics are often continued to produce probability types (b), which produce certain specific words [9]. Simply put, the LDA algorithm is based on the distribution of words contained in a document. Then look for whether the word is from the same topic in a document, and there are several topics.

D. Latent Dirichlet Allocation

Clustering considers an important approach to finding common ground in data and placing the same data into groups. Clustering divides data sets into groups where the similarity within a group is greater than between groups [10]. The idea of clustering has a simple nature and is close to human thinking, whenever we present this large amount of data into a small number of groups or categories to facilitate further analysis. Apart from that most of the data collected in many problems appears to have some inherent nature that has experienced natural groupings [11].

Document clustering is the process of grouping document datasets referring to the similarity (similarity) of document data patterns into a cluster, whereas those without similarities will be grouped into other clusters [12]. Clustering is defined as an effort to group data into clusters so that the data in the same cluster have more similarities compared to the data in different clusters [13]. Clustering is an unsupervised machine learning technique where the method does not need to be trained or there is no learning phase.



```
Input: $S$ (instance set), $K$ (number of clusters), $T_0$ (initial temperature),
        $T_f$ (final temperature), $c$ (temperature reducing constant)
Output: clusters
Randomly select $p_0$ which is a $K$-partition of $S$. Compute the squared
  error value $E(p_0)$.
while $T_0 > T_f$ do
    Select a neighbor $p_1$ of the last partition $p_0$.
    if $E(p_1) > E(p_0)$ then
        $p_0 \leftarrow p_1$ with a probability that depends on $T_0$
    else
        $p_0 \leftarrow p_1$
    end if
    $T_0 \leftarrow c * T_0$
end while
```

Fig. 2 Clustering Method [14]

Input: S (instance set), K (number of cluster)
Output: clusters

1. Initialize K cluster centers.
2. while termination condition is not satisfied do
3. Assign instances to the closest cluster center.
4. Update cluster centers based on the assignment.
5. end while

III. Methodology

In this study consists of several phases so as to obtain accurate analysis results. This study uses the Latent Dirichlet Allocation (LDA) method to model the topic and produce clustering in each topic discussion. The tools used are Python computer programs to process data with the LDA method algorithm. The research approach broadly consists of six stages, including:

1. Start or preparation. In the preparation phase of the study, the authors make observations (observations) in advance on the object directly by looking at how the LDA algorithm works in a python program to process text data tweets.
2. Data Crawler. After understanding the phenomenon, then the authors do a data crawler using the Python program.
3. Latent Dirichlet Allocation (LDA). LDA is a method used for filtering and tokenizing in the text of the crawler data that is represented in the form of clusters.
4. Results and Discussion. At this stage discussing the results
5. Conclusion. This stage, the authors draw conclusions from the results and discussions that have been carried out in the previous stage.

For more details, the authors include a step diagram in this study which is shown in Figure 3 below;:

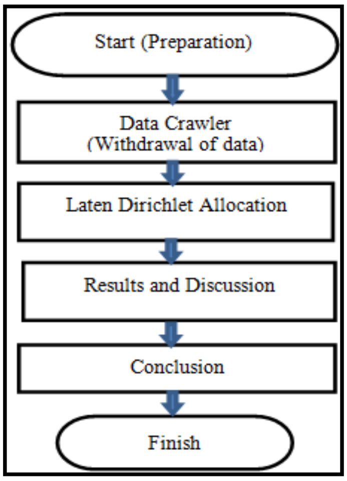


Fig. 3 Research Stage

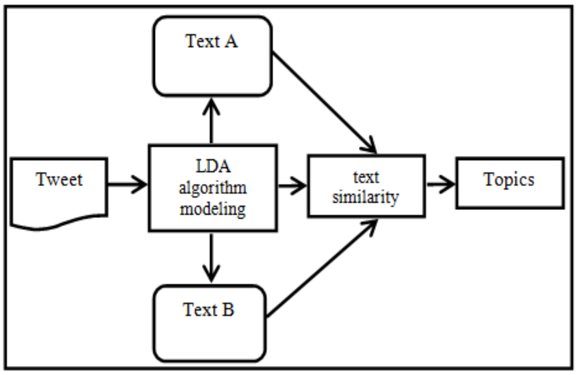
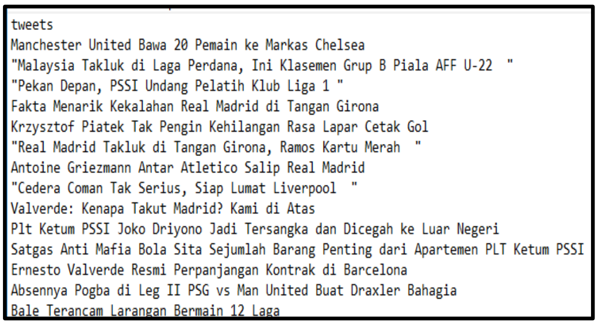


Fig. 4 The concept of Similarity of texts in a topic[15]

In Figure 4 above, a process / method of processing tweet data has been generated, first extracting tweet data, then making a model with the LDA algorithm that can produce text from each topic that can classify text, see index results and similarity in each cluster topic [15].

IV. Experimental and Result

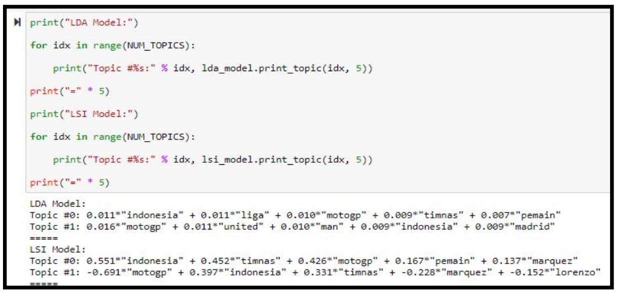
This research uses Twitter data that is crawled with python program code and access token provided by twitter, then the coding is entered into notepad ++ editor and run with command prompt. After the twitter crawler process was carried out successfully pulled Indonesian twitter data with the file.txt format as many as 4 different topics, and each tweet had a different number, namely topics about Economy totaling 3219 tweets, Military topics 2390 tweets, Sports topics 1260 tweets, and topics Technology 3215 successful tweets on the crawler. Twitter data retrieval using keywords in accordance with the topic, then stored in txt format in Figure 5.



```
tweets
Manchester United Bawa 20 Pemain ke Markas Chelsea
"Malaysia Takluk di Laga Perdana, Ini Klasemen Grup B Piala AFF U-22  "
"Pekan Depan, PSSI Undang Pelatih Klub Liga 1 "
Fakta Menarik Kekalahan Real Madrid di Tangan Girona
Krzysztof Piatek Tak Pengin Kehilangan Rasa Lapar Cetak Gol
"Real Madrid Takluk di Tangan Girona, Ramos Kartu Merah  "
Antoine Griezmann Antar Atletico Salip Real Madrid
"Cedera Coman Tak Serius, Siap Lumat Liverpool  "
Valverde: Kenapa Takut Madrid? Kami di Atas
Plt Ketum PSSI Joko Driyono Jadi Tersangka dan Dicegah ke Luar Negeri
Satgas Anti Mafia Bola Sita Sejumlah Barang Penting dari Apartemen PLT Ketum PSSI
Ernesto Valverde Resmi Perpanjangan Kontrak di Barcelona
Absennya Pogba di Leg II PSG vs Man United Buat Draxler Bahagia
Bale Terancam Larangan Bermain 12 Laga
```

Fig. 5 Topic tweets

The Latent Dirichlet Allocation (LDA) method as an algorithm used in this study was combined with coding python to get the results of representative topics in Indonesian tweet data using Indonesian corpus. In the process of running the gensim LDA model with Indonesian corpus and stopwords, it produces 5 words in the topic produced by the LDA algorithm with a range/index of each word in each topic. In topic LDA modeling, a comparison of the results of the indexing of words produced in the modeling of the topic with the Latent Semantic Indexing (LSI) method to see the comparison of the index results of each word. Before the emergence of the LDA method as an algorithm used to do modeling in a topic, LSI was a method used as an algorithm used to produce existing word indexing on each topic. The results of the word index in topic modeling using the LDA and LSI methods are shown in Figure 6.



```
print("LDA Model:")

for idx in range(NUM_TOPICS):

    print("Topic #%s:" % idx, lda_model.print_topic(idx, 5))

print("=" * 5)

print("LSI Model:")

for idx in range(NUM_TOPICS):

    print("Topic #%s:" % idx, lsi_model.print_topic(idx, 5))

print("=" * 5)
```

```
LDA Model:
Topic #0: 0.011*"indonesia" + 0.011*"liga" + 0.010*"motogp" + 0.009*"timnas" + 0.007*"pemain"
Topic #1: 0.016*"motogp" + 0.011*"united" + 0.010*"man" + 0.009*"indonesia" + 0.009*"madrid"
=====
LSI Model:
Topic #0: 0.551*"indonesia" + 0.452*"timnas" + 0.426*"motogp" + 0.167*"pemain" + 0.137*"marquez"
Topic #1: -0.691*"motogp" + 0.397*"indonesia" + 0.331*"timnas" + -0.228*"marquez" + -0.152*"lorenzo"
=====
```

Fig. 6 *LDA and LSI Modeling word index*

In the process of modeling topics from 4 different topics with 2 different methods used as an algorithm in the processing of data topics, tweets have been done one by one with different results. The results of each topic show that the LDA method works more effectively than the LSI method of carrying out the process of modeling topics as an algorithm in indexing words in topics with 5 words that have different indexes. In this case, LSI and LDA have the representation, accuracy, strengths, and weaknesses of each in the topic modeling process. The LSI method used in the modeling process produces words that often appear together in documents must be similar. The LSI covariance direction or main direction is obtained by using the singular value decomposition of□ ∈ R $^{d\,x\,N}$

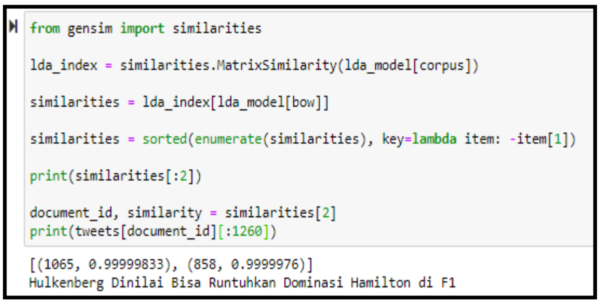
$\square = USV^{T}$, with$U^{T}U = I_{d}$and$V^{T}V = I_{N}$
and $S \in R^{dxN}$ a matrix with nonzero elements only on single diagonal values□, positive and ordered in descending order [5]
LSI index algorithm in words:
Expectation step

$$q_{ink}^{(t)} = p(z_{ink} = 1 \mid \mathbf{w}_{in}\,;\, \mathbf{d}_{i}^{(t-1)}, \mathbf{B}^{(t-1)}) = \frac{d_{ik}^{(t-1)}\, b_{j_{in}^{*}k}^{(t-1)}}{\sum_{k'=1}^{K} d_{ik'}^{(t-1)}\, b_{j_{in}^{*}k'}^{(t-1)}} \quad (1)$$

Maximization step

$$d_{ik}^{(t)} = \frac{\sum_{n=1}^{N^{(i)}} q_{ink}^{(t)}}{\sum_{n} \sum_{k'=1} \cdots} = \frac{\tilde{N}_{k}^{(i)}}{N^{(i)}} \quad \text{and} \quad b_{ik}^{(t)} = \frac{\sum_{i=1}^{M} \sum_{n=1}^{N^{(i)}} q_{ink}^{(t)}\, w_{inj}}{\sum_{i=1} \sum_{n=1} \cdots} \quad (2)$$

On the results of the topic with the LDA model, there is a similarity (similarity) in each topic produced. In each modeling with the LDA model produces 1 topic that is used as similarity in the discussion of tweet data topics. The results of the similarity in each topic produced are shown in Figure 7.

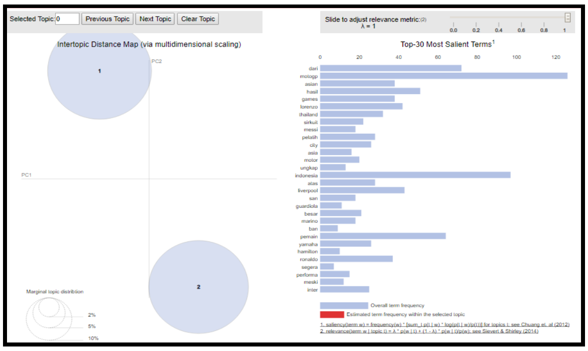


```
from gensim import similarities

lda_index = similarities.MatrixSimilarity(lda_model[corpus])

similarities = lda_index[lda_model[bow]]

similarities = sorted(enumerate(similarities), key=lambda item: -item[1])

print(similarities[:2])

document_id, similarity = similarities[2]
print(tweets[document_id][:1260])
```

```
[(1065, 0.99999833), (858, 0.9999976)]
Hulkenberg Dinilai Bisa Runtuhkan Dominasi Hamilton di F1
```

Fig. 7 Similarity in topic

Clustering of tweet data that is processed by the LDA model method produces 2 topic clusters of each tweet data topic, where each topic has different words that are interconnected. The purpose of the LDA Model which is run by coding python is to produce word keyword graphs in each tweet data topic. After processing with the LDA model, it enters the cluster stage from each word into a cluster of topics that represent each tweet topic. 4 tweet data topics have different clusters, by displaying a visualization of 2 topics in the cluster there are words generated from the computer using the Latent Dirichlet Allocation method. Cluster results for each topic of tweet data using the LDA method produce a visualization with pyLDAvis as shown in Figure 8.



Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
20
40
60
80
100
120
dari
motogp
asian
hasil
games
lorenzo
thailand
sirkuit
messi
pelatih
city
asia
motor
ungkap
indonesia
atas
liverpool
san
guardiola
besar
marino
ban
pemain
yamaha
hamilton
ronaldo
segera
performa
meski
inter
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 7 Topic cluster visualization

V. CONCLUSION AND FUTURE WORKS

Research on topic modeling using Twitter data and Latent Dirichlet Allocation (LDA) method as an algorithm to produce Indonesian tweet topic data modeling. In this research process, it produces word indexes, word clusters and similarity on each topic consisting of 4 different tweet data topics. In the process of modeling the index of words in a topic, LDA was tested by the Latent Semantic Indexing (LSI) method to see the performance of the LDA method in Topping indexing modeling as a comparison. LSI is a method for producing word indexing in topic modeling, then in the process of topic modeling testing is done by 2 methods, namely LDA and LSI. The results of testing the indexing process in topic modeling, LDA shows the performance in the process of word indexing in Sports topics totaling 1260 tweets with an accuracy of 98% better than the LSI method. Furthermore, the process of similarity with the LDA model produces 1 representative sentence in the topic of discussion on Sport 1260 tweets namely "Hulkenberg is considered able to Break down Hamilton's Domination in F1". The final results of this study display the visualization of wods that are considered relevant / often appear and related in the discussion of the topic, displayed two histogram visualization clusters with pyLDAviz words on the topic of Sports namely the words "MotoGP" and "Indonesian".

The LDA method used as an algorithm in topic modeling has been successfully carried out, in the process of text processing data tweets with very limited Indonesian, then it takes a new breakthrough to be able to create a large corpus / library in Indonesian to process Indonesian text data in the natural science field Language Processing (NLP).

REFERENCES

[1] Odewole, M.O., 2017. The Role of Librarian in Using Social Media Tools to Promote the Research Output of HIS/HER Clienteles. *Journal of Education and Practice*, *8*(27), pp.109-113..

[2] Becker, H., Naaman, M. and Gravano, L., 2011. Beyond Trending Topics: Real-World Event Identification on Twitter. *Icwsm*, *11*(2011), pp.438-441.

[3] Hurwitz, J.S., Nugent, A., Halper, F. and Kaufman, M., 2013. *Big data for dummies*. John Wiley & Sons.

[4] Landauer, T.K., Foltz, P.W. and Laham, D., 1998. An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), pp.259-284.

[5] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), pp.391-407.

[6] Blei, D.M., 2012. Probabilistic topic models. *Communications of the ACM*, *55*(4), pp.77-84.

[7] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

[8] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H. and Li, X., 2011, April. Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338-349). Springer, Berlin, Heidelberg.

[9] Campbell, J.C., Hindle, A. and Stroulia, E., 2014. Latent Dirichlet allocation: extracting topics from software engineering data. In *The art and science of analyzing software data* (pp. 139-159). Morgan Kaufmann.

[10] Xu, R. and Wunsch D., 2008. *Clustering* (Vol. 10). Jhon Wiley & Sons.

[11] Hammouda, K.M. and Kamel, M.S., 2003, October. Incremental document clustering using cluster similarity histograms. In *Proceedings IEEE/WIC International Confrence on Web Intelligence (WI 2003)* (pp. 597-601). IEEE

[12] Muflikhah, L. and Baharudin, B., 2009, November. Document clustering using concept space and cosine similarity measurement. In *Computer Technology and Development, 2009. ICCTD'09. International Conference on* (Vol. 1, pp. 58-62). IEEE.

[13] Gao, J. and Zhang, J., 2003, May. Sparsification strategies in latent semantic indexing. In *Proceedings of the 2003 Text Mining Workshop* (pp. 93-103).

[14] Al-Sultana, K.S. and Khan, M.M.,1996. Computational experience on four algorithms for the hard clustering problem. *Pattern recognition letters,* 17(3), pp.295-308.

[15] Chen, Q., Yao, L. and Yang, J., 2016, July. Short text classification based on LDA topic model. In *2016 International Conference on Audio, Language and Image Processing (ICALIP)* (pp. 749-753). IEEE

[16] Edi , S.N. and Ria, A., 2018. A Review on Overlapping and Non-Overlapping Community Detection Algorithms for Social Network Analytics. Far East Journal of Electronics and Communications, 18(1), pp.1-27.

[17] Edi, S.N., Djati, K., I Made, W. and Tubagus, M.K., 2017. Researchgate data analysis to measure the strength of Indonesian research. Far East Journal of Electronics and Communications, 17(5), pp.1177-1183.