Journal of Applied Engineering and Technological Science
Vol 4(1) 2022 : 390-399

TWITTER SOCIAL MEDIA CONVERSION TOPIC TRENDING ANALYSIS USING LATENT DIRICHLET ALLOCATION ALGORITHM

Musliadi KH[1], Hazriani Zainuddin[2], Yuyun Wabula[3]
Department of Computer System, STMIK Handayani Makassar, Makassar, Indonesia [123]
musliadi.esqway165@yahoo.co.id [1], hazriani@handayani.ac.id [2],
yuyunwabula@handayani.ac.id [3]

Received : 28 October 2022, Revised: 06 December 2022, Accepted : 06 December 2022
**Corresponding Author*

ABSTRACT
In Indonesia, Twitter is one of the most widely used social media platforms. Because of the diverse and frequently shifting message patterns on this social media, it is extremely challenging and time-consuming to manually identify topics from a collection of messages. Topic modeling is one method for obtaining information from social media. The model and visualization of the results of modeling topics that are discussed on social media by the Makassar community are the goals of this study. The Latent Dirichlet Allocation (LDA) algorithm is used to model and display the results of this study. The modeling results indicate that the eighth topic is the most frequently used word in a conversation. In the meantime, the 7th and 6th topics emerged as the conversation's core based on the spread of the words with the highest term frequency. The study's findings led the researchers to the conclusion that in the Makassar community's social media discussions, capitalization and visualization using the LDA method produced the words with the highest trend and the topic with the highest term frequency.

Keywords : *Topic Analysis, LDA, Trending Twitter Topics, Twitter Conversation Topics*

1. Introduction

The pace at which internet technology is developing has changed the way people live their lives. The way people communicate with one another in their day-to-day activities has changed as a result of the development of internet technology. While internet technology was initially thought to be complicated, it has since become something that the majority of people are familiar with (Yatabe et al., 2021).

Twitter is a smartphone application that has an impact on social interaction and culture. that Twitter is one of the social media platforms that can connect individuals from all over the world. The community's use of social media has a direct impact, both positive and negative. People who use social media frequently at certain times may disrupt their daily activities. For instance: Suddenly receiving a message from another person while working, which the recipient reads and responds to, can obviously disrupt their work (Ayora et al., 2021).

Twitter and other social media conversations among members of the community can provide data that can be used to examine how information changes over time. Analyses based on this information can make predictions about the Makassar community's events (Fraiwan, 2022).

Topic modeling is a clustering method which is included in unsupervised learning. No labels are used in unsupervised learning for an object. In unsupervised learning, there are three types of clusters that can be used to model data, namely hard clustering, hierarchical clustering, and soft/fuzzy clustering. To model the topic, you can use the soft/fuzzy clustering category, where each object can have more than one cluster with a certain level. Topic modeling with soft/fuzzy clustering category can use Latent Dirichlet Allocation (LDA) technique or algorithm. LDA is a method used to analyze very large documents. In addition to document analysis, LDA can also be used to summarize grouping, linking and processing data (Chauhan & Shah, 2021; Gurcan et al, 2021; Sharma & Sharma, 2022).

An observational study on the use of social media in determining the trend of discussion topics for the Makassar people from 12 to 27 September 2020 using the LDA algorithm is required based on the description of the problem's background. The study's title is research “Twitter Social Media Conversion Topic Trending Analysis Using Latent Dirichlet Allocation Algorithm”.

2. Literature Review

2.1 Grow of Internet Technology

The development of Information and Communication Technology has an impact and influence on community culture, both positive and negative impacts. The aspect of life that is most affected is in terms of cultural aspects (Irgashevich et al., 2022). From time to time, the development of communication technology continues to increase so that it affects the way humans communicate. From the results of research conducted by (Cook & Sayeski, 2022), it shows that, there is an influence of the use of Smartphone technology on the social interaction of adolescents in high school.

2.2 Social Media

Social media is digital media or the internet that has the potential as a medium for community empowerment. The presence of social media followed by the growing number of users every day provides interesting facts about how influential the internet is for life (Valkenburg et al., 2022).

Social media changes people's attitudes and behavior a lot because social media is used to create lies in society. Twitter is one of the social media that functions to find old friends which is applied by sending photos, videos, playing games, discussing, and much more (Singh et al., 2022). This social media was first founded by Mark Zuckerberg with his roommates and fellow Harvard University students, namely Eduardo, Saverin, Danrew McCollum, Dustin Moskovits and Chris Hughes (Haupt, 2021).

2.3 Data Mining

Data Mining is a scientific discipline that aims to find, explore, or mine knowledge from data or information. Data mining is an analytical step to find new knowledge from a database or knowledge discovery in a database, where knowledge can be in the form of valid data or relationships between data. Data mining can be applied in various fields that have a number of data so that data mining can be interpreted as a mixture of statistics, artificial intelligence, and database research. The application of data mining in various fields will certainly employ one or more computer learning techniques in analyzing or extracting knowledge automatically (Regin et al., 2021; Ageed et al., 2021; Oatley, 2022; Haoxiang & Smys, 2021).

Data mining has an important function to help obtain useful information in increasing knowledge for users. Basically, data mining has six functions which refer to Larose quoted, namely (Rusydiyah et al., 2021; Ewieda et al., 2021):

a. Description; aims to identify patterns that appear repeatedly in data and change these patterns into rules and criteria that are easy to understand so that they can be easily and effectively understood by the application domain so as to increase the level of knowledge in the system. This method is a data mining method that is needed by postprocessing techniques in validating and explaining the results of the data mining process. Postprocessing is a process used to ensure valid and useful results for use by interested parties.
b. Prediction; This method is used to predict what will happen in the future in a certain time based on examples of processed data.
c. Estimates; this method is similar to prediction, which distinguishes only the variable that is the target of the estimate is more in the numerical direction than in the categorical direction. The records used to perform estimates must be complete and provide the value of the target variable as the predicted value. Then, a review of the estimated value of the target variable is made based on the value of the predictive variable.
d. Classification; This method is a method used to describe and distinguish data into certain classes. This method performs the process of checking the characteristics of the object and then entering the object into one of the predefined classes.
e. Clustering; This method is a method of grouping data into the same object class without paying attention to certain data classes. Cluster is a collection of records that have similarities with each other and have dissimilarities with records in other clusters. The purpose of this process is to produce groupings of objects that are similar to each other. The greater the similarity of objects grouped in a cluster, the greater the difference between each cluster and the better the quality results from cluster analysis.

f. Association; The task of the association method in data mining is to find attributes that appear at a time. In the business world, this method is more often called a shopping basket analysis (market basket analysis).

2.4 Text Mining

Text mining is a process of exploring and analyzing large amounts of unstructured text data assisted by software in order to identify concepts, patterns, topics, keywords, and other attributes in the data. Text mining usually involves the process of structuring text input, where the input text is usually parsed together with the addition of linguistic features and deletion of words which then inserts them into the database and then derives patterns in structured data and finally evaluates and interprets the output. Text mining capabilities incorporated into AI chatbots and virtual agents are increasingly being used by companies in providing automated responses to customers as part of their marketing, sales, and customer service operations (Kumar et al., 2021; Hudaefi et al., 2021; Carracedo et al., 2021). In general, the stages carried out in text mining can be drawn as follows:

a. Tokenizing

The tokenizing stage is the stage of cutting the input string based on each word that composes it from the data source used. An example of the input string truncation stage is as follows:

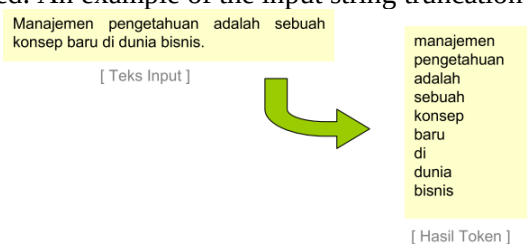

Fig. 1. Tokenization Stage

b. Filtering

The next step after tokenizing is taking important words from the token results and discarding unimportant words and storing important words. The removal of unimportant words can use the stop list algorithm or the word list algorithm. An example of the filtering stage can be seen in the following figure:

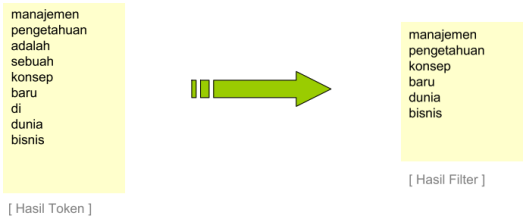

Fig. 2. Filter Stage

c. Stemming

The stemming stage is the stage carried out to find the root word of each filtered word. An example of the stemming process is more or less as follows:

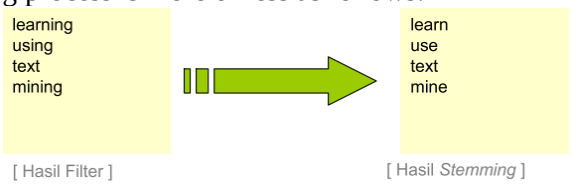

Fig. 3. Stemming Stage

d. Tagging

The tagging stage is the process of finding the initial or root form of each past word or word from stemming results. The results of the tagging process from the data taken from stemming are more or less as follows:

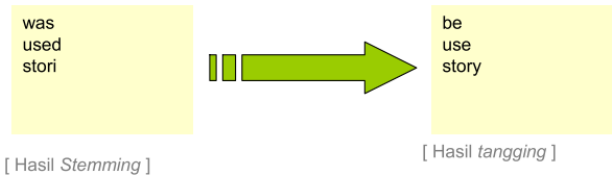

Fig. 4. Tagging Stage

e. Analyzing

The analyzing stage is the stage carried out to determine how far the connection between words between existing documents is. An example of the results of the analyzing stage is more or less like the following picture:

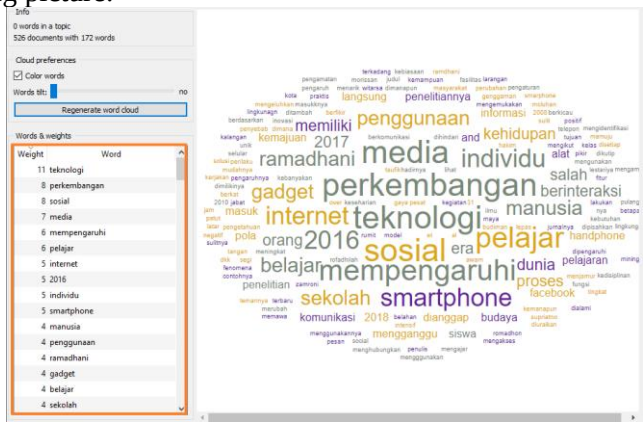

Fig. 5. Analysis Stage

2.5 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of the corpus. The basic idea of LDA is that each document is represented as a random mixture of hidden topics, where each topic is characterized by a distribution through the words in it (Gupta & Katarya, 202). Blei represents the LDA method as a probabilistic model visually (Habibi et al., 2021; Ning et al., 2022) :

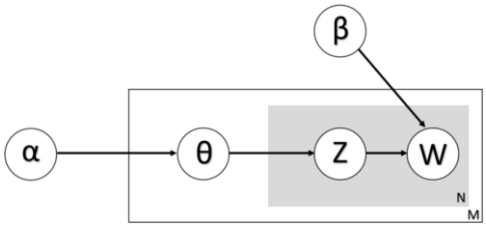

Fig. 6. Visualization of LDA according to Blei

Visualization of LDA according to Blei has three levels of LDA representation. Parameters and are corpus level parameters which are assumed to be sampled once in the process of producing a corpus. Variable is document level variable (M). Variables Zdn and Wdn are word level variables (N) and are sampled once for each word in a document. The parameter is used to determine the distribution of topics in a document, the greater the value of the greater the mix of topics in a document. On the other hand, the smaller the value of the smaller the mix of topics in a document. The parameter is used to determine the distribution of words in a topic, the greater the value of the more words there are in a topic. On the other hand, the smaller the value of the smaller the number of words in a topic. This variable represents the distribution of topics in a document, the greater the value of the more topics there are in the document. Conversely, the smaller the value of the fewer topics there are in the document.

In general, the way LDA works is by entering a document set and some specified parameters. Then the LDA process is carried out to produce a model consisting of weights that can be normalized to probability. Probability appears in 2 types: (a) the probability that a certain

document produces a certain topic in a position and (b) the probability that a certain topic produces a certain word from a collection of vocabulary.

2.6 Python

Python is a multipurpose interpretive programming language with a design philosophy that focuses on code readability. Python is an open source programming language. Python itself was launched in the community since 1991 by Guido van Rossum under the name of the Python Software Foundation vendor. Python is claimed to be a programming language that combines capabilities, capabilities with a clear code syntax and is equipped with a complete and comprehensive library.

3. Research Methods

Research methodology is a procedure or method used in conducting research along with steps that are systematically arranged to solve the problem being studied using a certain scientific basis. The research methodology framework used can be seen in Figure 7 below:

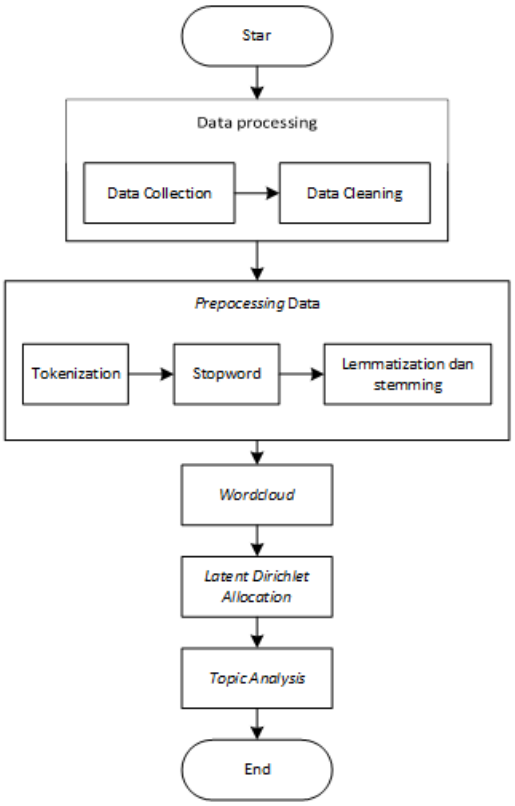

Fig. 7. Research methodology framework

4. Results and Discussions

The application of the Latent Dirichlet Allocation (LDA) algorithm to community twitter data on 12-27 September 2020 was carried out by following the research methodology framework with the following results:

a. Data collection

The results of data collection carried out on 12-27 September 2020 obtained data of 224,515 records from 29,668 users. Sample data from the results of data collection carried out can be seen in Table 1.

Table 1 – Sample Data.

Nama User	Username	Text	Date	Time	Location
Agnesia Hartono	agnesia_hartono	Keberhasilan adalah buah dari kerja keras + pantang menyerah, Bukan dari sekedar mimpi	2020-11-14	13.49.29	Makassar, Indonesia

Andi Muhammad Irham	Andirham	Risih juga di telpon terus..	2020-11-14	13.49.21	Makassar
Online Shop Makassar	TerkiniGaul	Hey\(â€™âˆ‡â€™)/ ranni24_ AyoBantu Retweet & Cekidot TokoTamz pinBB:2BB19B17 Jual Sepatu Baju Aksesoris Cewek di Makassar	2020-11-25	16.44.19	Kota Makassar
Makassar Event #Makassar Event	MakassarAcara	Hey(ã• £^â–¿^)ã• £ Pengurus_masjid AyoBantu Retweet & Cekidot TokoTamz pinBB:2BB19B17 Jual Sepatu Baju Aksesoris Cewek di Makassar	25/11/2020	16.44.18	Makassar

b. Data Cleaning

Data cleaning aims to remove parts of the data that are not used in the analysis process carried out to model the data. The focus of data analysis will only use data in the Text column which is the status of each user. Based on the results of data collection, in the data there are several columns that are not needed in the analysis process, then these columns will be cleaned.

```
# Data Cleaning
# Remove the columns
df = df.drop(columns=['No', 'Name User', 'Username', 'Date', 'Time', 'location'], axis=1).sample(100)
# Print out the first rows of papers
df.head()
```

Fig. 8. Data Cleaning

c. Data Preposing

The implementation of data preposing is carried out before analyzing the topic using LDA with the aim of structuring, tidying, and preparing the data before the data is analyzed. Data preprocessing is done sequentially, namely removing punctuation marks, removing numbers between spaces, case folding and removing stopwords. After preposing the data, punctuation marks, numbers between spaces, case folding and stopwords such as the word "yang" are removed.

```
30562         dsr bocah ka  l bis ngikutin cermah imam j...
145477                                      wu jelas hahaha
103494     sepertinya saya yang kamu mksd wkska bukan ba...
157987       you ever accidentally fuck w someone for 3 ...
61197     bercanda pakai bawa agama itu biar apa sih  ya...
Name: paper_text_processed, dtype: object
```

Fig. 9. Before Preprocessing Data

```
30562         dsr bocah  l bis ngikutin cermah imam jumb...
145477                                      wu jelas hahaha
103494     sepertinya kamu mksd wkska bukan bagaimana ma...
157987       you ever accidentally fuck w someone for years
61197          bercanda pakai bawa agama biar apa  yassalam
Name: paper_text_processed, dtype: object
```

Fig. 10. After Preprocessing Data

d. *Wordcloud*

After preposing the data, punctuation marks, numbers between spaces, case folding and stop words in the data are removed. To see the results of the data preposing process, the data is displayed in the form of a wordcloud visual representation of the most common words, so the results are as follows.

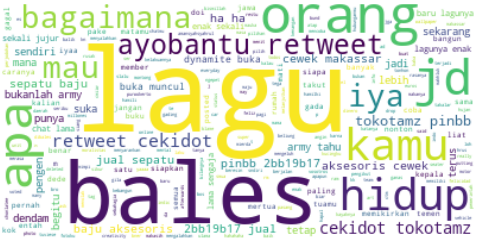

Fig. 11. Wordcloud Visual Representation

e. LDA Topic Modeling

Before doing topic modeling with LDA, it is necessary to determine the number of topic models. Determination of the number of topic models is done by looking at the coherence score. The coherence score is a measure used to evaluate topic modeling. Good modeling will produce topics with high topic coherence scores (Lestari, 2019)

Table 2 – Coherence Score

Topic	Alpha	Beta	Corehence
9	0.90	0.90	0.6821430186753112
10	0.90	0.90	0.6813112648237454
8	0.90	0.90	0.6485276773680742
9	0.90	0.90	0.6211660697995236
10	0.90	0.90	0.6185417503103652
4	asymmetric	00.01	0.6071873122344424
8	0.90	0.90	0.6031553436565154
2	00.01	00.01	0.5954692961848775
4	asymmetric	00.01	0.5910008238074393
5	asymmetric	0.90	0.5908647353392914

The coherence score generated in determining the number of topics is on topic 9 with a coherence score of 0.6821430186753112. The greater the coherence score obtained, the better the interpretation of the modeling topic will produce (Lestari, 2019). Based on the best results on the coherence score, this study used 9 (nine) topics.

```
from pprint import pprint
# number of topics
num_topics = 9
# Build LDA model
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
# Print the Keyword in the 9 topics
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

Fig. 12. Display coherence value

Table 3 - Topic Division and Its Coherence Value.

Topic	Terms
T0	0.016*"mas" + 0.016*"banget" + 0.016*"gaa" + 0.016*"abis" + 0.016*"seneng" + 0.016*"cust" + 0.016*"hasilnya" + 0.016*"adek" + 0.016*"trus" + 0.016*"ngejoki"
T1	0.033*"bales" + 0.017*"dendam" + 0.017*"chat" + 0.017*"sengaja" + 0.017*"draw" + 0.017*"tiddies" + 0.017*"adding" + 0.017*"bneran" + 0.017*"learn" + 0.017*"curut"
T2	0.017*"door" + 0.017*"kendaraan" + 0.017*"balas" + 0.017*"visor" + 0.017*"brader" + 0.017*"bupati" + 0.009*"trus" + 0.009*"orang" + 0.009*"lagunya" + 0.009*"ovo"

T3	0.024*"pagi" + 0.024*"indah" + 0.024*"bales" + 0.024*"hidup" + 0.013*"orang" + 0.013*"melepas" + 0.013*"cab" + 0.013*"eksport" + 0.013*"pangan" + 0.013*"soppeng"
T4	0.034*"berhenti" + 0.023*"pijat" + 0.023*"rakyat" + 0.012*"badan" + 0.012*"real" + 0.012*"bersahabat" + 0.012*"privasi" + 0.012*"fresh" + 0.012*"sehat" + 0.012*"terjamin"
T5	0.044*"lagu" + 0.030*"enak" + 0.016*"pas" + 0.016*"army" + 0.016*"banget" + 0.016*"jujur" + 0.016*"muncul" + 0.016*"kurus" + 0.016*"kek" + 0.016*"obatnya"
T6	0.023*"ways" + 0.020*"makassar" + 0.019*"cekidot" + 0.019*"aksesoris" + 0.019*"las" + 0.019*"ayobantu" + 0.019*"pinbb" + 0.018*"bb" + 0.018*"cewek" + 0.018*"baju"
T7	0.020*"tokotamz" + 0.020*"jual" + 0.019*"retweet" + 0.019*"sepatu" + 0.019*"baju" + 0.019*"cewek" + 0.019*"bb" + 0.019*"pinbb" + 0.019*"ayobantu" + 0.018*"aksesoris"
T8	0.043*"orang" + 0.015*"wheels" + 0.015*"album" + 0.015*"mood" + 0.015*"niggas" + 0.015*"cintailah" + 0.015*"st" + 0.015*"thursday" + 0.015*"percayai" + 0.015*"back"

Each topic generated shows the cohorence value of each word in each topic with a different value. Based on these results, we can see some examples of words with the highest cohorence/probability values, namely: “0.044_lagu” on topic T5, “0.043_orang” on topic T8, “0.034_berhenti” on topic T4, and on T1 there is the word “bales” with cohorence value 0.034.

f. Topic Modeling Visualization

The modeling visualization in the research after completing the modeling using LDA is saved in the form of pyLDAvis which can form a visualization of each topic and the most frequently occurring words.

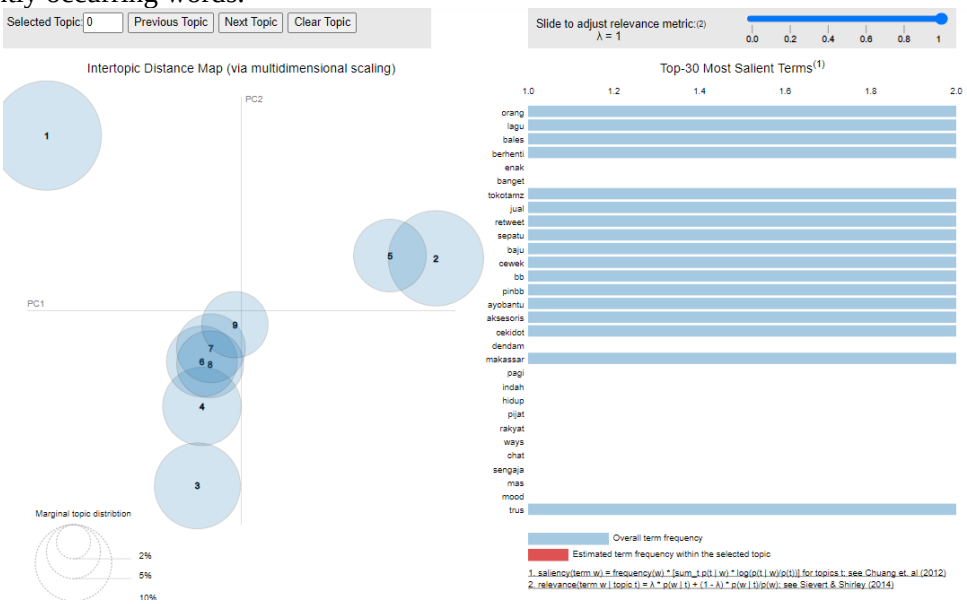

Fig. 13. pyLDAvis visualization

The pyLDAvis visualization results display 30 important words that appear in the corpus and display the dominant words discussed from 9 topics. In the right panel, the visualization displays the terms song, bales, makassar, you, sell, accessories and other words.

In addition to displaying 30 important words on all topics, the visualization results also display 30 important words from each topic. One topic with another topic may have the same word so that it overlaps with each other. For example, Topic 2 and Topic 5. The two topics overlap each other, because Topic 2 has words that are also found in Topic 5 or vice versa.

Based on the topic distribution data based on the cohorence value and the visualization results show the T8 topic to be the topic of conversation with the highest cohorence value because of the appearance of the word "People" in the visualization results.

Meanwhile, based on the results of matching the visualization with the words in the "Terms" column, the trending topics of conversation gathered on one topic, namely topics T7 and T6. In the T7 topic there are the words "tokotamz", "selling", "retweet", "shoes", "clothes", "girls", "bb", "pinbb", "ayobantu", and the word "accessories" which is a word that there are 30 important words, while on topic T6 there are the words "ways", "makassar", "cekidot", "accessories", "ayobantu", "pinbb", "bb", "girls" and the word "clothes".

5. Conclusion

Based on the results of the study, it was concluded that the capitalization and visualization with the LDA method produced the words with the highest trend and the topic with the highest term frequency in the discussion of the Makassar community on social media was on topic 8. women's accessories.

References

Ageed, Z. S., Zeebaree, S. R., Sadeeq, M. M., Kak, S. F., Yahia, H. S., Mahmood, M. R., & Ibrahim, I. M. (2021). Comprehensive survey of big data mining approaches in cloud systems. *Qubahan Academic Journal, 1*(2), 29-38.

Ayora, V., Horita, F., & Kamienski, C. (2021, January). Profiling Online Social Network Platforms: Twitter vs. Instagram. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (p. 2792).

Carracedo, P., Puertas, R., & Marti, L. (2021). Research lines on the impact of the COVID-19 pandemic on business. A text mining analysis. *Journal of Business Research, 132*, 586-593.

Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR), 54*(7), 1-35.

Ewieda, M., Shaaban, E. M., & Roushdy, M. (2021). Customer Retention: Detecting Churners in Telecoms Industry using Data Mining Techniques. *International Journal of Advanced Computer Science and Applications, 12*(3).

Fraiwan, M. (2022). Identification of markers and artificial intelligence-based classification of radical Twitter data. *Applied Computing and Informatics.*

Gupta, A., & Katarya, R. (2021). PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. *Computers in biology and medicine, 138*, 104920.

Gurcan, F., Ozyurt, O., & Cagitay, N. E. (2021). Investigation of emerging trends in the e-learning field using latent Dirichlet allocation. *International Review of Research in Open and Distributed Learning, 22*(2), 1-18.

Haoxiang, W., & Smys, S. (2021). Big data analysis and perturbation using data mining algorithm. *Journal of Soft Computing Paradigm (JSCP), 3*(01), 19-28.

Haupt, J. (2021). Facebook futures: Mark Zuckerberg's discursive construction of a better world. *New Media & Society, 23*(2), 237-257.

Hudaefi, F. A., Caraka, R. E., & Wahid, H. (2021). Zakat administration in times of COVID-19 pandemic in Indonesia: a knowledge discovery via text mining. *International Journal of Islamic and Middle Eastern Finance and Management.*

Irgashevich, S. T., Odilovich, O. A., & Mamadaliyevich, G. E. (2022). Internet Technologies In The Tourism Industry. *Web of Scientist: International Scientific Research Journal, 3*(9), 57-64.

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*(1), 100008.

Ning, W., Liu, J., & Xiong, H. (2022). Knowledge discovery using an enhanced latent Dirichlet allocation-based clustering method for solving on-site assembly problems. *Robotics and Computer-Integrated Manufacturing, 73*, 102246.

Oatley, G. C. (2022). Themes in data mining, big data, and crime analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1432.

Regin, R., Rajest, S. S., & Singh, B. (2021). Spatial data mining methods databases and statistics point of views. *Innovations in Information and Communication Technology Series*, 103-109.

Sharma, C., & Sharma, S. (2022). Latent DIRICHLET allocation (LDA) based information modelling on BLOCKCHAIN technology: a review of trends and research patterns used in integration. *Multimedia Tools and Applications*, 1-27.

Valkenburg, P. M., Meier, A., & Beyens, I. (2022). Social media use and its impact on adolescent mental health: An umbrella review of the evidence. *Current opinion in psychology*, *44*, 58-68.

Yatabe, J., Yatabe, M. S., & Ichihara, A. (2021). The current state and future of internet technology-based hypertension management in Japan. *Hypertension Research*, *44*(3), 276-285.