Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

Topic Modelling Using Non-Negative Matrix Factorization (NMF) for Telkom University Entry Selection from Instagram Comments

Alfajri*, Donny Richasdy, Muhammad Arif Bijaksana

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1,*]alfaajri@student.telkomuniversity.ac.id, [2] donnir@telkomuniversity.ac.id, arifbijaksana@telkomuniversity.ac.id
Submitted: **27/08/2022**; Accepted: **30/08/2022**; Published: **30/08/2022**

Abstract−The development of information technology is increasingly rapid, such as social media, which has much influence. Social media is a place or media used to express and express various opinions on a topic. One example is Instagram. Instagram is a social media platform with many features, such as posting photos, videos, comments, likes, and others. The comments feature that Instagram has contained much public opinion that can be used as data. Nothing but the post on the SMB Telkom University Instagram account about the entrance to the university. In posts about the entrance to Telkom university, many Instagram users comment on the post. This can be convenient for the marketing team to get topics or discussions that most followers need from Telkom University's Instagram account. Therefore, a topic modelling of Instagram users' perceptions of comments posted on the entrance to Telkom university was carried out using the Nonnegative Matrix Factorization (NMF) method. After doing several research scenarios, the best coherent value was obtained with a coherent value of 0.60628 and the best 4 topics.

Keyword: Topic Modelling; Non-negative Factorization Matrix; Coherent Topic; Instagram; Telkom University

1. INTRODUCTION

Social media has become the trend that most Internet users enjoy sharing information with the wider community. They can share using various kinds of content such as images, videos, or articles on social media. They also express their opinions or feelings through social media, such as Twitter, Facebook, Instagram, etc. One of the social media that is rapidly increasing in users is Instagram. On its website, Instagram states the number of active users on its site reaches 300 million users per month, with an average of 70 million photos uploaded every day and 2.5 billion comments every day[1]. It is none other than the SMB Telkom University Instagram account, where there are many comments given on each post. In this way, the Instagram comments can be used by the Telkom University marketing team to make it easier to find out what topics are contained in the SMB Telkom University Instagram comments, so that make it easier for the marketing team to make comparisons with Telkom University to many people, one of which is prospective new students. Therefore, a topic modelling is designed that automatically models the topic.

Topic modelling is a clustering method that is included in unsupervised learning. In unsupervised learning, there is no labelling for an object. There are three types of clustering, namely hard clustering, hierarchical clustering, and soft/fuzzy clustering. Topic modelling is included in soft/fuzzy clustering, where each object can have more than one cluster with a certain level [2].

To solve problems in topic modeling automatically, a topic modeling program will be made using the Non-negative Matrix Factorization method. There are several methods that can be used for topic modeling, such as LSA, NMF, and LDA. NMF is the set of algorithm in analyzing multivariate and linear algebra where An M matrix can generally be broken down into factors by providing two matrices W and H, with the properties that there are no negative elements in all three matrices[3]. Non-Negative Matrix Factorization (NMF) is very effective in finding the underlying topic in the corpora text. NMF is an unsupervised approach to reducing the dimensions of non-negative matrices [4]. The purpose of the NMF method to introduce text mining is done by using a partitional clustering technique that identifies the semantic features in a collection of documents and groups the documents into clusters based on their semantic features. NMF is used to organize a collection of texts into a structured form or grouped directly based on non-negative factorization.

There are several previous studies related to topic modelling with the Non-negative Matrix Factorization method, namely research (Guven, 2019) evaluation of the NMF and LDA methods using the Turkish Twitter dataset, which shows the NMF accuracy level is higher than LDA, where NMF produces an accuracy rate of 97.8 %, while LDA produces an accuracy rate of 96.0% [5]. Research (Nakhon Pathom, 2014) conducted topic modelling on a collection of comments contained on yahoo news social media using the Latent Dirichlet Allocation and Non-negative Matrix Factorization methods, which resulted in extracting three topics, with ten keywords for each topic and limited to a maximum keyword of 10,000 in comments body [6]. Research (Mifrah, 2020) conducted a comparison of topic modelling between the LDA and NMF methods using the corpus covid'19. From this study, it can be concluded that the LDA model is more relevant than the NMF model in the case of a corpus of 13,000 citations with documents or long texts in a coherent topic score [7]. As with research (Chirag, 2020) conducted topic modelling with NMF, whose dataset is a collection of Twitter tweets, and this research results that NMF not only produces an accurate topic model but also produces a much more stable output [8]. Research (O'Callaghan, 2015) comparing the value of the coherence of the NMF and LDA methods which in that research

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

used data from news tweets on the @BBCNews Twitter account, which amounted to 91,616 tweets, and proved that the NMF method produces a higher level of coherence and reduction which also exceeds LDA method [9].

This study focuses on modelling the Indonesian language Instagram commentary text contained on the SMB Telkom University Instagram account, on applying topic modelling with Non-negative Matrix Factorization to find out what topics are contained in the comments on the SMB Telkom University Instagram account, and getting an NMF evaluation based on an optimal number of topics by using coherence score measurement. The dataset used is the result of crawling from the SMB Telkom University Instagram account.

2. RESEARCH METHODOLOGY

2.1 Research Stages

In this study, a system design was built to make it easier to solve topic modelling problems on SBM Telkom University Instagram comments. There are several stages in this study, namely the first stage of data collection for SMB Telkom University Instagram comments, the second stage of pre-processing the data, the third stage of feature extraction, the fourth stage of topic modelling with non-negative matrix factorization, and the last stage of evaluation with a coherence score. The following is an overview of the system design, which can be seen in Figure 1:

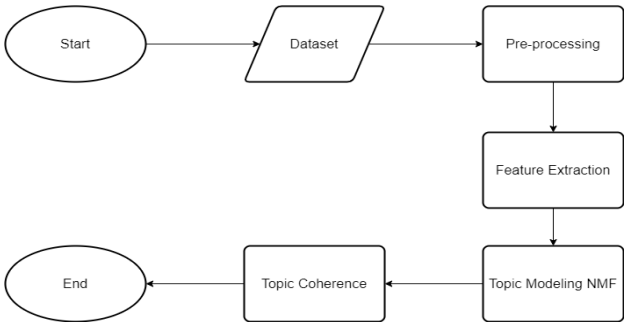

Figure 1 System Overvie

2.2 Dataset

This study used a dataset from the results of crawling comments on the SMB Telkom University Instagram account, which amounted to 1008 comment data collected in CSV format. The data crawling process is carried out using the APIFY web application, which is an application for automatic data scraping.

2.2 Pre-processing

Pre-processing is the initial stage of conducting topic modelling using Non-negative matrix factorization. The purpose of pre-processing is to standardize Instagram comment text data wherein the data there are various styles of writing or comments. Cleaning Instagram data at this pre-processing stage will include case-folding, tokenizing, filtering, stemming [10].

2.2.1. Case Folding

At the case-folding stage, changes are made to the text from uppercase to lowercase. In the search for text on Instagram comment data, you will find writing in lowercase and some in uppercase. The purpose of case-folding is to equate all words, to make it easier for the data modelling process later.

2.2.2. Tokenizing

Tokenizing is the second stage after carrying out the case-folding process. This stage is carried out to separate words in sentences into single-word pieces contained in Instagram comment text data.

2.2.3. Filtering

The next pre-processing stage is filtering, which is used to retrieve important words from the tokenizing results. Common words that usually appear and have no meaning are called stop words. For example, the use of conjunctions such as and, which, as well as, after, and others. Removing this stopword can reduce index size and processing time[11]

2.2.4. Stemming

The stemming process is carried out to filter the words used to produce basic terms. In the Instagram comment text document, there are many affixes/prefixes/suffixes [12]. Stemming is also a process that is included in the text

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

normalization stage so that the text is more abundant and consistent. The following is the result of pre-processing data, which can be seen in table 1.

Table 1 Data Prep**rocessing Results**

Process	Before	After
Case Folding	Masih menerima pendaftaran utk jurusan ini?(Still accepting applications for this course?)	masih menerima pendaftaran utk jurusan ini(still accepting applications for this course)
Tokenizing	masih menerima pendaftaran utk jurusan ini(still accepting applications for this course?)	[‘masih’(still),‘menerima’(accept),‘pendaftaran’(registration),’utk’(utk), ‘jurusan’(major), ‘ini’(ini)]
Filtering	masih menerima pendaftaran utk jurusan ini(still accepting applications for this course?)	masih menerima pendaftaran jurusan
Stemming	masih menerima pendaftaran utk jurusan ini(still accepting applications for this course?)	Masih(still), terima(accpet), daftar(list), utk(utk), jurus(stance), ini(this)

2.3 Term Frequency-Inverse Document Frequency (TF-IDF)

After going through the data processing process, it then enters the feature extraction process, and this process is carried out so that the dataset can be processed on topic modelling. The feature extraction process is carried out by calculating the term frequency using TF-IDF and producing documents in the form of term frequency against a collection of documents or corpus [13]. The results of the feature extraction are then processed in topic modelling using NMF. In the calculation of TF-IDF using the formula contained in the following equation:

a. Calculate the value of TF by using equation (1) as follows:

$$TF(i,j) = \frac{freq(i,j)}{max_{ki,kj}} \quad (1)$$

b. Calculate the IDF value using equation (2) as follows:

$$IDF(i) = \log_2\left(\frac{N+1}{n_1+1}\right) + 1 \quad (2)$$

c. Calculate the value of TF-IDF by using equation (3) as follows:

$$TF - IDF(i,j) = TF_{(i,j)} * IDF_{(i,j)} \quad (3)$$

2.4 Topic Modelling Non-negative Matrix Factorization

Non-negative Matrix Factorization is a method proposed by Lee and Seung as a method of decomposition of a data matrix. Non-negative Matrix Factorization is a word representation matrix with topics that have non-negative values so that this matrix is much easier to interpret [14]. NMF is a factorization method for the V matrix with size $m \times n$ into a $Wm \times k$ matrix and a non-negative value $Hk \times n$ matrix. NMF has a decomposition which can generally be expressed in terms of the equation proposed[15], where equation 1 is as follows:

$$V = W * H \quad (4)$$

Matrix V is a document word matrix that represents the document's text where each entry of the column vector describes the number of words in a document. The W matrix is a weight matrix in which each row vector represents the vector of each word on the topic. Matrix H is a feature matrix in which each column vector represents the corpus of each document on the topic. The following is an illustration of the NMF method in Figure 2.

$$\begin{array}{c} \\ k_1 \\ k_2 \\ \cdots \\ k_m \end{array}\!\!\begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ e_{m1} & e_{m2} & \cdots & e_{mn} \end{bmatrix} \end{array} \approx \begin{array}{c} \\ k_1 \\ k_2 \\ \cdots \\ k_m \end{array}\!\!\begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_k \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mk} \end{bmatrix} \end{array} \times \begin{array}{c} \\ t_1 \\ t_2 \\ \cdots \\ t_k \end{array}\!\!\begin{array}{c} \begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \\ \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{k1} & y_{k2} & \cdots & y_{kn} \end{bmatrix} \end{array}$$

Figure 2 Illustration of Non-negative Matrix Factorization method

2.2 Evaluasi Topic Coherence

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

A coherence score is a value that is used to measure the level of coherence of the topic. Topic Coherence calculates the score of a single topic by measuring the level of semantic similarity between words with high scores on topic modelling. This score measurement can help distinguish topics that can be interpreted semantically and topics that are the result of inference [16]. The score from the coherence calculation is a measure used to evaluate a topic model. Namely, the higher the coherence value or score, the more perfect the resulting model[17]. There are several coherence measurement techniques. In this study, using the measurement technique c_v, c_umass. The following is the formula for calculating the coherence value technique:

a. The calculation formula with the c_v technique is found in equation (5) as follows:

$$NPMI(w_i w_j) = \sum_{j}^{N-1} \frac{\log \frac{p(wi,wj)}{p(wi),p(wj)}}{-\log p(wi,wj)} \quad (5)$$

Where $P(wi)$ is the probability of random occurrence of wi in the document, $P(wi, wj)$ is the probability of two words wi and wj appearing in the document at random. N is the highest choice of the words $w1, w2, \dots, wn$ [18].

b. The calculation formula with the c_umass technique is found in equation (6) as follows:

$$C_{U\ Mass}(w_i w_j) = \log \frac{D_{(Wi,Wj)}+1}{D_{Wi}} \quad (6)$$

Where $D(wi, wj)$ indicates the number of times the words wi and wj appear together in the document, and $D(wi)$ is the number of times the word wi occurs alone. The bigger the number, the better the coherence score [18].

3. RESULTS AND DISCUSSION

In this study, several test scenarios have been carried out to obtain and find out the best results from topic modelling using the Non-negative Matrix Factorization method on the SMB Telkom University Instagram account comment dataset. The test is carried out using two test scenarios, namely, using stemming and stop removal, not using stemming and stop removal. This scenario was conducted to determine whether it affects the coherence score when using stemming and stop removal and not using stemming and stop removal, and the second test scenario centered on the technique of calculating the coherence value using the c_v technique and the c_umass technique. The purpose of this scenario is to determine whether the topics generated by the NMF method are optimal or not by looking at the topics that emerge with the best coherent value.

3.1 Results and Discussion using stemming and stop removal with the c_v . technique

In the first test, using stemming and stop removal with the c_v calculation technique. The use of stemming and stop removal serves to see the difference in the results of coherent values. The results of this test produce a coherent score with a value of 0.51582 with a maximum number of topics 4. The coherent value and the number of topics produced can be seen in table 2:

Table 2 Coherent Value and Number of First Test Topics

Pre-processing	Num of Topic	Coherence score (c_v)
Stemming and Stop Removal	2	0.30022
	3	0.32011
	4	**0.51582**
	5	0.39618
	6	0.43553
	7	0.50722
	8	0.47306
	9	0.40301
	10	0.45108

3.1.1 Topic Results from the first test

Table 3 is the result of topic modelling using stemming and stop removal with the calculation of coherent values using the c_v technique, resulting in a maximum of 4 topics with 10 keywords displayed to build the main topic.

Table 3 First Test Topic Results

No Topic	Keyword Topic	Topic
Topic 0	kapan(when), umum(general), utbk(utbk), utg(utg), jalur(track), berapa(how many), jam(o'clock), hasil(result), beasiswa(scholarship), tanggal(date)	Time for the opening of the Telkom University entrance selection path

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

No Topic	Keyword Topic	Topic
Topic 1	Jalur(track), buka(open), utbk(utbk), beasiswa(scholarship), telkom(telkom), rapor(report card), vokasi(vocation), masuk(enter), jpa(jpa), tahun(year)	Time for the opening of the Telkom University entrance selection path
Topic 2	Informasi(information), sistem(system), teknik(tecnique), teknologi(technology), grup(group), info(info), maba(maba), telekomunikasi(telecommunication), industri(industry), logististik(logistics)	Information on Departments at Telkom University
Topic 3	Daftar(Regis), errror(error), terus(Keep going), web(web), beasiswa(scholarship), eror(eror), pjj(pjj), akhir(end), reguler(regular), kapan(when)	Error Telkom University registration website

3.2 Results and Discussion do not use stemming and stop removal with the c_v . technique

The second result is not using stemming and stop removal. In this test, it is very influential on the coherent score, because the coherent value in this test is higher than the first test, which is worth 0.60628 with the maximum topic generated is 4. The results of the test not using stemming and step removal with the c_v technique can be seen in table 4:

Table 4 Coherent Value and Number of Second Test Topics

Pre-processing	Num of Topic	Coherence score (c_v)
No Stemming and no Stop Removal	2	0.50676
	3	0.55401
	4	**0.60628**
	5	0.54069
	6	0.59101
	7	0.53730
	8	0.51176
	9	0.60060
	10	0.56056

3.2.1 Topic Results from the second test

Table 5 is the result of topic modelling resulting from testing not using stemming and stop removal and using the calculation of coherent values with the c_v technique. So as to produce the same topic as the first test topic, the name has a different value in its coherent value.

Table 5 Second Test Topic Modelling Results

No Topic	Keyword Topic	Topic
Topic 0	kapan(when), umum(general), utbk(utbk), utg(utg), jalur(track), berapa(how many), jam(o'clock), hasil(result), beasiswa(scholarship), tanggal(date)	Time for the opening of the Telkom University entrance selection path
Topic 1	Jalur(track), buka(open), utbk(utbk), beasiswa(scholarship), telkom(telkom), rapor(report card), vokasi(vocation), masuk(enter), jpa(jpa), tahun(year)	Time for the opening of the Telkom University entrance selection path
Topic 2	Informasi(information), sistem(system), teknik(tecnique), teknologi(technology), grup(group), info(info), maba(maba), telekomunikasi(telecommunication), industri(industry), logististik(logistics)	Information on Departments at Telkom University
Topic 3	Daftar(Regis), errror(error), terus(Keep going), web(web), beasiswa(scholarship), eror(eror), pjj(pjj), akhir(end), reguler(regular), kapan(when)	Error Telkom University registration website

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

3.3 Results and Discussion using stemming and stop removal with the c_Umass technique

The next test uses stemming and stop removal by calculating the coherent value using the c_umass technique. The use of controlling and stop removal aims to check whether it affects the results of a coherent score. And the use of the c_umass technique to compare the results of topic modelling with the c_v technique. In this test, the highest coherent value was obtained with a value of -16.8606, and the best topics produced were 6 topics. The results of the coherent value and the number of topics generated can be seen in table 6 :

Table 6 Coherent Value and Number of Third Testing Topics

Pre-processing	Num of Topic	Coherence score (c_v)
Stemming and Stop Removal	2	-12.1969
	3	-15.2623
	4	-16.6023
	5	-14.5510
	6	**-16.8606**
	7	-15.4977
	8	-14.5732
	9	-17.3776
	10	-16.4325

3.3.1 Modelling Results Third test topic

Table 7 is the result of topic modelling using stemming and stop removal with the technique of calculating the coherent value of c_umass so that it produces a maximum topic of 6 with 10 keywords that are displayed to build the main topic.

Table 7 Third Test Topic Modelling Results

No Topic	Keyword Topic	Topic
Topic 0	kapan(when), umum(general), utbk(utbk), berapa(how many), jalur(track), hasil(result), jam(o'clock), tanggal(date), nilai(score), kpn(kpn)	Time for the opening of the Telkom University entrance selection path
Topic 1	jalur(track), buka(open), utbk(utbk), rapot(rapot), vokasi(vacation), telkom(telkom), masuk(enter), reguler(regular), tahun(year), tanya(asking)	Time for the opening of the Telkom University entrance selection path
Topic 2	informasi(information), sistem(system), teknologi(technology), akuntansi(accountancy), mbti(mbti), seleksi(selection), ekstensi(extension), grupp(grupp), pjj(pjj), prodi(study program)	Information on Departments at Telkom University
Topic 3	daftar(list), pjj(pjj), kapan(when), terus(keep going), buka(open), reguler(regular), eror(eror), web(web), akhir(end), error(error)	Error Telkom University registration website
Topic 4	teknik(technique), info(info), grup(group), maba(maba), telekomunikasi(telecommunication), industry(industry), logistic(logistics), biomedis(biomedical), informatika(informatics), angkat(lift)	Information on Departments at Telkom University
Topic 5	utg(utg), jpa(jpa), please(please), minn(minn), bakal(will), minnn(minnn), sama(same), usm(usm), buka(open), sih(sih)	Questions Regarding the Entrance Path to Telkom University

3.4 Results and Discussion do not use stemming and stop removal with the c_Umass technique

The next test is not using stemming and stop removal by calculating the coherent value using the c_umass technique. The use of stemming and stop removal functions to see changes in the coherent values and the use of the c_umass technique is to compare the best topics generated. This test produces a coherent value of -13.9704, and the best topic is 7. The results of the test can be seen in table 8:

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

Table 8 Coherent Value and Fourth Topic Test

Pre-processing	Num of Topic	Coherence score (c_v)
No Stemming and no Stop Removal	2	-6.1710
	3	-6.6750
	4	-13.5832
	5	-9.8362
	6	-11.6354
	7	**-13.9704**
	8	-12.8745
	9	-10.0600
	10	-12.0818

3.4.1 Modelling Results Fourth test topic

Table 7 is the result of topic modelling not using stemming and stop removal with the c_umass coherent value calculation technique so that it produces a maximum topic of 6 with 10 keywords that are displayed to build the main topic.

Table 9 Fourth Test Topic Modelling Results

No Topic	Keyword Topic	Topic
Topic 0	Kapan(when), umum(general), utbk(utbk), berapa(how many), jalur(track), hasil(result), jam(o'clock), tanggal(date), nilai(score), kpn(kpn)	Time for the opening of the Telkom University entrance selection path
Topic 1	Jalur(track), buka(open), utbk(utbk), rapot(rapot), vokasi(vacation), telkom(telkom), masuk(enter), reguler(regular), tahun(year), tanya(asking)	Time for the opening of the Telkom University entrance selection path
Topic 2	Informasi(information), sistem(system), teknologi(technology), akuntansi(accountancy), mbti(mbti), seleksi(selection), ekstensi(extension), grupp(grupp), pjj(pjj), prodi(study program)	Information on Departments at Telkom University
Topic 3	Daftar(list), pjj(pjj), kapan(when), terus(Keep going), buka(open), reguler(regular), eror(eror), web(web), akhir(end), error(error)	Error Telkom University registration website
Topic 4	Teknik(technique), info(info), grup(group), maba(maba), telekomunikasi(telecommunication), industry(industry), logistic(logistics), biomedis(biomedical), informatika(informatics), angkat(lift)	Information on Departments at Telkom University
Topic 5	Utg(utg), jpa(jpa), please(please), minn(minn), bakal(will), minnn(minnn), sama(same), usm(usm), buka(open), sih(sih)	Questions Regarding the Entrance Path to Telkom University

4. CONCLUSION

Based on the research that has been done. Has successfully crawled the comment data for the SMB Telkom University Instagram account, which amounted to 1008 comment data. The data will be used as a dataset to conduct topic modelling using Non-negative Matric Factorization with the help of c_v and c_umass techniques for the evaluation of topic modelling. From the results of the evaluation of the coherent value with the c_v technique, the value is 0.60628 without using stemming and stop removal. And the value is 0.51582 by using stemming and stop removal and produces the best 4 topics. The results of the evaluation of the coherent value using the c_umass technique were -13.9704 with no stemming and stopped removal, and a value of -16.8606 by using stemming and stop removal, and the best topics were 6 and 7, but in the c_umass technique, the topics were ambiguous because the keywords generated could not be found. Used to build the main topic. The reason why stemming and stop removal can affect the coherence score is because stemming and stop removal change the original text a lot in the data set. From these results, topic modelling with Non-negative Matrix Factorization using an evaluation with a coherent value of the c_v technique without stemming and removal produces the best results. And the 4 best topics

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 485−492
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2212

that were produced were topics 1 and topic 2 regarding the opening time of the Telkom University entrance selection path, topic 3 regarding the information on majors at Telkom University, and topic 4 regarding the wrong Telkom University registration website.

REFERENCES

[1] S. Scholz and C. Winkler, "How to Engage Followers: Classifying Fashion Brands According to Their Instagram Profiles, Posts and Comments," Dec. 2020, pp. 29–50. doi: 10.5121/csit.2020.101704.

[2] C. B. Asmussen and C. Møller, "Smart literature review: a practical topic modelling approach to exploratory literature review," *J Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0255-7.

[3] M. Belford, B. mac Namee, and D. Greene, "Stability of topic modeling via matrix factorization," *Expert Syst Appl*, vol. 91, pp. 159–169, Jan. 2018, doi: 10.1016/j.eswa.2017.08.047.

[4] E. OKKALI, H. ATAMTÜRK, and Z. H. KİLİMCİ, "Evaluation of Society Response to Violence against Women in Turkey via Twitter using Topic Modeling," *Kocaeli Journal of Science and Engineering*, Nov. 2021, doi: 10.34088/kojose.907333.

[5] Z. A. Guven, B. Diri, and T. Cakaloglu, "Comparison Method for Emotion Detection of Twitter Users," Oct. 2019. doi: 10.1109/ASYU48272.2019.8946435.

[6] ICT International Student Project Conference 3. 2014 Nakhon Pathom, J. L. Mitrpanont, ICT International Student Project Conference 3 2014.03.26-27 Nakhon Pathom, ICT-ISPC 3 2014.03.26-27 Nakhon Pathom, and ISPC 3 2014.03.26-27 Nakhon Pathom, *Proceedings of the 2014 Third ICT International Senior Project Conference (ICT-ISPC2014) March 26-27, 2014, Faculty of ICT, Mahidol University, Nakhon Pathom, Thailand*. IEEE, 2014.

[7] S. Mifrah, "Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5756–5761, Aug. 2020, doi: 10.30534/ijatcse/2020/231942020.

[8] M. Chirag and K. Pathela, "Exploring the space of Topic Modelling and Topic Coherence on short and long text corpora," 2020.

[9] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Syst Appl*, vol. 42, no. 13, pp. 5645–5657, Aug. 2015, doi: 10.1016/j.eswa.2015.02.055.

[10] R. Churchill and L. Singh, "textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data." [Online]. Available: https://github.com/GU-DataLab/topic-modeling-

[11] A. Oussous, A. A. Lahcen, and S. Belfkih, "Impact of text pre-processing and ensemble learning on Arabic sentiment analysis," in *ACM International Conference Proceeding Series*, 2019, vol. Part F148154. doi: 10.1145/3320326.3320399.

[12] R. Churchill and L. Singh, "textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data," 2021. [Online]. Available: https://github.com/GU-DataLab/topic-modeling-

[13] E. Zamiraylova and O. Mitrofanova, "Dynamic Topic Modeling of Russian Prose of the First Third of the XXth Century by Means of Non-Negative Matrix Factorization *."

[14] R. Vangara *et al.*, "Finding the Number of Latent Topics with Semantic Non-negative Matrix Factorization," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3106879.

[15] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications," Mar. 2018, doi: 10.1109/MSP.2018.2877582.

[16] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, Y. Sahria, and D. Hatta Fudholi, "Terakreditasi SINTA Peringkat 2 Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation)," *masa berlaku mulai*, vol. 1, no. 3, pp. 336–344, 2017.

[17] Zoya, S. Latif, F. Shafait, and R. Latif, "Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling," *IEEE Access*, vol. 9, pp. 127531–127547, 2021, doi: 10.1109/ACCESS.2021.3112620.

[18] S. Bellaouar, M. M. Bellaouar, and I. E. Ghada, "Topic modeling: Comparison of LSA and LDA on scientific publications," in *ACM International Conference Proceeding Series*, Feb. 2021, pp. 59–64. doi: 10.1145/3456146.3456156.