Combined Text Mining: Fuzzy Clustering for Opinion Mining on the Traditional Culture Arts Work

Elta Sonalitha[1], Anis Zubair[2], Priyo Dari Molyo[3], Salnan Ratih Asriningtias[4]
Bambang Nurdewanto[5], Bondhan Rio Prambanan[6], Irfan Mujahidin[7]
Electrical Engineering, Merdeka University Malang[1, 7]
Information Systems Faculty of Information Technology, Merdeka University Malang[2, 5]
Information Technology, Brawijaya University[4]
Communication Engineering, Merdeka University Malang[5]
Master of management, Merdeka University Malang, Malang, Indonesia[6]

***Abstract*—The Indonesian government is currently intensifying work programs in the field of traditional arts and culture. In order to realize the promotion of the country's culture, the government has enacted a law on cultural promotion. One indicator of the achievement of the promotion of culture, among others, with the collection of data on traditional culture, the data mapping and data inventory can be processed into information and knowledge. In this research, indicators of performance indicators were compiled from connoisseurs of traditional works of art using data in the city of Malang, East Java, Indonesia. The results of the audience's opinion on cultural offerings can be used as a benchmark for the success of the promotion of traditional culture. When the culture is explored and tried to be displayed again, it is important to know the audience's satisfaction and understanding of the display that has just been witnessed. The results of the description of respondents in the form of opinions on the artwork will be collected as data processed using Text Mining with the Clustering of Fuzzy C-Means method to determine the audience's opinion about Feeling , which is related to feelings when viewing the beauty of the artwork, Value is related to the assessment to an art work that can be in the form of art weights contained in the work of art, and Emphasizing , which is related to empathy or respect for the art world, including professions related to the world such as dancers, musicians and others. The results achieved from this study show that has good performance on the proposed method. This can be known from the results of data testing using cluster variance V = 0.00000217901. Based on these values it can be concluded that the value of all cluster variants is good.**

***Keywords*—Text mining; opinion mining; fuzzy clustering; arts work**

I. Introduction

Cultural shifts began in the millennia era. Advances in technology led to the spread of a very diverse modern spectacle had shifted the traditions and traditional culture of the Indonesian nation. Generation who was born in the millennia era, is less familiar with the traditions and cultural heritage of the ancestors. This is very worrying and threatens the extinction of the native culture of the country [1][2]. The right step taken by the government is to pass a law on cultural promotion.

Traditional culture is the ultimate wealth of the Republic of Indonesia Unitary State, if it is not preserved properly, Indonesia's wealth will eventually disappear. For this reason, the Indonesian government promulgates Law No. 5 2017 concerning the promotion of culture. Namely the efforts that will be made to explore, record, reorganize the traditional works of the archipelago. In addition, this is reinforced by the UUD 194;32 (1); which reads culture is a future investment in the nation's civilization. There are 10 cultural fields that are defined in the presentation of the laws on cultural advancement, including: Oral Traditions, Traditional Knowledge, Manuscripts, Language, Customs, Sites, Traditional Sports, Art, Folk Games and Traditional Technology. This research raises the field of art culture.

In the world of art, there are three components that support the arts, among other artists, his appreciation for the connoisseur or awards, namely, Spectator, and Art itself as a product. This research focuses on “Audience” as a measure of the success of a show [3][4]. The opinion of the audience or viewers of a performance of art is very important for the evaluation material of the show itself. Good and bad impressions, understanding or not, enjoying or rejecting the show, whether or not satisfied will be the subject of study and input to be followed up on the next display [5].

The importance of public opinion and appreciation must receive more attention. Data collection and mapping need to be done to overcome this problem. Indicators of achieving cultural progress include the collection of data on traditional culture, data mapping, and data inventory [6][7]. The indicator of connoisseurs of art is a measure of the success of cultural promotion. When the culture is explored and tried to be displayed again, it is important to know the audience's satisfaction and understanding of the display that has just been witnessed [8][9].

Especially in the increasingly rare Traditional Art Works, requiring data and information from the public, how much attention is paid to the traditional art. The more the audience understands and understands the storyline, meaning, philosophy contained in a work of art, then the work of art can be said to be good. This understanding is a positive value from the achievement of the objectives of the cultural promotion effort. In an effort to realize the law on cultural promotion, indicators of the success of a cultural dish are mainly needed from the viewers or the public.

Based on these problems, we need a decision support system that is able to provide an overview and conclusions that can be used to benchmark policy decisions [10][11]. This research focuses on measuring the level of satisfaction of the audience of a show by using an assessment instrument that will be distributed when the audience has finished watching the show.

This study proposes the combined method of Text Mining Fuzzy Clustering as a method. Fuzzy was chosen in this study because fuzzy has the advantage of being easier to implement in various problems. Fuzzy is also often used in various kinds of problems related to forecasting, control, and clustering. F Fuzzy used in this study is Clustering of Fuzzy C-Means. Several previous studies have successfully used C-Means as a method in clustering, including Collazo-Cuevas. Text mining is used as a method of feature extraction that is filled in by the audience or the community of art lovers. While Fuzzy Clustering as a machine of machine learning to produce information [12][13]. The information generated from this decision support system is in the form of an audience's level of understanding and satisfaction with the work of art. This level of understanding and satisfaction becomes the basis of evaluation for the process of promoting culture, especially art. The achievement of the success of the appearance of a work can be measured accurately. Fuzzy Clustering can measure the level of audience satisfaction with an art performance so that it is very important to more massive promote the Indonesian culture.

The structure of this paper consists of an introduction which contains an introduction to traditional art in Indonesia as a medium to achieve the original culture and orientation of combined text mining - fuzzy clustering for this research, then the implementation process of Text Mining which is maximized by combining fuzzy c-Means to improve the results of measuring accuracy. The level of audience satisfaction with an art performance, after that result and analysis, contains an analysis of optimization analysis and improvement results from the implementation of combined text mining-fuzzy c-Means.

II. COMBAINED TEXT MINING-FUZZY C-MEANS

This study uses a combination of the Text Mining and Fuzzy C-Means methods as a method. Text mining is used as a method of feature extraction from the results of questionnaires or instruments that have been filled in by the audience after watching a traditional art performance. In Fig. 1, clustering data with Fuzzy C-Means or information that has gone through a weighting process [14][15].

The stages in this research starting from extracting information to producing information are shown in the figure below.

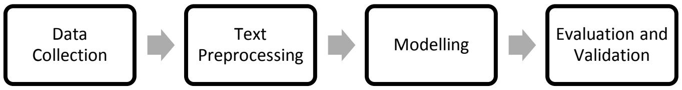

Fig. 1. Flow Diagram of the Text Mining - Fuzzy C-Means Combined Method.

A. The Data Collection

Data used in the study were 139 opinions of data from viewers of art performances in the city of Malang, East Java, Indonesia. Data in the form of text or paragraph data obtained from the questionnaire form. The data collection process is carried out by giving questionnaires or research instruments to the audience who have just seen a web-based work of art. After the data is collected, the process will then be performed Text Preprocessing [16]. The data that has been collected is shown in Table I.

B. Text Preprocessing

Text mining is a knowledge intensive process where users interact and work with a group of documents using several analysis tools. Text mining aims to get useful information from data sources in the form of documents consisting of unstructured text through identification, and exploration of a pattern. The stage in text mining is text preprocessing, namely changing unstructured data into structured data. Steps in the preprocessing text include Bag of Words and Term Weighting [17].

Bag of words consists of several processes, including folding cases to eliminate characters other than letters and change all letters to lowercase letters. Next Tokenizing which aims to break the sentence into separate words known as term or token. Furthermore filtering by removing punctuation, changing capital letters into lowercase letters and eliminating stop word that aims to delete words that are not useful or have no influence in the process [18]. Finally, stemming is to get the basic words from words that have received affixes or other information.

TABLE I. COLLECTION OF DATA ON AUDIENCE OPINIONS PERFORMING ART

Audience ID	Opinion
1	have seen traditional art performances often watch traditional art performances ... Very important to improve or develop
2	have seen traditional art performances often watch traditional art performances in a year Awareness of the importance of traditional arts and local culture to continue to be supported and cultivated Also maintain and support the existence of traditional art ... It is important to improve or develop
3	ever see traditional art performances sometimes watch traditional art performances in a year Because love traditional art entertainment As a form of self-appreciation More have the awareness to continue to participate in preserving and preserving traditional art ... Marketing Aspects Very Important to be improved or developed
4	ever see traditional art performances rarely watch traditional art performances in a year Idly, invited or fill free time Participate in preserving and supporting the existence of traditional art More having the awareness to continue to participate in preserving and preserving traditional arts ... Important to be repaired or developed Social security of the performers Very important to be repaired or developed
…	…
139	ever see traditional art performances sometimes watch traditional art performances in a year Awareness of the importance of traditional arts and local culture to continue to be supported and cultivated as a form of appreciation Balance, ... Social security of actors is very important to be repaired or developed

Term Weighting aims to Identify the value or heft of a term based on the level of interest in the document. Period Frequency and Inverse Document Frequency (TF-IDF) is a weighting that is often used in information retrieval and text mining [19][20]. TF is the frequency of the case of a term in the document. IDF is the frequency of the case of a term in the whole document. The more frequently the term comes up in a document, the more-large the TF value and the smaller the IDF value.

$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times \log\left(\frac{N}{df_j}\right)$$

Where w_{ij} is the the weight of the j term for document i, tf_{ij} is the number of the case of term j in document i, N is the number of literature,df_j is the number of literature containing term j.

C. Evaluation

The testing process is carried out using cluster variance. Equation for calculate the cluster variance shown in (2).

$$v_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} \left(d_i - \bar{d}_i\right)^2$$

Where v_c^2 is the variance in cluster c, c is the 1 ... k, k is the number of clusters, n_c is the the amount of data in cluster c, d_i is the data i on a cluster, $\bar{d}_i$ is the average of data in a cluster. This variant is used to see the results of the variance of data distribution in a cluster. The smaller the value of Vw, the better the cluster. The equation for calculating Vw shown in (3).

$$v_w = \frac{1}{N - k} \sum_{i=1}^{k} (n_i - 1) v_i^2$$

Where N is the value point of all data, k is the value point of clusters, n_i is the value point of data members in cluster i. This variant is used to see the results of variance in data distribution between clusters. The greater the value of V_b, the better the cluster[21][22]. The equation for calculating V_b is shown in (4).

$$v_b = \frac{1}{k - 1} \sum_{i=1}^{k} n_i \left(d_{ij} - \bar{d}\right)^2$$

Where k is the value point of clusters d_{ij} is the data to j in a cluster to i, $\bar{d}$ is the average of $\bar{d}_i$. To see the variance of all clusters, it can be measured by (5). If the V value is getting smaller, the cluster value is getting better.

III. Result and Analysis

In this section, the results of the tests that were carried out in the previous chapter are explained, starting from Text Preprocessing, Modeling, and Evaluation. After the data is collected, the process carried out is Text Preprocessing, the results of which are shown in Table II. In this process, punctuation has been successfully removed.

Furthermore, after the punctuation in the previous process has been lost, the data is processed using Tokenizing, which breaks the sentence into separate words. The results of this test are shown in Table III.

TABLE II. The Result of Case Folding Process in Text Preprocessing

Audience ID	Opinion
1	ever see traditional art performances often watch traditional art performances in a year of awareness of the importance of traditional arts and local culture to continue to be supported and cultivated to take care ... it is very important to improve or develop
2	... adequate audio system sound stage and accompaniment that you have ever enjoyed has adequate quality accompaniment equipment from the show that you have ever seen conditions are very adequate equipment from the show includes costume property ... developed social security for artists is very important to be repaired or developed
3	... the educational values of messages or social criticism that you can catch enough to understand traditional art very well understand the meaning and meaning of the verbal language used to understand the meaning of conversations in the accompaniment narrative dialogue used during the performance can capture the message conveyed through the stage layout includes lighting and the use of property in staging messages of value education and social criticism delivered can be understood and understood messages delivered beautifully and deserve to strongly approve or agree on the message and values ...
4	... about traditional art, lack of understanding of the meaning and meaning of the verbal language that is lacking in understanding the meaning of conversational songs, narrative accompaniment used during ... developed price standards are very important to be improved or developed social security of the actors is very important to be repaired or developed
…	…
139	... through word of mouth rarely get information about the show schedule through the leaflet flyer leaflet brochure etc. sometimes get information about the show schedule through announcements in a line conversation groups etc ...

TABLE III. The Result of Tokenizing and Filtering Process in Text Preprocessing

Audience ID	Opinion
1	ever traditional art performances often watch traditional art performances a year awareness of traditional arts supported local culture cultivated keeping supporting the existence of traditional art having awareness of preserving wayang people condition ...
2	... adequate stage lighting quality ever visited adequate audio stage sound system accompaniment enjoy adequate quality accompaniment equipment performances have seen conditions are very adequate performance equipment including costume property ...
3	... supporting equipment has been watched very adequate location of the show quite easy to reach publication information about the event is available often information schedule social media shows often information schedule shows word of mouth sometimes schedule information ...
4	... marketing is very important to be improved developed production management is very important to be improved developed price standards are very important to be improved developed social security is very important actors to improve developed ...
…	…
139	... the ability of presenters is quite important to be improved developed facilities are very important to be improved developed equipment is very important to be improved developed quality of cultivation is quite important to be improved developed creative process of exploration is quite important to be improved developed use ...

The last Text Preprocessing is Steptemming. This process aims to get the basic words from words that have received affixes or other information. The results of this process are shown in Table IV.

After obtaining the results in Table IV, the weighting process is then performed using the TF-IDF Term Weighting technique. Weighting is done based on the level of importance in the document. The number of weights for each audience is 34. The figure is obtained based on the results of the formation of opinion variables. In Table V, these opinion variables are formulated into Keywords. So, each audience has 34 keywords.

Based on Table VI, the technique for determining data clusters that are Fuzzy clustering in which the existence of each data point in a cluster is determined by the degree of membership. The determination of data clusters is based on the Euclidean Distance form to measure the proximity between data. The concept of fuzzy clustering is to indicate the center of the cluster in advance which is each cluster center on average location. In the cluster center of the initial condition, the value level is still not precision, repetitive improvements are made to the membership degree and the center of the cluster in each point of data so that the center of cluster moves to the right position. At this stage, cluster testing has been carried out 5 times. Tests carried out using an initial cluster of 2 clusters with an error e <0.001. The results of this test are shown in Table VII.

To determine the accuracy of the cluster formed, the cluster variance testing technique Vw, Vb, and V. Based on the results shown in Table VII, it can be seen that the value of Vw = 0.284667123 the smaller, the better the distribution of data in the cluster. Meanwhile, to find out the accuracy of the variants of all clusters by using the calculation of the value of V. It can be seen in Table VII that the value of V is getting smaller so it can be concluded that the better cluster value V = 0.00000217.

TABLE IV. The Result of Stemming Process in Text Preprocessing

Audience ID	Opinion
1	ever traditional art performances often watch traditional art performances in the year of awareness of traditional arts local culture support cultivated keep supporting the existence of traditional arts awareness of preserving sustainable wayang people the condition of the building ever visited ...
2	... stage lighting has never been enough adequate stage sound system ever enjoyed the adequate quality of accompaniment equipment shows have seen very adequate conditions of equipment performances costumes property equipment support ever ...
3	the show is quite easy to reach the publication of event information is quite available often information schedule social media shows often information schedule shows word of mouth sometimes information schedule shows announcements by message group conversations etc. quality
…	…
139	... developed information dissemination reviewing news narratives is quite important developed marketing aspects are quite important developed production management is quite important developed price standards are quite important developed social guarantees art behavior is quite important to be developed

TABLE V. Keywords that have been Formed based on Audience Opinion

No	**Keyword**	**No**	**Keyword**
1	never	18	very bad
2	ever	19	bad
3	often	20	very understanding
4	rarely	21	quite understand
5	sometimes	22	not understand
6	awareness	23	do not understand
7	cheer up	24	understand
8	prankster	25	really understand
9	appreciation	26	quite understand
10	keep	27	do not completely understand
11	support	28	do not understand
12	lesson	29	understand
13	introspection	30	very catch
14	sustainable	31	just catch it
15	very good	32	less catch
16	quite good	33	not catch
17	very nice	34	catch it

TABLE VI. TF-IDF Calculation Results

No	Keyword	DF	IDF
1	appreciation	26	1.676
2	very nice	103	0.3
3	bad	0	0
4	quite good	55	0.927
5	quite understand	53	0.964
6	quite understand	76	0.604
7	just catch it	51	1.003
8	support	137	0.014
9	cheer up	65	0.76
10	introspection	12	2.45
…	…	…	…
34	not catch	2	4.241

TABLE VII. Results of the Audience Opinion Clustering Test

No.	**Experiment …**	**Vc**	**Vw**	**Vb**	**V**
1	Experiment 1	Cluster_1 = 11.833549726654	0.284671533	130582.4281	2.18001E-06
		Cluster_2 = 10.111108133333			
2	Experiment 2	Cluster_1 = 10.111108133333	0.715328467	130582.4281	5.47798E-06
		Cluster_2 = 11.833549726654			
3	Experiment 3	Cluster_1 = 11.833549726654	0.284671533	130582.4281	2.18001E-06
		Cluster_2 = 10.111108133333			
4	Experiment 4	Cluster_1 = 11.833549726654	0.284671533	130582.4281	2.18001E-06
		Cluster_2 = 10.111108133333			
5	Experiment 5	Cluster_1 = 10.111108133333	0.284667123	130582.4281	2.17901E-06
		Cluster_2 = 11.833549726654			
Average			0.456934307	130582.4281	3.4992E-06

IV. Conclusion

Combined Text Mining Fuzzy C-Means Clustering method that is proposed in this research. This method can be implemented in the clustering opinion of an audience of art performances to assess the level of audience satisfaction with an art show. This step is carried out as an effort to promote Indonesian culture. The novelty of this research that can measure the level of audience satisfaction with an art performance so that it is very important to more massive promote the Indonesian culture. The results of testing the value of V = 0.00000217 show that the combined method of Text Mining and Fuzzy C-Means Clustering has good performance. This is indicated by the decreasing value of V in each test. This can indicate that all cluster variants are getting better.

References

[1] S. C. Satapathy, V. Bhateja, and A. Joshi, "Proceedings of the International Conference on Data Engineering and Communication Technology : ICDECT 2016. Volume 2," Int. J. Eng. Res. Technol., 2020.

[2] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," IEEE Access, 2019.

[3] A. Abdul Aziz and A. Starkey, "Predicting Supervise Machine Learning Performances for Sentiment Analysis Using Contextual-Based Approaches," IEEE Access, 2020.

[4] I. Mujahidin, D. A. Prasetya, Nachrowie, S. A. Sena, and P. S. Arinda, "Performance tuning of spade card antenna using mean average loss of backpropagation neural network," Int. J. Adv. Comput. Sci. Appl., 2020.

[5] T. Traylor, J. Straub, Gurmeet, and N. Snell, "Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator," in Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019, 2019.

[6] X. Peng, Y. Bao, and Z. Huang, "Perceiving Beijing's 'city Image' across different groups based on geotagged social media data," IEEE Access, 2020.

[7] I. Mujahidin, D. A. Prasetya, A. B. Setywan, and P. S. Arinda, "Circular Polarization 5.5 GHz Double Square Margin Antenna in the Metal Framed Smartphone for SIL Wireless Sensor," in 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2019, pp. 1–6.

[8] X. Chen, Y. Zhao, Z. Cui, G. Meng, Y. Liu, and Z. Wang, "Large-Scale Empirical Studies on Effort-Aware Security Vulnerability Prediction Methods," IEEE Trans. Reliab., 2020.

[9] R. Yuwono and I. Mujahidin, "Rectifier using UWB microstrip antenna as electromagnetic energy harvester for GSM, CCTV and Wi-Fi transmitter," J. Commun., 2019.

[10] Z. Li, C. Zhang, S. Jia, and J. Zhang, "Galex: Exploring the evolution and intersection of disciplines," IEEE Trans. Vis. Comput. Graph., 2020.

[11] I. Mujahidin, S. H. Pramono, and A. Muslim, “5.5 Ghz Directional Antenna with 90 Degree Phase Difference Output,” 2018.

[12] D. Buenano-Fernandez, M. Gonzalez, D. Gil, and S. Lujan-Mora, “Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach,” IEEE Access, 2020.

[13] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, “Twitter and Research: A Systematic Literature Review through Text Mining,” IEEE Access, 2020.

[14] H. Gao, Y. Xu, Y. Yin, W. Zhang, R. Li, and X. Wang, “Context-Aware QoS Prediction with Neural Collaborative Filtering for Internet-of-Things Services,” IEEE Internet Things J., 2020.

[15] R. Yuwono, I. Mujahidin, A. Mustofa, and Aisah, “Rectifier using UFO microstrip antenna as electromagnetic energy harvester,” Adv. Sci. Lett., 2015.

[16] I. Hafeez, M. Antikainen, A. Y. Ding, and S. Tarkoma, “IoT-KEEPER: Detecting Malicious IoT Network Activity Using Online Traffic Analysis at the Edge,” IEEE Trans. Netw. Serv. Manag., 2020.

[17] A. Gavioli, E. G. de Souza, C. L. Bazzi, K. Schenatto, and N. M. Betzek, “Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods,” Biosyst. Eng., 2019.

[18] A. Seyedzadeh, S. Maroufpoor, E. Maroufpoor, J. Shiri, O. Bozorg-Haddad, and F. Gavazi, “Artificial intelligence approach to estimate discharge of drip tape irrigation based on temperature and pressure,” Agric. Water Manag., 2020.

[19] E. H. Kim, S. K. Oh, W. Pedrycz, and Z. Fu, “Reinforced fuzzy clustering-based ensemble neural networks,” IEEE Trans. Fuzzy Syst., 2020.

[20] H. Gan, “Safe Semi-Supervised Fuzzy C -Means Clustering,” IEEE Access, 2019.

[21] M. B. Ferraro, P. Giordani, and A. Serafini, “Fclust: An R package for fuzzy clustering,” R J., 2019.

[22] A. Alsarhan, Y. Kilani, A. Al-Dubai, A. Y. Zomaya, and A. Hussain, “Novel Fuzzy and Game Theory Based Clustering and Decision Making for VANETs,” IEEE Trans. Veh. Technol., 2020.