



PERBANDINGAN *FEATURE SELECTION WORD FREQUENCIES* DAN *TF/IDF* PADA PROSES *FUZZY CLUSTERING*

Elta Sonalitha¹⁾, Anis Zubair²⁾, Priyo Dari Mulya³⁾, Salnan Ratih Asriningtias⁴⁾ Bambang Nurdewanto⁵⁾

¹⁾Teknik Elektro, Fakultas Teknik, Universitas Merdeka Malang
Email: elta.sonalitha@unmer.ac.id

²⁾ Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka Malang
Email: anis_zubair@yahoo.com

³⁾ Sistem Komunikasi, Fakultas Ilmu Sosial & Politik, Universitas Merdeka Malang
Email: priyo.dari@unmer.ac.id

⁴⁾Teknologi Informasi, Pendidikan Vokasi, Universitas Brawijaya
Email: salnanratih@gmail.com

⁵⁾Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka Malang
Email: nurdewa@unmer.ac.id

Abstrak

Perundang-undangan yang diberlakukan di Indonesia tentang Pemajuan Budaya baru saja dicanangkan tahun 2017. Opini penonton terhadap sajian budaya merupakan salah satu data dasar yang sangat penting digunakan untuk referensi apakah suatu tontonan atau sajian seni itu masih diminati, dipahami dan diterima oleh masyarakat. Data opini penonton ini akan dikategorisasikan/diklusterisasi sehingga akan diperoleh informasi tentang kepuasan, kephahaman dan ketertarikan di bidang seni sehingga informasi ini dapat digunakan sebagai pendukung keputusan tentang sajian seni itu sendiri. Metode klasifikasi yang akan diterapkan di penelitian ini adalah *Fuzzy Clustering*. Hasil klasifikasi menggunakan *Fuzzy Clustering* dengan *Feature Selection TF/IDF* terhadap data opini penonton lebih pada menguji dan membandingkan *Feature Selection Word Frequencies* dengan *TF/IDF*. Hasil ini akan diperbandingkan dengan *Fuzzy Clustering* dengan *Feature Selection Word Frequencies* menghasilkan kesimpulan bahwa metode WF memiliki nilai *cluster variance* lebih kecil dibandingkan dengan metode *TF-IDF*, sehingga hasil cluster dari metode *TF-IDF* lebih baik.

Kata kunci: Budaya, Fuzzy Clustering, Opini, *TFIDF*, Word Frequencies

Abstract

Laws that have been put into effect in Indonesia regarding Cultural Advancement have just been launched in 2017. Audience opinion on cultural offerings is one of the most important basic data used to reference whether a spectacle or art presentation is still in demand, understood and accepted by the public. This audience opinion data will be categorized / clustered so that information about satisfaction, understanding and interest in the arts will be obtained so that this information can be used as supporting decisions about the art presentation itself. The classification method that will be applied in this research is Fuzzy Clustering. The results of the classification using Fuzzy Clustering with TF / IDF Feature Selection on audience opinion data is more on testing and comparing Word Frequency Feature Selection with TF / IDF. These results will be compared with Fuzzy Clustering with Word Frequencies Feature Selection resulting in the conclusion that the WF method has a smaller cluster variance value than the TF-IDF method, so that the cluster results from the TF-IDF method are better.

Keyword: Culture, Fuzzy Clustering, Opinion, *TFIDF*, Word Frequencies



I. PENDAHULUAN

Perundang-undangan Pemajuan Budaya harus dilaksanakan oleh seluruh bidang yang berkaitan, salah satunya adalah pendataan. Pendataan meliputi segala fakta di bidang seni dan budaya. Undang-undang pemajuan kebudayaan memprioritaskan budaya tradisional. Sebelum adanya undang-undang, data yang ada hanya bersifat regional per kelompok seni dan budaya saja. Maka diperlukan perancangan sistem otomatisasi pengolahan data dasar sebagai fakta yang dapat digunakan sebagai acuan untuk langkah-langkah pemajuan budaya.

Data tentang opini masyarakat terhadap suatu sajian budaya, semakin lama akan semakin banyak. Sehingga *big data* akan terus menerus bertambah. Sistem klasterisasi otomatis harus dirancang secara terstruktur. Salah satu metode untuk menganalisa hasil isian instrumen opini penonton yang berupa *text* adalah menggunakan *text mining*. Hasil dari pengelompokan test akan diproses menggunakan *Fuzzy Clustering*.

Text Mining yang diberlakukan pada data mentah untuk diproses *fuzzy clustering* ada beberapa metode dan biasa disebut *Feature Selection*. Dua diantaranya *Feature Selection* antara lain *Word Frequencies* dan *Term Frequency Inverse Document Frequency (TF-IDF)*. Penelitian ini lebih pada pengujian kedua metode ini dalam proses *Fuzzy Clustering*. Data yang digunakan dalam pengujian adalah data Opini Penonton pada sajian budaya tradisional di Malang Raya.

II. KAJIAN LITERATUR

A. *Text Mining*

Text mining memberikan sebuah set metodologi dan *tool* untuk menemukan, menggambarkan dan mengevaluasi pengetahuan dari kumpulan besar *text* dokumen. *Text mining* setara dengan analisis *text* yaitu proses untuk memperoleh

informasi yang berkualitas tinggi dari *text* yang dapat diperoleh dengan cara pengenalan pola, penandaan (*tagging*), ekstraksi informasi, serta mempelajari distribusi frekuensi kata. Beberapa proses dalam *text mining* diantaranya proses *text preprocessing*, *feature selection*, *modelling* dengan *fuzzy clustering* dan evaluasi.

Text Preprocessing merupakan sebuah fungsi untuk mengubah data *text* yang tidak terstruktur ke dalam bentuk data terstruktur. Beberapa tahap proses *text preprocessing* diantaranya sebagai berikut :

- *Case folding*
Menghilangkan karakter selain huruf dan merubah semua huruf menjadi huruf kecil
- *Tokenizing*
Memotong kalimat menjadi kata sejumlah kata penyusunannya
- *Filtering*
Mengambil kata-kata penting dari hasil *tokenizing* menggunakan teknik *stop list* yaitu membuang kata tidak penting
- *Stemming*
Mencari kata dasar dari hasil *filtering* dengan membuang imbuhan

B. *Feature Selection*

Terdapat dua pendekatan *feature selection* yang dapat digunakan untuk proses *text mining* diantaranya adalah sebagai berikut :

- *Word Frequency (WF)*
WF merupakan salah satu metode untuk menghitung bobot *term* dalam *text*. Dalam metode ini setiap *term* diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan *term* tersebut pada *text*. Nilai WF diperoleh dengan menghitung jumlah kemunculan atau *frekuensi term* (TF) dalam dokumen dan dinormalisasikan dengan jumlah total kata pada dokumen (N) dan dikalikan dengan 10.000, sesuai persamaan (1).



$$WF = \left(\frac{TF}{N} \right) \times 10000 \quad (1)$$

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

Dalam pencarian informasi TF-IDF bertujuan untuk menentukan nilai atau bobot pada sebuah kata (*term*) berdasarkan tingkat kepentingannya dalam dokumen. TF merupakan frekuensi kemunculan sebuah *term* dalam dokumen. IDF merupakan frekuensi kemunculan sebuah *term* pada keseluruhan dokumen. Semakin sering *term* muncul pada suatu dokumen maka nilai TF semakin besar dan nilai IDF semakin kecil. Persamaan untuk menghitung TF-IDF dapat dilihat dalam (1)[2]. Pada persamaan (1) D merupakan jumlah keseluruhan dokumen dan DF merupakan jumlah dokumen yang mengandung *term*.

$$TFIDF = TF \times IDF = TF \times \log_e \frac{|D|}{DF} \quad (2)$$

C. Fuzzy Clustering

Fuzzy clustering yang merupakan salah satu teknik untuk menentukan *cluster* data yang mana keberadaan tiap-tiap titik data dalam suatu *cluster* ditentukan oleh derajat keanggotaan. Penentuan *cluster* data di dasarkan pada bentuk *Euclidean Distance* untuk mengukur jarak kedekatan antar data. Konsep dari *fuzzy clustering* adalah menentukan pusat kluster terlebih dahulu yang merupakan lokasi rata-rata untuk tiap pusat kluster. Pada kondisi awal nilai pusat kluster masih belum akurat, maka dilakukan perbaikan secara berulang pada pusat kluster dan derajat keanggotaan di tiap-tiap titik data sehingga menghasilkan pusat kluster yang bergerak ke posisi yang tepat. Perulangan ini didasarkan pada minimisasi fungsi objektif yang menggambarkan jarak dari titik data ke pusat kluster yang terbobot oleh derajat keanggotaan titik data tersebut. Hasil *output* berupa deretan pusat kluster dan beberapa derajat keanggotaan untuk tiap-tiap data. Algoritma dari *fuzzy clustering* adalah sebagai berikut :

1. Menentukan parameter dari *fuzzy clustering* diantaranya :
 - a. Jumlah *cluster* (c)
 - b. Pangkat (w)
 - c. Maximum iterasi (*MaxIter*)
 - d. *Error* terkecil yang diharapkan (ϵ)
 - e. Fungsi objektif awal ($P_0=0$)
 - f. Iterasi awal ($t=1$)
2. Membangkitkan bilangan acak (μ_{ik}) 0-1 (matrik partisi U) sejumlah n x m dimana n adalah jumlah data dan m adalah jumlah kluster
3. Menentukan derajat keanggotaan untuk tiap-tiap data dan atribut (matrik partisi U)
 - a. Menghitung jumlah setiap kolom sesuai persamaan (3)

$$Q_i = \sum_{k=1}^c \mu_{ik} \quad (3)$$

- b. Menghitung derajat keanggotaan tiap-tiap kluster sesuai persamaan (4)

$$\mu_{ik} = \frac{\mu_{ik}}{Q_i} \quad (4)$$

dimana,
 i = data ke-i
 c= jumlah kluster.

4. Menghitung pusat kluster sesuai persamaan (5)

$$V_{kj} = \frac{\sum_{i=1}^n ((\mu_{ik})^2 * x_{ij})}{\sum_{i=1}^n (\mu_{ik})^2} \quad (5)$$

dimana,
 n = jumlah data,
 i = data ke-i,
 k = kluster ke-k,
 j = atribut ke-j

5. Menghitung fungsi objektif pada iterasi ke-t, sesuai persamaan (6)

$$P_t = \sum_{i=1}^n \sum_{k=1}^c \left(\left[\sum_{j=1}^m (x_{ij} - V_{jk})^2 \right] (\mu_{ik})^w \right) \quad (6)$$

dimana,



- n = jumlah data
- c = jumlah kluster
- m = jumlah atribut,
- X_{ij} = data ke-i atribut ke-j
- V_{kj} = pusat kluster ke k atribut ke j
- μ_{ik} = bobot data ke-i kluster ke-k
- w = pangkat

6. Menghitung perubahan matrik partisi sesuai persamaan (7)

$$\mu_{ik} = \frac{\left[\sum_{j=1}^m (x_{ij} - v_{kj})^2 \right]^{\frac{-1}{w-1}}}{\sum_{k=1}^c \left[\sum_{j=1}^m (x_{ij} - v_{kj})^2 \right]^{\frac{-1}{w-1}}} \quad (7)$$

7. Cek kondisi berhenti, jika $(|Pt - P(t-1)| < \epsilon)$ atau $I > \text{MaxIter}$ maka berhenti. Jika tidak maka $t=t+1$ ulangi langkah ke 4.

D. Evaluasi Fuzzy Clustering

Evaluasi untuk pengujian *fuzzy clustering* dilakukan dengan menggunakan *cluster variance* sesuai persamaan (8)

$$v_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (d_i - \bar{d}_i)^2 \quad (8)$$

Dimana :

- v_c^2 = variance pada cluster c
 - c = 1...k, k adalah jumlah cluster
 - n_c = jumlah data pada cluster c
 - d_i = data ke i pada suatu cluster
 - \bar{d}_i = rata-rata dari data pada suatu cluster
- Terdapat dua jenis *cluster variance* yaitu :

- *Variance within cluster* (V_w)

Varian ini digunakan untuk melihat hasil variansi penyebaran data yang ada dalam sebuah cluster. Semakin kecil nilai V_w maka semakin bagus *clusternya*. Persamaan untuk menghitung V_w dapat dilihat pada persamaan (9).

$$v_w = \frac{1}{N - k} \sum_{i=1}^k (n_i - 1)v_i^2 \quad (9)$$

Dimana,

- N = jumlah semua data
- k = jumlah cluster
- n_i = jumlah data anggota dalam cluster i

- *Variance between cluster* (V_b)

Varian ini digunakan untuk melihat hasil variansi penyebaran data antar cluster.

Semakin besar nilai V_b maka semakin bagus *clusternya*. Persamaan untuk menghitung V_b dapat dilihat pada persamaan (10)

$$v_b = \frac{1}{k - 1} \sum_{i=1}^k n_i (d_{ij} - \bar{d})^2 \quad (10)$$

Dimana,

- k = jumlah cluster
- d_{ij} = data ke j pada suatu cluster ke i
- \bar{d} = rata-rata dari \bar{d}_i

Untuk melihat varian semua cluster maka dapat diukur dengan persamaan (11)

$$v = \frac{V_w}{V_b} \quad (11)$$

Jika nilai V semakin kecil maka nilai cluster semakin baik

III. METODE PENELITIAN

Pada penelitian ini, dilakukan proses klasifikasi data opini penonton dengan *text mining* terhadap pertunjukkan seni yang diperoleh dari instrumen. Tahapan proses pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Proses *Text Mining* Data Opini Penonton.

- *Data Acquisition*

Terdapat 139 data yang dapat dilihat pada Tabel 1.

Tabel 1. Data Opini Penonton

No	Deskripsi
1	pernah melihat pertunjukan seni tradisional sering menonton pertunjukan seni tradisional ... sangat Penting untuk diperbaiki atau dikembangkan
2	pernah melihat pertunjukan seni tradisional sering menonton pertunjukan seni tradisional dalam setahun Kesadaran akan pentingnya seni tradisional dan budaya lokal untuk terus didukung dan dibudayakan Turut menjaga dan mendukung eksistensi seni tradisional ... Penting untuk diperbaiki atau dikembangkan
3	pernah melihat pertunjukan seni tradisional terkadang menonton pertunjukan seni tradisional dalam setahun Karena menyukai hiburan seni tradisional Sebagai wujud apresiasi diri Lebih memiliki kesadaran untuk terus turut menjaga dan melestarikan seni tradisional ... Aspek Pemasaran Sangat Penting untuk diperbaiki atau dikembangkan
4	pernah melihat pertunjukan seni tradisional jarang menonton pertunjukan seni tradisional dalam setahun Iseng, diajak atau mengisi waktu kosong Turut menjaga dan mendukung eksistensi seni tradisional



	Lebih memiliki kesadaran untuk terus turut menjaga dan melestarikan seni tradisional ... Penting untuk diperbaiki atau dikembangkan Jaminan sosial para pelaku seni sangat Penting untuk diperbaiki atau dikembangkan
...	...
139	pernah melihat pertunjukan seni tradisional terkadang menonton pertunjukan seni tradisional dalam setahun Kesadaran akan pentingnya seni tradisional dan budaya lokal untuk terus didukung dan dibudayakan Sebagai wujud apresiasi Keseimbangan, ... Jaminan sosial para pelaku seni sangat Penting untuk diperbaiki atau dikembangkan

- **Text Preprocessing**
 Tahapan proses *text processing* diantaranya proses *case folding*, *tokenizing*, *filtering*, dan *stemming*
- **Feature Selection**
 Tahap ini digunakan untuk memberikan pembobotan *term* dari hasil *text preprocessing*. Tahap ini dilakukan dengan dua pendekatan yang berbeda yaitu dengan *Word Frequency (WF)* dan *Term Frequency Inverse Document Frequency (TF-IDF)*
- **Modelling**
- Tahap ini digunakan untuk proses *cluster* data opini penonton dari hasil pembobotan *term* atau *Feature Selection* Menggunakan *fuzzy clustering*
- **Evaluasi**
 Tahap ini untuk melakukan evaluasi hasil *clustering* untuk dua jenis *Feature Selection* yang berbeda menggunakan *cluster variance*.

IV. HASIL DAN PEMBAHASAN

Term atau *keyword* yang diperoleh dari proses *text preprocessing* dari data opini penonton pertunjukkan dapat dilihat pada Tabel 2.

Tabel 2. *Term* atau *Keyword*

No	Deskripsi
1	tidak pernah
2	pernah
3	sering
4	jarang
5	terkadang
6	kesadaran
7	hibur
8	iseng
9	apresiasi
10	sering

11	dukung
12	pelajaran
13	introspeksi
14	lestari
15	sangat bagus
16	cukup bagus
17	bagus
18	sangat buruk
19	buruk
20	sangat paham
21	cukup paham
22	kurang paham
23	tidak paham
24	paham
25	sangat mengerti
26	cukup mengerti
27	kurang mengerti
28	tidak mengerti
29	mengerti
30	sangat tangkap
31	cukup tangkap
32	kurang tangkap
33	tidak tangkap
34	tangkap

Pengujian dilakukan untuk mengukur keefektifan metode *feature selection* antara WF dan TF-IDF. Pengujian dilakukan sebanyak lima kali untuk masing-masing metode. Terdapat beberapa jenis pengujian yang dilakukan diantaranya sebagai berikut:

- **Perbandingan Jumlah Iterasi**
 Perbandingan jumlah iterasi yang dibutuhkan kedua metode *feature selection* untuk melakukan proses *clustering* dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan Jumlah Iterasi

Nama Pengujian	Metode <i>Feature Selection</i>	
	WF	TF-IDF
Pengujian 1	7	7
Pengujian 2	6	6
Pengujian 3	8	7
Pengujian 4	7	7
Pengujian 5	7	7
Rata-rata	7	7

Berdasarkan Tabel 3. Kedua metode *feature selection* membutuhkan jumlah iterasi yang hampir sama yaitu membutuhkan kurang lebih 7 iterasi.

- **Perbandingan Waktu**
 Perbandingan waktu yang dibutuhkan kedua metode *feature selection* untuk melakukan proses *clustering* dapat dilihat pada Tabel 4.

Tabel 4. Perbandingan Waktu

Nama Pengujian	Metode <i>Feature Selection</i>	
	WF (detik)	TF-IDF (detik)



Pengujian 1	374	610
Pengujian 2	374	589
Pengujian 3	386	579
Pengujian 4	346	562
Pengujian 5	376	544
Rata-rata	371	577

Berdasarkan Tabel 4. metode WF membutuhkan waktu rata-rata 371 detik dan metode TF-IDF membutuhkan waktu rata-rata 577 detik. Hal ini menunjukkan bahwa metode WF membutuhkan waktu lebih cepat dibandingkan metode TF-IDF.

- Perbandingan *Error*
 Perbandingan *error* dari kedua metode *feature selection* untuk melakukan proses *clustering* dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan Error

Nama Pengujian	Metode <i>Feature Selection</i>	
	WF	TF-IDF
Pengujian 1	0.0008653634476	0.00081793797696
Pengujian 2	0.0000996973270	0.00085932204686
Pengujian 3	0.0005508434284	0.00092766544704
Pengujian 4	0.0003248212748	0.00000422767331
Pengujian 5	0.0007396104170	0.00004669813037
Rata-rata	0.000516067	0.00053117

Berdasarkan Tabel 5. metode WF memiliki nilai *error* rata-rata 0.000516067 dan metode TF-IDF memiliki nilai *error* rata-rata 0.00053117. Hal ini menunjukkan bahwa metode WF memiliki nilai *error* yang lebih kecil dibandingkan dengan metode TF-IDF.

- Perbandingan Cluster Variance
 Perbandingan hasil *cluster variance* dari kedua metode *feature selection* untuk melakukan proses *clustering* dapat dilihat pada Tabel 6.

Tabel 6. Perbandingan Cluster Variance

Nama Pengujian	Metode <i>Feature Selection</i>	
	WF	TF-IDF
Pengujian 1	0.00000547798411	0.0008653634
Pengujian 2	0.00000218001408	0.0000996973
Pengujian 3	0.00000547798411	0.0005508434
Pengujian 4	0.00000218001408	0.0000000019
Pengujian 5	0.00000547798411	0.0000000056
Rata	0.000004158	0.0003031824

Berdasarkan Tabel 6. metode WF memiliki nilai *cluster variance* rata-rata 0.000004158 dan metode TF-IDF memiliki nilai *cluster variance* rata-rata 0.0003031824. Hal ini menunjukkan bahwa metode WF memiliki nilai *cluster variance* lebih kecil dibandingkan dengan metode TF-IDF, sehingga hasil *cluster* dari metode TF-IDF lebih baik.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan proses pengujian yang dilakukan maka dapat ditarik kesimpulan bahwa :

1. Metode *feature selection* WF lebih optimal dibandingkan metode TF-IDF dalam proses *fuzzy clustering*.
2. Hasil *cluster* menggunakan metode WF lebih baik dibandingkan dengan metode TF-IDF, yang ditunjukkan dengan nilai *cluster variance* pada WF lebih kecil pada TF-IDF.

B. Saran

Hasil pengujian menunjukkan metode WF dan TF-IDF memiliki hasil yang hampir sama, maka perlu ditambahkan data uji dan penambahan metode *feature selection* untuk pembandingan sehingga diperoleh hasil yang lebih optimal.

REFERENSI

Arifin, A.Z., Mahendra, P.A.K., Ciptaningtyas, H.T., *Enhanced Confix Stripping Stemmer And Ant Algorithm For Classifying News Document in Indonesian Language*, International Conference on Information and Communication Technology and System, 2009, ISSN: 2085-1944, pp 149-158

Asriningtyas, Salnan Ratih; Sonalitha, Elta. Aplikasi K-Means Clustering untuk Pengelompokan Siswa ke Dalam Kelas Berdasarkan Nilai Akademis, Jenis Kelamin, Perilaku dan Nama Siswa di SMA Negeri 1 SRENGAT. Seminar Nasional Sistem Informasi



- (SENASIF). 2018. Vol 1. Issue 1.
Halaman 1179-1187
- Hapsari, R.K., Santoso, Y.J., *Stemming Artikel Berbahasa Indonesia dengan Pendekatan Confix Stripping*, 2015, Prosiding Seminar Nasional Manajemen Teknologi XX12
- Prasidhatama, A., Suryaningrum, K.M., *Perbandingan Algoritma Nazief dan Adriani dengan Algoritma Idris untuk Pencarian Kata Dasar*, Jurnal Teknologi dan Manajemen Informatika, 2018, Vol. 4, No. 1, pp 192-195.
- Program Studi MMT-ITS, ISBN:978-602-70604-1-8, pp 1-8
- Purbo, O.W., 2019, *Text Mining*, Yogyakarta : ANDI
- Sonalitha, Elta; Nurdewanto, Bambang; Ratih, Salnan; Sari, Nadia Roosmalita; Setiawan, Aries Boedi; Tutuko, Pindo. *Comparative Analysis of Tsukamoto and Mamdani Fuzzy Inference System on Market Matching to Determine the Number of Exports for MSMEs*. 2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS). Page 440-445.
- Sruthi, S., Shalini, S., *Sentence Clustering in Text Document Using Fuzzy Clustering Algorithm*, 2014, ICCICCT.
- Suresh, H., Raj., G., *An Unsupervised Fuzzy Clustering Method for Twitter Sentiment Analysis*, 2016, Internasional Conference on Computational System and Information Systems for Sustainable Solution.
- Tahitoe, A.D., Purwitasari, D., *Implementasi Modifikasu Enhanced Confix Stripping Stemmer untuk Bahasa Indonesia dengan Metode Corpus Based Stemming*.