Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 11, November 2019, hlm. 10623-10629 http://j-ptiik.ub.ac.id

Ekstraksi Topik Dokumen Berita Menggunakan Term-Cluster Weighting dan Clustering Large Application (CLARA)

Rizal Maulana[1], Sigit Adinugroho[2], Sutrisno[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]rizal_maulana@student.ub.ac.id, [2]sigit.adinu@ub.ac.id, [3]trisno@ub.ac.id

Abstrak

Perkembangan teknologi mempermudah untuk mendapatkan informasi dan informasi yang sering digunakan adalah media berita. Seiring perkembangan teknologi, berita dapat disebarkan melalui portal berita dalam bentuk *web-base* seperti Kompas, Detik, Tempo, dan lain lain. Pengguna teknologi informasi ada kalanya tidak memiliki waktu untuk membaca berita secara seksama dan sebagian tidak bisa mendapatkan berita yang diperlukan. Salah satu cara untuk menyelesaikan masalah tersebut adalah melakukan *clustering* dokumen berita setelah itu dilakukan ekstraksi topik untuk mendapatkan topik penting dari kelompok berita. Pada penelitian ini menggunakan *Clustering Large Application* (CLARA) untuk proses *clustering* karena CLARA merupakan optimasi dari *k-medoid* yang lebih baik dari *k-means* dari berbagai aspek dan pada ekstraksi topik menggunakan *term-cluster weighting* untuk menghitung bobot *term* pada *cluster*. Proses dari penelitian ini melakukan *text proprocessing* untuk mengubah dokumen menjadi data terstruktur, setelah itu melakukan *Singular Value Decomposition* (SVD) untuk mendekomposisi fitur. Kemudian melakukan *clustering* menggunakan CLARA dan untuk ekstraksi topik menggunakan *term frequency-inverse cluster frequency* (TF-ICF). Data yang digunakan pada penelitian ini merupakan data sekunder yang didapatkan dari *web* Kaggle yang merupakan dokumen berita berbahasa inggris. Hasil dari penelitian ini yaitu dengan jumlah dokumen 226 dan menggunakan 2 *cluster* menghasilkan nilai *silhoette score* 0,005. Sedangkan untuk akurasi dari hasil *clustering* sebesar 1 dengan jumlah pengambilan topik dari rentang 1 sampai 10

Kata kunci: *ekstraksi topik, clustering, CLARA, term-cluster weighting, TF-ICF, silhoette score.*

Abstract

The growth of technology makes it easy to get informations and a kind of informations is often used is news media. As technology growth, news can be spread through news portals in form of web-bases such as Kompas, Detik, Tempo, and many others. Users of information technology sometimes don't have time to read news all the time and sometime can't get the news that they need. One of many solution to solve the problem is to do clustering news documents and after that topic extraction is used to get get important topics from the news cluster. In this research using Clustering Large Application (CLARA) for the clustering algorithm because CLARA is an optimization of k-medoid which is better than k-means from various aspects and on topic extraction uses term-cluster weighting to calculate term weights in the cluster. The proses of this research is used text preprocessing documents so it become structured data, after that Singular Value Decomposition (SVD) used to decomose features. Then CLARA is used to clustering documents and for topic extraction is using term frequency-inverse cluster frequency (TF-ICF). Data in this research is secondary data that obtained from Kaggle website which is an English language news documents. The result of silhoette sore from using 226 documents and 2 clusters is 0,005. As for accuracy topic extraction is 1 with taken number topic from 1 to 10.

Keywords: *topic extraction, clustering, CLARA, term-cluster weighting, TF-ICF, silhoette score.*

1. PENDAHULUAN

Saat ini penyebaran informasi melalu dunia digital sangat cepat dan beragam seperti sosial media, berita, internet, majalah dan lain lain. Perkembangan teknologi mempermudah penyebaran informasi dan yang sering digunakan oleh pengguna teknologi informasi adalah media berita (Menda, 2017). Manusia memerlukan informasi dan berita mulai dari pendidikan, olahraga, politik, ekonomi, dan lain lain. Seiring perkembangan teknologi, berita dapat disebarkan melalui portal berita dalam bentuk *web-base* seperti Kompas, Detik, Tempo, dan lain lain. Pengguna teknologi informasi ada kalanya tidak memiliki waktu untuk membaca berita secara seksama dan sebagian tidak bisa mendapatkan berita yang diperlukan. Oleh karena itu perlu dilakukan ekstraksi topik.

Salah satu cara untuk mengatasi permasalahan tersebut adalah melakukan *clustering* untuk mengelompokan berita dan ekstraksi topik dari setiap *cluster* yang ada. Dengan begitu akan mengetahui topik penting dari sebuah *cluster* tanpa perlu membaca seluruh berita. Pada penelitian Menda (2017) dan Prihatini, et al. (2017) menggunakan *Non-negative Factorization* (NNF) dan *Latent Dirichlet Allocation* (LDA), untuk melakukan ektraksi topik dengan cara dekomposisi matrik *document-term matrix* (DTM) serta melakukan pembobotan kata tertinggi hingga terendah dan menggambil 5 kata tertinggi. Pada penelitian Abdurasyid, et al. (2018) dan Hudin, et al. (2018) menggunakan *improved k-means* dan *k-means* untuk *clustering* dokumen yang bertujuan untuk mencari dokumen-dokumen yang sejenis.

Pada penelitian ini akan menggunakan *Clustering Large Application* (CLARA) untuk proses clustering dan *term-cluster weigthting* untuk menghitung bobot kata pada setiap cluster. Penelitian ini menggunakan CLARA dibandingkan *k-means* karena jumlah data yang akan digunakan cukup besar maka diperlukan algoritma yang kuat untuk melakukan clustering. CLARA merupakan optimasi dari *k-medoid*, menurut Arora, et al. (2016) *k-medoid* lebih baik dari *k-means* hampir pada segala sisi, seperti waktu eksekusi, tidak sensitif pada *outlier*, dan mengurangi *noise* tetapi dengan kelemahan kompleksitas yang tinggi bila dibandingkan dengan *k-means*. Proses untuk ekstraksi topik menggunakan *term frequency-inverse cluster frequency* (TF-ICF) yang dilakukan pada setiap cluster dan mengambil kata (*term*) dengan nilai bobot terbesar.

Pada penelitian Prihatini, et al. (2017) dan Menda (2017) melakukan ekstraksi topik terhadap seluruh dokumen yang ada, sehingga hasil dari ekstraksi topik terlalu umum. Pada penelitian ini akan melakukan *clustering* supaya mendapatkan dokumen yang sejenis dan setelah itu melakukan ekstraksi topik. Ektraksi topik setiap cluster bisa dilakukan menggunakan TF-ICF yang berguna untuk mengitung bobot *term* yang muncul pada setiap cluster, TF-ICF merupakan *term frequency-invers document frequency* (TF-IDF) hanya saja bekerja pada level *cluster* (Ayad & Kamel, 2002).

2. DASAR TEORI

2.1. *Text Mining*

Text mining merupakan suatu proses ekstraksi informasi dari sekumpulan *text* yang tidak terstruktur (Abdurasyid, et al., 2018). Perbedaan antara *text mining* dan data *mining* hanya pada data yang digunakan, pada *text mining* menggunakan *text* atau data yang tidak terstuktur, sedangkan pada data *mining* menggunakan data yang sudah terstruktur (Menda, 2017). Pada dasarnya *text mining* berguna untuk merubah *text* supaya bisa dianalisis.

2.2. *Text Preprocessing*

Text preprocessing merupakan tahap awal dari *text mining*. Tujuan dari *text preprocessing* adalah mempersiapkan *text* untuk dapat diolah ketahap selanjutnya. *Text prepocessing* juga memproses semua *text* yang ada pada korpus dan merubahnya menjadi data terstruktur (Menda, 2017). Tahapan dari *text prepocessing* secara umum seperti *tokenizing*, *filtering*, dan *lemmatization.*

2.2.1. *Tokenizing*

Tokenizing merupakan proses memecah data pada korpus menjadi *token*. Pada proses ini tanda baca, spasi, dan karakter selain huruf akan dihilangkan (Hudin, et al., 2018). Cara yang mudah untuk melakukan *tokenizing* adalah dengan *regular expressions* (*regex*) atau bisa juga memisahkan dengan *whitespace*.

2.2.2. *Filtering*

Filtering merupakan proses untuk hapus *token* yang terlalu banyak muncul dan tidak memiliki arti signifikan dalam *text* (Hudin, et al., 2018). *Filtering* bisa juga disebut sebagai penghapusan kata *stopword* yang berada pada *stoplist.*

2.2.3. *Lemmatization*

Lemmatization hampir sama dengan stemming, yaitu membuat *term* menjadi kata dasar dengan cara menghilangkan imbuhan depan (prefiks), imbuhan belakang (suffiks), imbuhan tengah (infiks), atau imbuhan depan dan belakang (konfiks). Perbedaan dari *lemmatization* dan *stemming* adalah pada *lemmatization* adalah *lemma* (*output* dari *lemmatization*) lebih halus dibandingkan *stem word* (*output* dari *stemming*), karena pada *lemmatization* prosesnya menambahkan pengecekan *term* kedalam kamus atau *word net* (Manning, et al., 2009).

2.3. Pembobotan *Term*

Pembobotan *term* berfungsi untuk mengetahui suatu *term* muncul berapa kali dalam sebuah korpus. Pembobotan *term* memiliki 3 proses yaitu, *term frequency* (TF), *invers document frequency* (IDF), dan TF-IDF. Semakin besar nilai TF maka semakin penting sebuah *term*, semakin besar nilai IDF maka semakin penting sebuah *term* dan relevan terhadap dokumen (Xu & Qiu, 2015).

2.3.1. TF-IDF

TF-IDF merupakan perkalian dari TF dan IDF. Bobot pada TF-IDF mewakili jumlah kemunculan *term* pada suatu korpus, semakin tinggi bobot maka semakin sering *term* muncul pada korpus. Persamaan TF-IDF dapat dilihat pada Persamaan (1).

$$TF - IDF = 1 + log_{10} tf_{t,d} \times log_{10} N/df_t \quad (1)$$

$tf_{t,d}$ merupakan kemunculan *term* pada dokumen, *N* merupakan jumlah dokumen, dan df_t merupakan kemunculan dokumen pada korpus.

2.4. *Feature Transformation*

Feature transformation berguna untuk mendekomposisi sebuat matrik. Pada DTM nilai yang lebih dari 0 berjumlah sedikit, dikarenakan banyak *term* yang bersifat sinonim atau *polysemy*. Pada penelitian ini akan menggunakan *Single Value Decomposition* (SVD) yang masuk kedalam kelompok *approximation method. Approximation method* menggunakan pendekatan perkiraan solusi awal dan dilanjutkan dengan iterasi yang memberikan solusi lebih optimal (Menda, 2017).

2.5. *Clustering Large Application* (CLARA)

CLARA merupakan pengembangan dari metode *k-medoid* dimana membuat *subsample* dengan jumlah yang cukup tetapi tetap merepresentasikan seluruh data yang ada (Schubert & Rousseeuw, 2019). Pada CLARA tidak melakukan *k-medoid* terhadap semua data karena akan membuat kompleksitas bertambah, oleh karena itu CLARA melakukan *subsample* dengan jumlah $n' << n$ atau disarankan menggunakan $n' = 40 + 2k$ (*k* merupakan jumlah *cluster*), setelah itu objek yang lain akan dikelompokan dengan *medoid* yang terdekat sampai mencapai kondisi yang diinginkan (Vukčević, et al., 2019). Berbeda dari *k-means*, pada *k-medoid centroid* dari *cluster* merupakan salah satu data, pada *k-medoid centroid* biasa disebut dengan *medoid*. Karena CLARA merupakan *unsupervised algorithms* maka diperlukan perhitungan jarak. Perhitungan jarak pada penelitian ini menggunakan *euclidean distance.* Persamaan *euclidean distance* dapat dilihat pada Persamaan (2).

$$dist(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

x dan *y* merupakan data yang ingin diitung jaraknya, sedangkan i merupakan fitur yang digunakan.

2.6. *Topic Weighting*

Topic weighting berguna untuk menentukan bobot topik yang nantinya akan digunakan untuk menjadi hasil akhir dari ekstraksi topik. Karena pada penelitian ini melakukan *clustering* sebelum melakukan ekstraksi topik, maka proses *topic weighting* akan dilakukan pada setiap *cluster* yang telah dibuat. Bobot *term* akan dihitung kembali menggunakan TF-IDF yang bekerja pada *level cluster* yang bisa disebut dengan TF-ICF (Ayad & Kamel, 2002). Persamaan *topic weighting* dapat dilihat pada Persamaan (3).

$$w_{i,j} = tf_{i,j} \times log_{10} C/cf_j \quad (3)$$

$tf_{i,j}$ merupakan kemunculan *term* pada

cluster, C merupakan jumlah *cluster*, dan Cf_j merupakan kemunculan *term* pada *cluster*.

2.7. Evaluasi

Pada penelitian ini memiliki 2 tahap evaluasi, yaitu evaluasi *cluster* dan evaluasi ekstraksi topik yaitu evaluasi *cluster* dan evaluasi ekstraksi topik.

2.7.1. Evaluasi *Cluster*

Evaluasi *cluster* berguna untuk mengetahui berapa baik hasil dari *clustering* yang dilakukan. Pada menelitian ini menggunakan *silhouette coefficient* sebagai evaluasi *cluster*. Nilai dari *silhouette coeficient* mengukur berapa mirip objek dengan *cluster* sendiri (*cohesion*) dan membandingkan dengan *cluster* lain (*separation*), bila nilai *silhouette coefficient* mendekati 1 maka cocok dengan *cluster*-nya dan berbeda jauh dengan *cluster* lain, bila nilai *silhouette coefficient* mendekati -1 maka objek tidak sesuai dengan *cluster*-nya (Hudin, et al., 2018). Persamaan (4) merupakan persamaan *cohesion*.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \epsilon C_i, i \neq j} d(i,j) \quad (4)$$

Persamaan (5) merupakan persamaan *separation*.

$$b(i) = \frac{min}{k \neq i} \frac{1}{|C_k|} \sum_{j \epsilon c_k} d(i,j) \quad (5)$$

Persamaan (6) merupakan persamaan *silhoette score*

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, if\ |C_i| > 1 \quad (6)$$

2.7.2. Evaluasi Ekstraksi Topik

Evaluasi ekstraksi topik berguna untuk mengetahui akurasi yang dihasilkan dari ekstraksi topik. Karena hasil dari ekstraksi topik merupakan vektor yang didalamnya *term* dan *weight*, maka setiap *term* yang mewakili ekstraksi topik akan dilakukan evaluasi (Ayad & Kamel, 2002). Persamaan dari evaluasi ekstraksi topik dapat dilihat pada Persamaan (7)

$$A(i,j) = \frac{w(i,j)}{\sum_{k=1}^{C} w(k,j)} \quad (7)$$

Persamaan *overall topic accuracy* dapat dilihat pada Persamaan (8)

$$Overall\ Topic\ Accuracy = \frac{1}{C} \sum_{i=1}^{C} \frac{1}{l} \sum_{j=1}^{l} A(i,j) \quad (8)$$

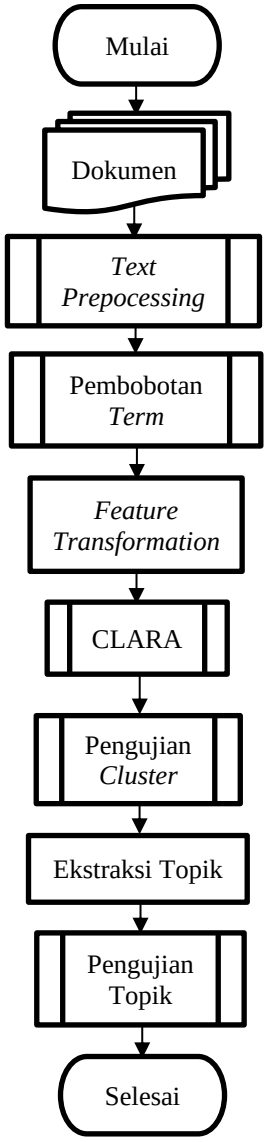
3. METODOLOGI PENELITIAN

3.1. Pengumpulan Data

Pengumpulan data pada penelitian ini menggunakan data sekunder yang disediakan pada *website* Kaggle. Pada data tersebut terdapat 9 fitur dan hanya akan diambil 1 fitur yaitu *content*, jumlah data yang digunakan sebanyak 107 dokumen berita dari rentang waktu 01 Januari 2017 sampai dengan 02 Januari 2017.

3.2. Metodologi Secara Umum

Metode yang digunakan dalam penelitian ini, direpresentasikan dalam Gambar 1.



Gambar 1. Metode Secara Umum

Metode pada penelitian ini diawali dengan dokumen yang akan digunakan. Setelah itu melakukan *text preprocessing* untuk bisa diproses untuk tahap selanjutnya. Kemudian pemboobotan *term* untuk menghitung kemunculan *term*. *Feature transformation* bertujuan untuk dekomposisi matrik. Lalu CLARA dilakukan untuk *clustering* dan setalah itu melakukan ekstraksi topik.

4. PENGUJIAN DAN ANALISIS

4.1. Pengujian Variasi Jumlah *Cluster*

Pada pengujian variasi jumlah dokumen menggunakan 107 dokumen dengan percobaan mulai dari 2 *cluster* sampai 9 *cluster*. Pemilihan percobaan *cluster* hanya sampai 9 *cluster* dikarenakan tidak ada nilai signifikan setelah 9 *cluster* teratas. Tabel 1 merupakan hasil dari pengujian variasi jumlah *cluster*.

Tabel 1. Hasil Pengujian Jumlah *Cluster*

Silhoette Score	**Rata-rata**
2 *Cluster*	**-0,011**
3 *Cluster*	-0,019
4 *Cluster*	-0,013
5 *Cluster*	-0,02
6 *Cluster*	-0,03
7 *Cluster*	-0,042
8 *Cluster*	-0,043
9 *Cluster*	-0,043

Hasil dari Tabel 1 menunjukan nilai rata rata terbesar pada *cluster* 2 dengan nilai rata-rata -0,011 yang menunjukan pembentukan *cluster* tidak bagus.

4.2. Pengujian Variasi Jumlah Dokumen

Pada pengujian variasi jumlah dokumen, jumlah dokumen yang digunakan sejumlah 107, 226, dan 337 dengan jumlah 2 *cluster*. Tabel 2 merupakan hasil dari pengujian variasi jumlah dokumen.

Tabel 2. Hasil Pengujian Jumlah Dokumen

Jumlah Dokumen	**Rata-rata**
107	-0,011
226	**0,005**
337	-0,006

Tabel 2 menunjukan bahwa dengan jumlah dokumen 226 menghasilkan nilai rata-rata yang tertinggi dengan nilai 0,005 yang menunjukan pembentukan *cluster* tidak terlalu bagus. Pada pengujian variasi jumlah dokumen tidak memiliki nilai yang signifikan antara penggunaan jumlah dokumen, dapat dilihat dari nilai rata-rata *silhoette score* yang perbedaannya hanya sedikit.

4.3. Pengujian Variasi Jumlah Topik

Pada pengujian variasi jumlah topik akan menggunakan jumlah dokumen 226 dengan 2 *cluster*. Pada pengujian ini menggunakan variansi jumlah topik dari 1 sampai 10 untuk melihat pengaruh jumlah topik terhadap hasil akurasi. Tabel 3 merupakan hasil dari pengujian variasi jumlah topik.

Tabel 3. Hasil Pengujian Jumlah Topik

Jumlah Topik	**Akurasi**
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1

Tabel 3 memiliki hasil yang bernilai 1 pada semua jumlah pengambilan topik. Hal tersebut dikarenakan bila ada *term* yang muncul pada semua *cluster* maka akan bernilai 0 dan karena jumlah *cluster* yang digunakan hanya 2 maka nilai bobot hanya akan terdapat pada satu *cluster* sedangkan *cluster* lainnya bernilai 0. Oleh karena itu pengambilan jumlah topik pada 2 *cluster* akan memiliki nilai akurasi 1 berapa pun jumlah topiknya.

4.3. Analisis dan Kesimpulan

Hasil penelitian menunjukan nilai *silhouette score* terbesar pada jumlah dokumen 226 dengan 2 *cluster*. Untuk pengambilan jumlah topik bisa mengambil jumlah mana saja dari Tabel 3 dikarenakan semua nilai akurasi berjumlah 1. Hasil dari penelitian ini menunjukan bahwa algoritma CLARA tidak terlalu baik untuk melakukan *clustering* dengan data yang digunakan pada penelitian ini.

Cluster yang terbentuk pada proses pengujian sering muncul *singleton cluster* yang menyebabkan nilai dari *silhoette score* menurun dan pembentukan *cluster* tidak maksimal. *Singleton cluster* terbentuk karena pada CLARA pemilihan data *subsample* secara *random* yang menyebabkan seluruh data tidak direpresentasikan pada data *subsample. Singleton cluster* juga bisa disebabkan dari data yang digunakan, pada penelitian ini data yang digunakan tidak memiliki perbedaan yang signifikan pada setiap dokumen. Karena data yang digunakan sebagian besar merupakan berita tentang politik atau ekonomi yang tidak jauh berbeda.

Pada pengujian jumlah variasi topik didapatkan nilai 1 untuk semua akurasi, karena pada perhitungan TF-ICF bila terdapat *term* yang muncul pada semua *cluster* maka bobotnya akan bernilai 0. Karena pada penelitian ini menggunakan 2 *cluster* maka nilai bobot hanya akan ada pada satu *cluster* sedangkan pada *cluster* lain akan bernilai 0. Pengaruh dari *singleton cluster* pada pengambilan jumlah topik yaitu topik yang diekstraksi hanya menggunakan satu dokumen.

5. PENUTUP

5.1. Kesimpulan

Bedasarkan penerapan algoritma CLARA dan *term-cluster weighting* dapat disimpulkan bahwa:

1. Pengaruh dari jumlah data yang digunakan membuat nilai rata-rata dari *silhoette score* tidak jauh berbeda satu sama lain, seperti yang dapat dilihat pada Tabel 2. Hal tersebut dikarenakan pada percobaan pengujian jumlah dokumen yang dilakukan akan terbentuk *singleton cluster* yang menyebabkan nilai *silhoette score* tidak terlalu baik.
2. Jumlah *cluster* berpengaruh terhadap pembentukan *cluster* semakin banyak jumlah *cluster* maka nilai *silhoette score* akan menurun. Penerunan nilai *silhoette score* terjadi karena semakin banyak *cluster* yang terbentuk maka jarak antara *cluster* semakin dekat atau bahkan beririsan. Semakin banyak jumlah *cluster* yang dilakukan maka waktu komputasi akan bertambah lama juga. Kinerja dari CLARA tergantung dari pemilihan *subsample* yang tepat atau yang merepresentasikan keseluruhan data, bila pemilihan *subsample* tidak tepat maka nilai *silhoette score* akan menurun.
3. Akurasi dari *clustering* menggunakan CLARA menghasilkan nilai *silhoette score* sebesar 0,005 dengan 2 *cluster*, menandakan pembentukan *cluster* kurang baik. Sedangkan pada akurasi topik didapatkan nilai 1 untuk jumlah topik dengan rentang 1 sampai 10.

5.2. Saran

Dari hasil analisis dan kesimpulan, dapat dikatakan penggunaan algoritma CLARA pada dokumen berita berbahasa inggris menghasilkan *cluster* yang kurang baik. Untuk penelitian selanjutnya dapat menggunakan data yang memiliki variasi yang beragam untuk menghindari pembentukan *cluster* yang kurang baik dan menghindari terjadinya *singleton cluster*. Pemilihan *subsample* pada CLARA mempengaruhi hasil akhir, oleh karena itu perlu adanya optimasi untuk pemilihan *subsample.*

6. DAFTAR PUSTAKA

Abdurasyid, Muhammad, Indriati, dan Rizal Setya Perdana. 2018. "Implementasi Metode Improved K-Means Untuk Mengelompokkan Dokumen Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 2 (10).

Ayad, Hanan, dan Mohamed Kamel. 2002. "Topic Discovery from Text Using Aggregation of Different Clustering Methods." *Springer-Verlag Berlin Heidelberg.*

Hudin, Muhammad Sholeh, M Ali Fauzi, dan Sigit Adinugroho. 2018. "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya)." *Jurnal Pengembangan Teknologi Informasi dan Ilmu*

Komputer 2 (11).

Manning, Christopher D, Prabhakar Ragghavan, dan Hinrich Schütze. 2009. "The term vocabulary and postings." Dalam *An Introduction to Information Retrival*, 32. Cambridge University.

Menda, Clara Sri. 2017. "Ekstraksi Tren Topik Portal Berita Online Menggunakan Non-Negative Matrix Factorization."

Schubert, Erich, dan Peter J Rousseeuw. 2019. "Faster K-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms."

Vukčević, M., V. Popović-Bugarin, dan E. Dervić. 2019. "DBSCAN and CLARA Clustering Algorithms and their usage for the Soil Data Clustering." Montenegro: IEEE.

Xu, Guixian, dan Lirong Qiu. 2015. "Technology Research of Tibetan Hot Topics Extraction." IEEE.