# Klastering Dokumen dengan Menambahkan *Metadata* Menggunakan Algoritma *COATES*

Indri Nurandini[1, a)] Arief Fatchul Huda[1, 2, b)]

[1]*Jurusan Matematika, UIN Sunan Gunung Djati, Bandung, Indonesia 40614*
[a)]*email: indrynurandiny@gmail.com*
[b)] *email: afhuda@gmail.com*

**Abstrak**

*Text mining* adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data tak terstruktur. Salah satu perkembangan *text mining* adalah ruang lingkup perbaikan dari pemanfaatan sebuah *"side information"* yang digunakan untuk membantu proses klastering yang lebih efisien. *"side information"* yang dimiliki data dapat membantu proses *text mining* jika "side information" tersebut bersifat informatif. Di dalam *"side information"* , *metadata* merupakan bagian dari *"side information"* yang dimiliki oleh data. Oleh karena itu, algoritma klastering partisi klasik dan model probabilistik dalam *text mining* telah dikembangkan untuk memproses data bersama "*side information*" dengan menggunakan algoritma *Content and Auxiliary attribute Based Text Clustering* (*COATES*). Adapun proses klastering ini menggunakan inisialisasi klaster dengan algoritma k-means berdasarkan perhitungan jarak *euclidean distance*.

*Kata kunci: text mining, metadata, klastering teks, algoritma k-means, algoritma COATES*

**Abstract**

*Text mining* is a process of pattern extraction in the form of useful information and knowledge from a large number of unstructured data sources. One of the development of *text mining* is the scope of improvement of utilizing a "*side information*" that is used to help more efficient clustering process. "*Side information*" owned data can help the process of *text mining* if "*side information*" is informative. In "*side information*", *metadata* is part of "*side information*" owned by data. Therefore, classical partition clustering algorithms and probabilistic models in *text mining* have been developed to process data along *side information* using the *Content and Auxiliary Attribute Based Text Clustering* (COATES) algorithm. The clustering process uses cluster initialization with k-means algorithm based on euclidean distance calculation.

*Keywords: text mining, metadata, clustering text, k-means algorithm, COATES algorithm*



## Pendahuluan

*Text mining* adalah penemuan informasi baru dan sebelumnya tidak diketahui dengan cara mengekstrak secara otomatis informasi dari sejumlah besar data teks tekstual yang tidak terstruktur. *Text mining* merupakan teknik yang digunakan untuk menangani masalah klasifikasi, klastering, *information extraction*, dan *information retrieval* [5]. *Text mining* telah banyak dikembangkan oleh para ilmuan dalam bidang komputasi. Diantaranya adalah Charu .C Aggarwal, Philip S, dan Yuchen Zhao yang berjudul "*On Text Clustering with Side Information*", dilakukan percobaan 50.080 jurnal, data ditambahkan *"side information"* dengan menggunakan algoritma *COATES (Content and Auxilary Based on Text Clustering)*

dan mengahasilkan algoritma klastering yang dianggap effisien untuk klastering dengan menggunakan *"side information"* [12].

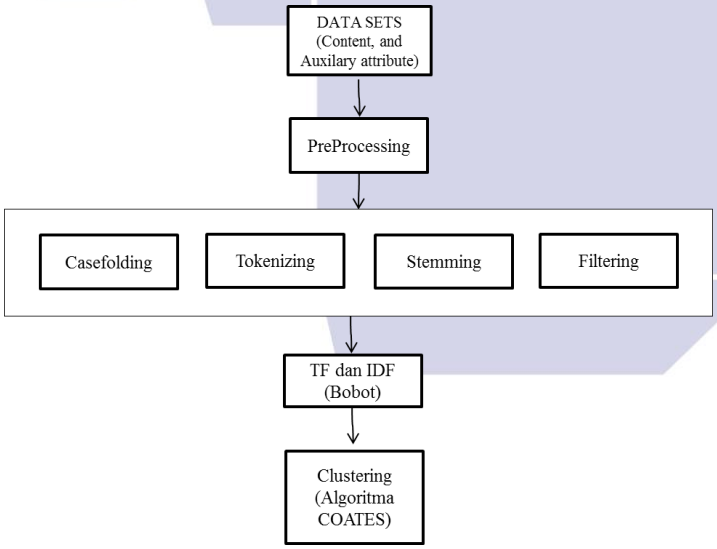
Pada Jurnal "*An Effective Clustering Approach for Mining Text Data Using Side Information*" yang ditulis oleh Monica. M dan Ganesh. J , mengenalkan Algoritma *COATES (Content and Auxilary Based on Text Clustering)* untuk klastering dan COLT untuk klasifikasi [8]. Pada Jurnal yang ditulis oleh Shilpa S. Raut dan Prof. V. Maral yang berjudul " *Text Clustering and Classification on The Use of Side Information*", Pengenalan Algoritma *COATES* yang dianggap sangat efektif dalam penggunaan *"side information"* sebagai attribut tambahan [9]. Pada jurnal yang ditulis oleh Ms. Neha Tiwari dan Prof. Gaima Singh yang berjudul " *A Framework For Mining Of Text Data With The Application Of Side Information*", ditambahakan fungsi JACARD dalam untuk menghitung jarak minimumnya dalam algoritma *COATES* [6]. Pada jurnal yang ditulis oleh Mrunal V. Uspani , dan Rucha C. Samant yang berjudul *"Meta Information Based On Text Clustering and Classification with the Use of COATES and COLT Algorithm"* [11], dan " *Clustering and Classification based on Meta Information using COATES and COLT Algorithm*", dikembangkan *"side information"* yang berupa meta informasi yang digunakan untuk membantu proses klastering dan klasifikasi [7].

Pada jurnal yang ditulis oleh Nikhil Patankar dan Sailee Salkar yang berjudul *"On the use of Side Information Based Improved K-means Algorithm for Text Clustering*" telah dikembangkan penggunaan *"side information"* berdasarkan pada algoritma *COATES* dengan menggunakan algoritma *k-means* dan menggunakan data sebagai objek percobaannya [10]. Dan pada jurnal yang ditulis oleh Shraddha S. Bhanuse, Shailes D. Kamble, Sandeep M. Kakde yang berjudul "*Text mining using Metadata for Generation of Side Information*", dijelaskan bahwa *metadata* merupakan bagian dari *"side information"* dan merupakan meta informasi dari data yang bersifat informatif yang dapat membantu proses klastering dalam *text mining*. Oleh karena itu dalam paper tersebut diusulkan untuk mengaplikasikan teknik klasterisasi dengan menggunakan algoritma *Content and Auxiliary based Text Clustering* (*COATES*) pada *metadata* dalam paper sebagai *"side information"* data. Dijelaskan bahwa *metadata* dapat berupa judul, abstrak, publisher, *keyword* dari sebuah paper [1].

Adapun alasan utama untuk merancang Algoritma *Text mining* yang efektif adalah meningkatnya jumlah data tekstual di sekitarnya. Dalam penambangan teks, banyak masalah yang diangkat karena beberapa domain aplikasi seperti informasi web, data digital, dan jaringan yang berbeda Dalam domain ini, sejumlah besar *"side information"* dikaitkan dengan dokumen. Tetapi cukup sulit untuk menghitung pentingnya *"side information"* karena penggabungan *"side information"* dapat meningkatkan kualitas proses penambangan. Untuk itu adanya ruang lingkup perbaikan dalam *"side information"* ini adalah berupa *metadata* [2]. Oleh karena itu, dengan menggunakan pendekatan yang memastikan pengelompokan karakteristik dari *"side information"* dengan isi teks. Hal ini akan memperbesar efek pengelompokan keduanya, dimana isi teks dan *"side information"* teks memberikan petunjuk yang sama tentang sifat pengelompokan yang mendasarinya, dan mengabaikan aspek-aspek di dalamnya. Untuk mencapai tujuan tersebut, akan digunakan algoritma *Content and Auxiliary attribute Based Text Clustering* (*COATES*) untuk menambahkan metadata dalam proses klastering dan melihat hasil klaster yang dibentuk ketika ditambahkan metadata.

**Metode**

Pada bagian ini akan dijelaskan bagaimana mengolah data set yang akan digunakan untuk mengklasteringkan data menggunakan algoritma *Content and Auxilary Based on Text Clustering* (*COATES*).



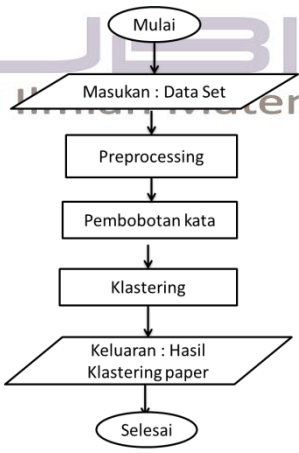

Gambar 1. Alur Penelitian

Data yang digunakan dalam penerapan *algoritma Content and Auxiliary attribute Based Text Clustering* (*COATES*) dalam text mining adalah jurnal/ paper ilmiah yang diambil dari web Hindawi.com yang terdiri dari 4 jenis kategori yaitu *artificial intelligence, data structures*, *information retrieval*, dan *text mining*. jumlah paper yang diambil adalah sebanyak 50 paper terdiri dari 10 paper *Artificial intelligence*, 10 paper data structure, 11 paper *information retrieval* 19 paper *text mining* .

Penelitian ini bertujuan untuk mengklaster data dengan memanfaatkan *metadata* dalam prosesnya. Dengan Menggunakan algoritma *COATES*, pendekatan dapat memberikan keuntungan dalam pengolahan data dengan menggunakan *metadata* sebagai *"side information"* yang dimiliki oleh data.

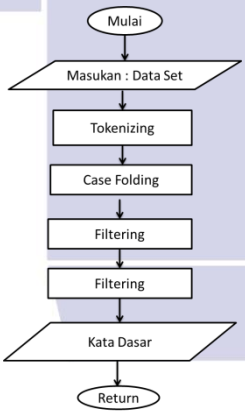
Sebelum data set dapat diklasterkan, data set harus dipersiapkan terlebih dahulu dengan menggunakan *machine learning* yang digunakan yaitu *python* dengan *package* yaitu NLTK dan dilakukan proses persiapan data *(preprocessing)*, yang dilanjutkan dengan pembobotan kata. Kemudian dilakukan proses klastering data dengan menggunakan algoritma *Content and Auxiliary attribute Based Text Clustering* (*COATES*). Gambar 2. menunjukan alur proses klastering.






Gambar 2. Alur klastering

*Text preprocessing* bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk proses selanjutnya. Gambar 3. menunjukkan potongan diagram alur proses *preprocessing*.




Gambar 3. Alur *Preprocessing*

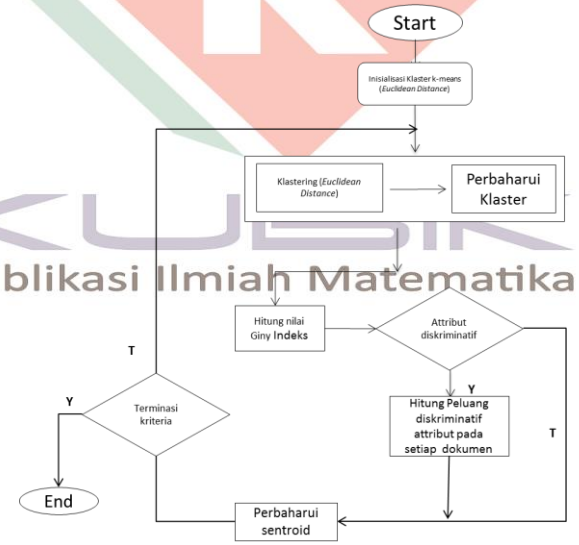
Dari langkah *preprocessing* ini menghasilkan kata (*term*). *Term-term* yang telah melalui proses *stemming* kemudian dihitung bobotnya dengan menggunakan TF-IDF dengan menggunakan persamaan

$$W_t = Tf_t \; x \; Idf_t \quad (1)$$

Ket :

- $W_t$: Bobot dari setiap *term*.
- $Tf_t$ : Jumlah dari setiap *term*.
- $Idf_t$: *Inverse Document frequency* dari setiap term.

Hasil dari perhitungan bobot kemudian disimpan [4] untuk proses selanjutnya yaitu klastering dengan menggunakan algoritma *COATES*. Untuk Alurnya sendiri dapat dilihat pada Gambar 4. Alur Algoritma *COATES*






Gambar 4. Alur Algoritma *COATES*

Didefinisikan bahwa *auxilary attribute* $X_i$ mempunyai sebuah peluang untuk bisa mendiskriminasi sebuah dokumen $T_i$ yang akan dihitung seberapa besar probabilitasnya sehingga akan mempengaruhi keberadaan dokumen $T_i$ yang didiskriminasi dalam klaster sebelumnya. Nilai *Giny Indeks* [3] diperoleh dari kebradaan relatif attribut dalam klaster yang dinotasikan dengan $p_{rj}$. Dengan melihat nilai *Giny Indeks* ($G_i$) akan diambil beberapa nilai yang memang dianggap dapat mendiskriminasi kedalam himpunan $R_i$. Dan selanjutnya attribut dalam himpunan $R_i$. akan dibandingkan satu persatu pada semua dokumen yang dimiliki. Adapun keberadaan relatif attribut pada dokumen-dokumen dalam klaster ke-j didefinisikan:

$$p_{rj} = \frac{f_{rj}}{\sum_{m=1}^{k} f_{rm}} \quad (2)$$

Keterangan :

$f_{rj}$ : kata dalam auxilary attribut yang tidak terdapat dalam corpus klaster ke-i

$f_{rm}$ : Banyak kata auxilary attribut yang tidak terdapat dalam semua klaster

Untuk *Giny Indeks* dari attribut r didefinisikan oleh:

$$G_r = \sum_{j=1}^{k} P^2{}_{rj} \quad (3)$$

Adapun inisialisasi klaster yang digunakan meliputi algoritma k-means berdasarkan perhitungan jarak *euclidean distance*.

**Analisa Hasil**

Hasil percobaan dengan dataset 50 paper, yang terdiri dari 4 jenis kategori yaitu *Artificial intelligence*, *Data Structures, Information Retrieval, dan Text Mining* Hindawi dengan metode *COATES* yang didalamnya menggunakan Algoritma *k-means*.

Diperoleh hasil klastering dataset adalah sebagai berikut :

Tabel 1. Hasil inisialisasi klaster dengan Euclidean Distance

| Klaster | Node |
|---|---|
| 1 | Doc 32 Doc 50 |
| 2 | Doc 15 |
| 3 | Doc 26 |
| 4 | Doc 1 Doc 2 Doc 3 Doc 4 Doc 5 Doc 6 Doc 7 Doc 8 Doc 9 Doc 10 Doc 11 Doc 12 Doc 13 Doc 14 Doc 16 Doc 17 Doc 18 Doc 19 Doc 20 Doc 21 Doc 22 Doc 23 Doc 24 Doc 25 Doc 27 Doc 28 Doc 29 Doc 30 Doc 31 Doc 33 Doc 34 Doc 35 Doc 36 Doc 37 Doc 38 Doc 39 Doc 40 Doc 41 Doc 42 Doc 43 Doc 44 Doc 45 Doc 46 Doc 47 Doc 48 Doc 49 |

Tabel 2. Hasil Klaster *COATES*

| Klaster | Node |
|---|---|
| 1 | Doc 32 Doc 50 |
| 2 | - |
| 3 | - |
| 4 | Doc 1 Doc 2 Doc 3 Doc 4 Doc 5 Doc 6 Doc 7 Doc 8 Doc 9 Doc 10 Doc 11 Doc 12 Doc 13 Doc 14 Doc 15 Doc 16 Doc 17 Doc 18 Doc 19 Doc 20 Doc 21 Doc 22 Doc 23 Doc 24 Doc 25 Doc 26 Doc 27 Doc 28 Doc 29 Doc 30 Doc 31 Doc 33 Doc 34 Doc 35 Doc 36 Doc 37 Doc 38 Doc 39 Doc 40 Doc 41 Doc 42 Doc 43 Doc 44 Doc 45 Doc 46 Doc 47 Doc 48 Doc 49 |



Terlihat perubahan klaster yang dipengaruhi oleh penambahan "*side information*" berupa *metadata* yang memberikan "*side* information" dalam analisis klaster. Adapun Hasil Perubahan klaster yang dipengaruhi oleh penambahan data berupa *metadata* pada algoritma *COATES* berdasarkan pada perhitungan jarak dengan menggunakan *euclidean distance* yaitu berpindahnya dokumen 15 dari klaster 2 kedalam klaster 4 dan dokumen 26 kedalam klaster 4.

**Kesimpulan**

Algoritma *Content and Auxiliary attribute Based Text Clustering* (*COATES*) digunakan untuk mengklasterkan data dengan memanfaatkan *"side information"* yang akan membantu proses

pengelompokan yang lebih koheren dan efisien. Inisialisasi klaster yang digunakan dapat mempengaruhi hasil akhir klaster dan dapat mempengaruhi proses pengelompokan data.

**Referensi**


[1] Shraddha S. Bhanuse, Shailesh D. Kamble, Sandeep M. Kakde. "*Text Mining using Metadata for Generation of Side Information*".in Proc. ICISP(2015) .pp 807-814.

[2] Wikipedia, "*Metadata*" (online), (https://id.wikipedia.org/wiki/*Metadata*. diakses tanggal 5 september, pukul 13.20)

[3] C. C. Aggarwal and H. Wang, *Managing and Mining Graph Data*.New York, NY, USA: Springer, 2010

[4] C. Silverstein and J. Pedersen, "*Almost-constant time clustering of arbitrary corpus sets*," in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 60–66

[5] C. C. Aggarwal and C.-X. Zhai, *Mining Text Data*. New York, NY, USA: Springer, 2012.

[6] Ms. Neha Tiwari dan Prof. Gaima Singh. " *A Framework For Mining Of Text Data With The Application Of Side Information*". 2015

[7] Mrunal V. Uspani , dan Rucha C. Samant. " *Clustering and Classification based on Meta Information using COATES and COLT Algorithm*".2015

[8] Monica. M dan Ganesh. J. "*An Effective Clistering Approach for Mining Text Data Using Side Information*". 2014

[9] Shilpa S. Raut dan Prof. V. Maral. dul " *Text Clustering and Classification on The Use of Side Information*" . 2014

[10] Nikhil Patankar dan Sailee Salkar. "*On the use of Side Information Based Improved K-Means Algorithm for Text Clustering*". 2015

[11] Mrunal V. Uspani , dan Rucha C. Samant. "*Meta Information Based On Text Clustering and Classification with the Use of COATES and COLT Algorithm*". 2015

[12] C. C. Aggarwal and P. S. Yu, "*On text clustering with side information*," in Proc. IEEE ICDE Conf. Washington, DC, USA,2012.