# ABSTRACT

Online news websites have massive data flows, but clusterization of news texts haven't appropriate with the content itself. Document clusterization is unsupervised text processing technique which there is no defined category or class in grouping data text. In order to do the clustering of documents, we need data representation that able represent the data based on the important of the words in documents.

Vector space model (VSM) uses term frequency and inverse document frequency (TF-IDF) as a representation of word weighting and able to represent data appropriately. For representing the data texts, TF-IDF works in vector space as well as cosine similarity which is a measure for finding closeness among the document. This Algorithm measure the angle of two vektors (a couple of document) in a vector space. The usage of TF-IDF weighting feature and the combination of that two method have an ability for determining features in clustering process. For knowing the usage influence of that methods in clustering documents can be found by examine the comparation of that two methods in clustering documents.

Evaluation of clusterization result has done by measure silhouette coefficient value from the result of document clustering. Based on the experiment, it is found that the calculation of cosine similarity not always improve the document clustering result. In some feature selection parameters, the use of TF-IDF without cosine similarity has better clustering results. Maximum and minimum limit values of document frequency (df) affect the calculation of the best cluster results. The result of the experiment obtained the best number of cluster is on maksimum df 0.8 and minimum df 0.3. The best number of cluster has reached medium structure with silhouette coefficient value 0,604287007 for clustering process with TF-IDF weighting and 0,618543225 for clustering process with combination of TF-IDF weighting and cosine similarity.

**Keywords :** *Document clustering*, , TF-IDF, *feature*, *cosine similarity measure*, K-means, *Indonesian text*

# INTISARI

Portal berita *online* memiliki aliran data yang sangat besar, namun pengelompokan berita belum sesuai dengan konten atau isi berita. Klasterisasi dokumen merupakan teknik pengolahan teks tidak terbimbing dimana tidak didefinisikan terlebih dahulu kategori atau pembagian kelas di dalam pengelompokkan data teks. Di dalam melakukan pengolahan teks guna mengelompokkan dokumen diperlukan representasi data yang mampu mewakili data berdasarkan tingkat pentingnya kata di dalam dokumen tersebut.

*Vector space model* (VSM) menggunakan *term frequency and inverse document frequency* (TF-IDF) sebagai representasi untuk pembobotan kata dan mampu merepresentasikan data dengan baik. Untuk merepresentasikan data teks tersebut TF-IDF bekerja pada ruang vektor sama halnya dengan kesamaan kosinus yang merupakan perhitungan untuk mengetahui nilai kedekatan antar dokumen. Algoritme ini menghitung sudut dua vektor (pasangan dokumen) di dalam ruang vektor. Penggunaan fitur pembobotan TF-IDF dan kombinasi dari kedua metode tersebut memiliki kemampuan sebagai penentu fitur dalam proses klasterisasi dokumen. Untuk mengetahui pengaruh terhadap penggunaan kedua metode tersebut dalam mengklaster dokumen dilakukan pengujian terhadap perbandingan antara kedua proses tersebut dalam mengklaster dokumen.

Evaluasi hasil klasterisasi dilakukan dengan menghitung nilai *silhouette coefficient* dari hasil klasterisasi dokumen. Berdasarkan percobaan yang dilakukan, didapatkan bahwa perhitungan kesamaan kosinus tidak selalu meningkatkan hasil klasterisasi dokumen. Pada beberapa parameter pemilihan fitur, penggunaan TF-IDF tanpa kesamaan kosinus memiliki hasil klasterisasi yang lebih baik. Nilai batas maksimal dan minimal dokumen frekuensi (df) berpengaruh pada pencarian jumlah klaster terbaik. Dari hasil percobaan didapatkan jumlah klaster terbaik pada nilai maksimum df 0,8 dan nilai minimum df 0,3. Jumlah klaster terbaik sudah mencapai kelompok *medium structure* dengan nilai *silhouette coefficient* 0,604287007 untuk proses klasterisasi dengan fitur pembobotan TF-IDF dan 0,618543225 untuk proses

klasterisasi dengan kombinasi TF-IDF dan perhitungan kesamaan kosinus.

**Kata kunci –** Klasterisasi dokumen, TF-IDF, perhitungan kesamaan kosinus, K-means, teks bahasa Indonesia