Klasifikasi Dokumen Menggunakan Metode *k*-Nearest Neighbor (kNN) dengan *Information Gain*

Document Classification using *k*-Nearest Neighbor (*k*NN) Method with *Information Gain*

Pratama Dwi Nugraha[1], Said Al Faraby[2], Adiwijaya[3] [1,2,3]

Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom

[1]pratamadwinugraha@gmail.com, [2]saidalfaraby@telkomuniversity.ac.id, [3] adiwijaya@telkomuniversity.ac.id

ABSTRAK

Pada saat ini, informasi sangatlah penting bagi semua orang, kebutuhan akan informasi semaking meningkat seiring dengan semakin canggihnya teknologi sekarang ini. Informasi yang dibutuhkan saat ini semakin tinggi, baik informasi bersifat umum maupun informasi bersifat khusus. Tapi terkadang informasi yang didapat tidak sesuai dengan apa yang diinginkan. Sehingga muncul sebuah permasalah pada saat pencarian data yang dibutuhkan. Sehingga diperlukan sebuah cara untuk memperoleh data yang valid. *Document Classification* (Klasifikasi dokumen) dapat membantu dalam proses pencarian sebuah data atau dokumen yang valid sesuai dengan apa yang kita butuhkan. Penggunaan klasifikasi dokumen tidak lain untuk membantu dalam proses pencarian data dengan cepat, tepat dan valid. Klasifikasi dokumen mengelompokan dokumen yang sesuai dengan kategori yang terkandung pada dokumen tersebut. Untuk menyelesaikan permasalahan yang ada, metode yang akan digunakan pada penelitian ini yaitu Metode *K-Nearest Neighbot* (KNN) dan *Information Gain*.

Kata kunci : *Document Classification, K-Nearest Neighbor, Information Gain.*

1. PENDAHULUAN

Pada sekarang ini kebutuhan akan informasi semakin meningkat seiring dengan berkembangnya teknologi dalam menyebarkan informasi kepada masyarakat. Informasi yang dibutuhkan mengalami banyak perkembangan mulai dari informasi yang bersifat umum hingga informasi yang bersifat khusus. Banyaknya informasi dan dokumen yang tersedia mendorong pengguna untuk mencari cara lebih cepat dalam mendapatkan informasi dan dokumen yang dibutuhkan. Jika waktu pencarian terlalu lama, maka manfaat dari informasi yang diperoleh dapat berkurang. Hal ini dikarenakan informasi yang diperoleh sudah masuk waktu yang sudah tidak berguna atau tidak valid.

Klasifikasi dokumen dapat membantu proses pencarian sebuah dokumen dengan cepat dan tepat. Klasifikasi dokumen mengelompokan dokumen yang sesuai dengan katergori yang terkandung pada dokumen tersebut. Permasalahan klasifikasi dokumen bisa diselesaikan dengan banyak metode, salah satu diantaranya adalah *K-Nearest Neighbor (KNN)*. Algoritma KNN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data yang jaraknya paling dekat dengan objek tersebut. Algoritma KNN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut[6]. Algoritma KNN adalah algoritma yang kompleks, tapi membutuhkan proses yang lama dalam pengklasifikasiannya. Karena kompleksitas yang tinggi, oleh karena itu dibutuhkan metode untuk meningkatkan performanya kecepatan waktu untuk proses klasifikasinya.

Salah satu metode untuk meningkatkannya adalah *Information Gain*. Dimana pada pemilihan metode ini bertujuan untuk menyelesaikan masalah tentang klasifikasi data atau dokumen dimana pada saat metode ini berjalan, performasi kecepatan waktu pada saat ada data yang masuk pada proses klasifikasi akan meningkat. Sehingga apa yang menjadi masalah tentang klasifikasi data atau dokumen bisa diselesaikan. *Information Gain* adalah ukuran efektifitas suatu attribut dalam menglasifikasikan data. *Information gain* juga merupakan pengurangan yang diharapkan dalam *entropy*. Dalam *machine learning*, ini dapat digunakan untuk menentukan urutan atribut atau mempersempit atribut yang dipilih.

2. TINJAUAN PUSTAKA

2.1. Document Classifciation

Document Classification (klasifikasi dokumen) adalah sebuah proses untuk menanggulangi munculnya sebuah masalah sederhana akan jumlah dokumen yang setiap hari semakin bertambah jumlahnya. Manfaat dari klasifikasi dokumen adalah untuk pengorganisasian dokumen[2]. Penggunaan klasifikasi dokumen tidak lain untuk membantu proses pencarian sebuah dokumen dengan cepat dan tepat. Klasifikasi dokumen mengelompokan dokumen yang sesuai dengan katergori yang terkandung pada dokumen tersebut.

2.2. *K*-Nearest Neighbor

KNN adalah salah satu metode dimana metode ini melakukan klasifikasi berdasarkan data *training* atau data pembelajaran dilihat dari jarak yang paling dekat dengan objek berdasarkan nilai k. Metode ini bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan training sample. Diberikan suatu titik query, selanjutnya akan ditemukan sejumlah K objek atau titik training yang paling dekat dengan titik query. Nilai prediksi dari query akan ditentukan berdasarkan klasifikasi tetanggaan (Tri, 2010).

Sebelum melakukan perhitungan dengan metode *K-Nearest Neighbor*, terlebih dahulu harus menentukan data latih dan data uji. Kemudian akan dilakukan proses perhitungan untuk mencari jarak menggunakan *Euclidean*. Teknik ini sangat sederhana dan mudah diimplementasikan. Mirip dengan teknik clustering, yaitu mengelompokkan suatu data baru berdasarkan jarak data baru itu ke beberapa data/tetangga terdekat. Pertama sebelum mencari jarak data ke tetangga adalah menentukan nilai K tetangga (neighbor). Lalu, untuk mendefinisikan jarak antara dua titik yaitu titik pada data training dan titik pada data testing, maka digunakan rumus Euclidean.

$$d(x_i x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2} \quad [1]$$

Keterangan :

$d(x_i, x_j)$: Jarak *Euclidean (Euclidean Distance).*

$(x_i), (x_j)$: *record* ke-*i* , *record* ke-*j*

(a_r) : data ke-*r*

i,j : 1,2,3,...n

n : dimensi objek[3]

2.7. *Information Gain*

Informasi Gain dan *Entropy* adalah fungsi dari distribusi probabilitas yang mendasari teori komunikasi. Ini merupakan penurunan diharapkan *entropy* disebabkan oleh partisi sampel sesuai dengan atribut ini[2]. *Entropy* mengukur suatu ketidakpastian dari suatu variabel acak. Berdasarkan *entropy*, untuk fitur pilihan ukuran yang disebut "*Informasi Gain*" didefinisikan.[9]

Entropy adalah suatu parameter untuk mengukur tingkat keberagaman (heterogenitas) dari kumpulan data. Semakin heterogenitas, nilai *entropy* semakin besar[5]. Fitur yang dipakai dalamFungsi *entropy* dituliskan sebagai berikut :

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

[5]

Keterangan :

c : jumlah nilai yang ada pada atribut target(jumlah kelas klasifikasi)

p_i : jumlah proporsi sampe (peluang) untuk kelas i

Information Gain merupakan ukuran efektifitas suatu attribut dalam menglasifikasikan data.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|s_v|}{|s|} Entropy(s_v)$$

[5]

Keterangan :
A : attribut
V : nilai mungkin untuk attribut A
Values(A): himpunan nilai yang mungkin untuk attribut A
$|S_v|$: jumlah sampel untuk nilai v
$|S|$: jumlah seluruh sampel data
Entropy(S_v) : Entropy untuk sampel yang memiliki nilai v

3. PERANCANGAN SISTEM

3.1. Dataset

Data yang digunakan merupakan data *text categorization* yang diperoleh dari *R8 of Reuters-21578 Text Categorization Collection Data Set*. Berikut ini adalah sampel data yang akan digunakan pada tugas akhir ini [14].

1. **Sampel *data training*** :
   - earn champion products ch approves stock split champion products inc said its board of directors approved a two for one stock split of its common shares for shareholders of record as of april the company also said its board voted to recommend to shareholders at the annual meeting april an increase in the authorized capital stock from five mln to mln shares reuter

2. **Sampel *data testing* :**
   - interest the unilever spokesman declined to say how much the group expected to receive for stauffer chesebrough s footwear and tennis racket businesses are also likely to be disposed of he added immediately available financial information on stauffer which is wholly owned was limited he added nine month sales to september were about billion dlrs unilever aquired chesebrough for billion dlrs in order to benefit from its well known toiletry brands and food products reuter

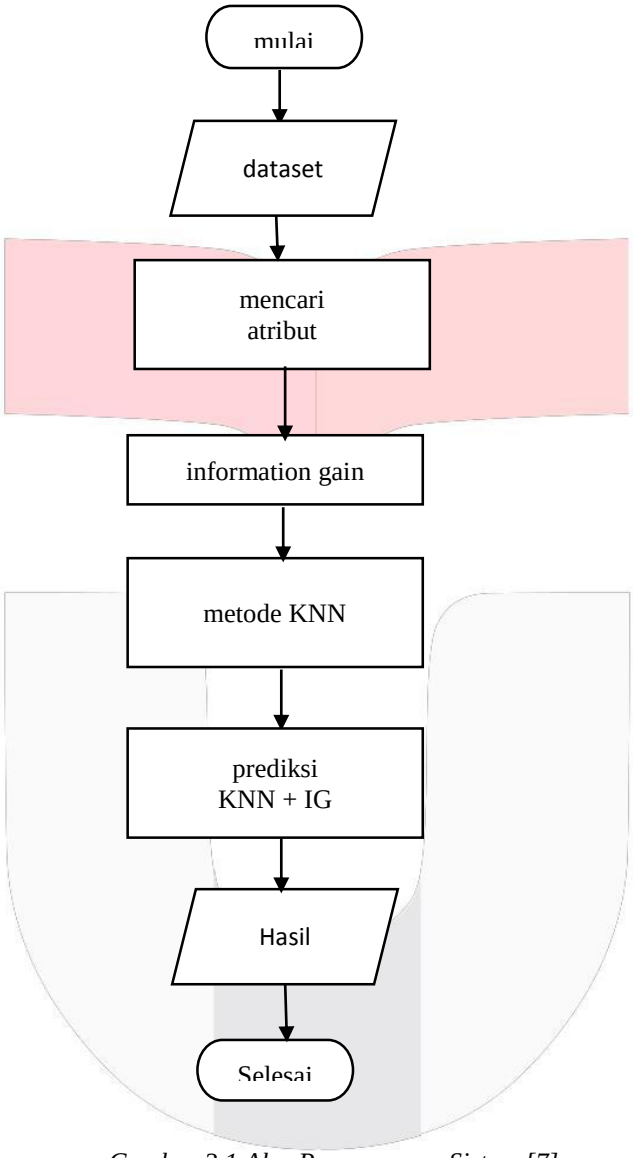
Tabel 3.1 Total Dokumen R8 of Reuters-21578

R8			
Class	s		cs
acq	1596	696	2292
crude	253	121	374
earn	2840	1083	3923
grain	41	10	51
interest	190	10	271
money-fx	206	81	293
ship	108	36	144
trade	251	75	326

Total	**5485**	**2189**	**7674**

3.2. Alur Pembuatan Model

Pada pembuatan sistem ini ada beberapa tahapan alur perancangan sistem yang harus dilakukan untuk menjamin bahwa proses pembuatan sistem sesuai dengan kebutuhan. Tahapan dari alur perancangan sistem yang akan dibangun pada sistem bisa dilihat pada gambar 3.1.



Gambar 3.1 Alur Perancangan Sistem [7]

Dari alur yang ada diatas, dapat dijelaskan bahwa untuk membangun sebuah sistem dengan kebutuhan tertentu diperlukan atau harus melakuka tahapan-tahapan untuk menjamin bahwa proses pembuatan sistem sesuai dengan kebutuhan, tahapan-tahapan tersebut terdefinisikan sesuai dengan apa yang dibutuhkan, antara lain :

1. Dataset
2. Mencari Atribut
3. Fitur Seleksi *Information Gain*
4. Metode KNN
5. Prediksi KNN dan IG
6. Hasil

4. Hasil Pengujian dan Analisis

4.1. Skenario

Skenario pengujian pada tugas akhir ini meliputi :

1. Klasifikasi KNN dengan metode pencarian nilai k pada dataset, menggunakan nilai k yang beragam (1,3,5,7).
2. Klasifikasi KNN dengan metode pencarian nilai k pada dataset, menggunakan nilai k yang beragam (1,3,5,7) dengan *featuring selection Information gain.*
3. Klasifikasi KNN dengan metode pencarian nilai gain pada dataset, menggunakan nilai gain yang beragam dengan *featuring selection Information gain.*

4.2. Pengujian Klasifikasi kNN dengan *Information Gain*

Tabel 4.1 Hasil Pengujian Klasifiakasi KNN

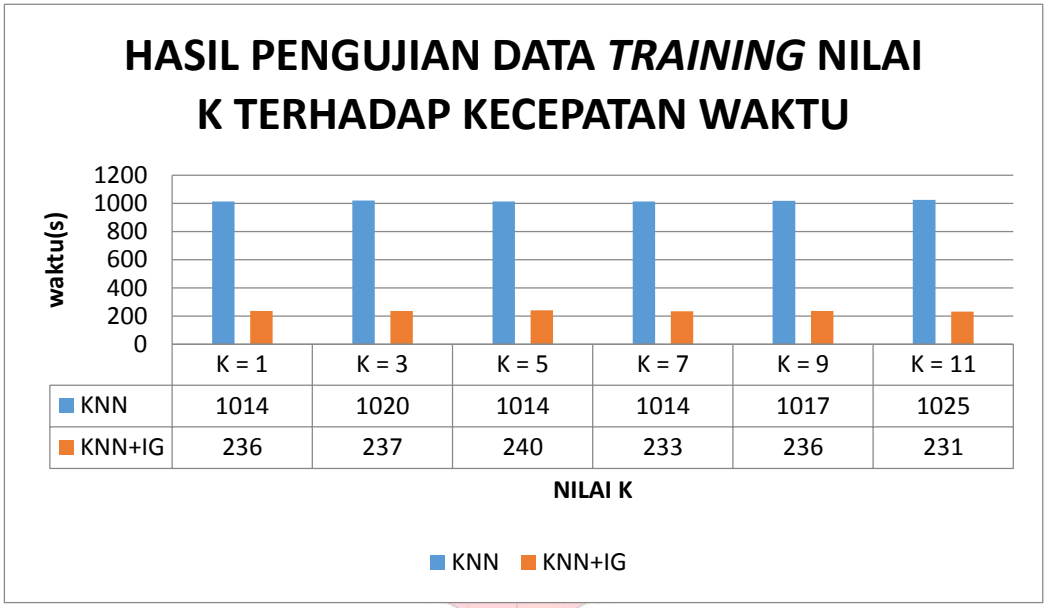
Kombinasi	Jenis Data	Nilai k	Akurasi(%)	Waktu(s)
1	Training	1	99,8889	1014
2		3	94,4444	1020
3		5	93	1014
4		7	92,333	1014
5		9	92,1111	1017
6		11	91,8889	1025
7	Testing	1	90	119
8		3	92	117
9		5	94	117
10		7	94	120
11		9	92	118
12		11	90	118

Berdasarkan tabel 4.1 semua kombinasi pada data *training* memiliki akurasi lebih tinggi dari pada data testing. Berdasarkan hasil dari seluruh percobaan diatas, maka metode KNN tanpa *Information Gain* memiliki rata-rata nilai akurasi yaitu sebesar 93,94438% pada seluruh dokumen *training* dengan berbagai parameter-parameter. Pada data hasil diatas, nilai akurasi tertinggi yaitu pada k = 1 sebesar 99,8889 %.

Pada data *training*, nilai akurasi paling tinggi saat k = 1 yaitu 99,8889%, dan nilai akurasi paling rendah saat k = 11(nilai k tertinggi) yaitu 91,8889%. Berbeda dengan data testing, nilai akurasi paling tinggi yaitu saat k = 5 dan k = 7 yaitu 94% dan nilai akurasi paling rendah saat k = 1 dan k = 11 yaitu 90%.

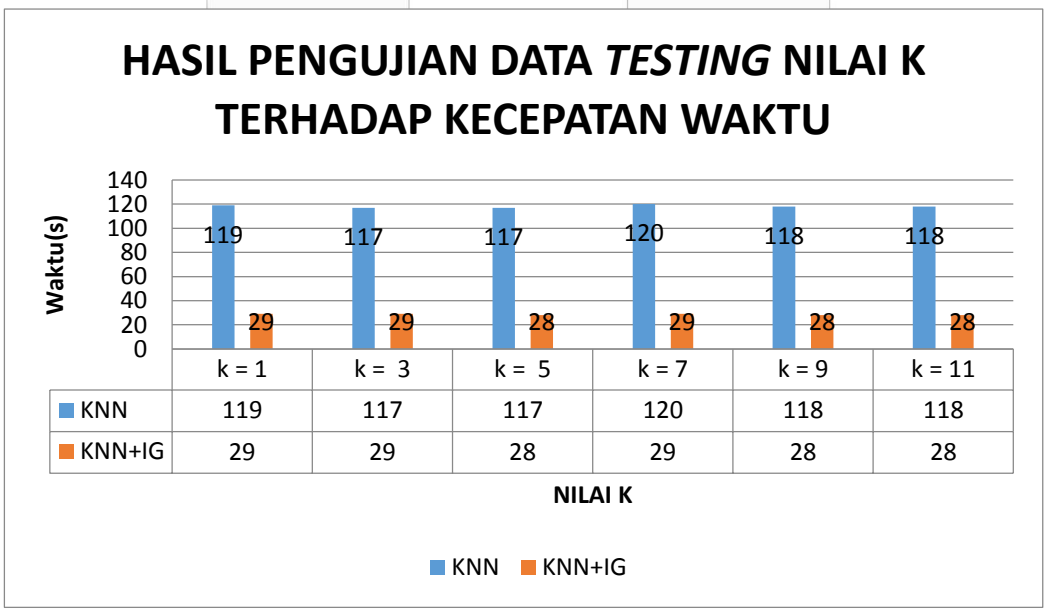
Dari hasil percobaan tabel diatas, metode klasifikai KNN terhadap nilai akurasinya sudah cukup baik, ini didukung oleh *dataset* yang bagus dan juga metode klasifikasi KNN ini sangat cocok untuk data yang *noisy* ataupun data yang jumlahnya cukup besar.

4.3 **Hasil Perbandingan Pengujian KNN dan KNN+IG Terhadap Tingkat Kecepatan Waktu**



gambar 4.3 hasil pengujian data training nilai K terhadap kecepatan waktu

Berdasarkan hasil dari seluruh percobaan diatas, maka metode KNN tanpa *Information Gain* memiliki rata-rata kecepatan waktu yaitu sebesar 1017(s) pada seluruh dokumen *training* dengan berbagai parameter-parameter. Dan dengan menggunakan kombinasi metode KNN dengan *Information Gain* memiliki rata-rata kecepatan waktu sebesar 235,5(s) pada seluruh dokumen *training* dengan berbagai parameter-parameter. Hal ini menunjukan bahwa, kecepatan waktu dalam proses klasifikasi KNN+IG jauh lebih baik dibandingkan hanya memakai KNN saja. Ini dipengaruhi oleh pereduksian atribut yang dilakukan saat *featuring selection* menggunakan *Information Gain*.



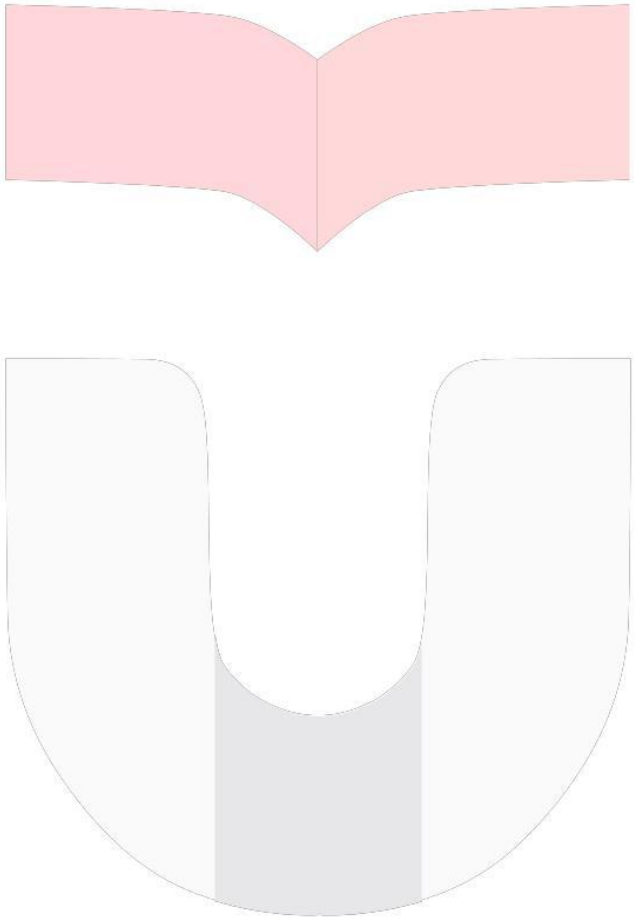
gambar 4.4 hasil pengujian data testing nilai K terhadap kecepatan waktu

Berdasarkan hasil dari seluruh percobaan diatas, maka metode KNN tanpa *Information Gain* memiliki rata-rata kecepatan waktu yaitu sebesar 118,6667(s) pada seluruh dokumen *testing* dengan berbagai parameter-parameter. Dan dengan menggunakan kombinasi metode KNN dengan *Information Gain* memiliki rata-rata kecepatan waktu sebesar 28,5(s) pada seluruh dokumen *testing* dengan berbagai parameter-parameter. Hal ini menunjukan bahwa, kecepatan waktu dalam proses klasifikasi KNN+IG jauh lebih baik dibandingkan hanya memakai KNN saja. Ini dipengaruhi oleh pereduksian atribut yang dilakukan saat *featuring selection* menggunakan *Information Gain*.

5. Kesimpulan

Dari hasil pengujian dan analisis dapat disimpulkan bahwa:

1. Berdasarkan hasil pengujian, metode KNN tanpa *Information Gain* memiliki rata-rata nilai akurasi yaitu sebesar 93,94438% pada seluruh dokumen *training* dengan berbagai parameter-parameter. Dan dengan menggunakan kombinasi metode KNN dengan *Information Gain* memiliki rata-rata nilai akurasi sebesar 93,4999% pada seluruh dokumen *training* dengan berbagai parameter-parameter.
2. Berdasarkan hasil pengujian, metode KNN tanpa *Information Gain* memiliki rata-rata nilai akurasi yaitu sebesar 92% pada seluruh dokumen *testing* dengan berbagai parameter-parameter. Dan dengan menggunakan kombinasi metode KNN dengan *Information Gain* memiliki rata-rata nilai akurasi sebesar 90,5% pada seluruh dokumen *testing* dengan berbagai parameter-parameter.
3. Dari *dataset* yang ada, terdapat 19.985 atribut. Dengan menggunakan *featuring selection Information Gain* atribut direduksi menjadi 3.185, dengan batas gain rata-rata yaitu 0,0025.
4. Penggunaan parameter yang berbeda membuat nilai akurasi dan kecepatan waktu yang dihasilkan bervariasi, namun kecepatan waktu pada kombinasi KNN+IG hasilnya jauh lebih baik.



DAFTAR PUSTAKA

[1] Avelita, B. (2016). Klasifikasi K-Nearest Neighbor. 1.

[2] Nobertus Krisandi, H. B. (2013). *Algoritma KNN dalam klasifikasi data hasil produksi kelapa sawit pada PT.MINAMAS kecamatan parindu*, 1-2.

[3] Indonesia, U. (t.thn.). Pemanfaatan dokumen-Literatur. *Klasifikasi Dokumen*, 1-11.

[4] Suyanto. (2009, November 11). *Decision Tree Learning.* Diambil kembali dari http://file.upi.edu/Direktori/FPMIPA/PRODI._ILMU_KOMPUTER/LALA/Materi_Kuliah/Kecerdasan_Buatan/9._Decision_Tree.pdf

[5] Wikipedia. (2013, Maret 18). *Algoritma K-Nearest Neighbor*. Diambil kembali dari Wikipedia: https://id.wikipedia.org/wiki/KNN

[6] Yadi, N. (2015). Tugas Akhir Data Mining. *Iterative Dichotomiser 3(ID3)*, 1-62.

[7] Mrs. Leena. H. Patil, D. M. (2014). International Jurnal of Advance Research in Artificial ntelligence. *A Multistage Feature Selection Model for Document Classification Using Information Gain and Rough Set*, 1-7.

[8] Rafael B. Pereira, A. P. (2015). Journal of Information ad Data Management. *Information Gain Featue Selection for Multi-Label Classification*, 1-11.

[9] Daniel I.Morariu, R. G. (t.thn.). *Seleksi Fitur Dokumen klasifikasi*, 1-9.

[10] UCI. (2012, 10 19). *Legal Case Report Data Set*. Diambil kembali dari UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports#

[11] Raka, R. D. (2016). *Academi.edu*. Diambil kembali dari Academi.edu: http://www.academia.edu/7448540/Praproses_data_meliputi

[12] Joko Samodra, S. S. (2009). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naive Bayes.

[13] Shweta Taneja, C. G. (t.thn.). An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering. 1-5.

[14] Reuters-21578 Text Categorization Collection Datasets for single-label text categorization, http://www.cs.umb.edu/~smimarog/textmining/datasets/

[15] Asriyanti Indah Pratiwi, Adiwijaya. 2018. On The Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis, Applied Computational Intelligence and Soft Computing 2018. Hindawi

[16] Adiwijaya. 2014. Aplikasi Matriks dan Ruang Vektor. Yogyakarta: Graha Ilmu

[17] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta

[18] Mubarok, M.S., Adiwijaya and Aldhi, M.D., 2017. Aspect-based sentiment analysis to review products using Naïve Bayes. In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060). AIP Publishing.

[19] Aziz, R.A., Mubarok, M.S. and Adiwijaya, A., 2016, September. Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. In Indonesia Symposium on Computing (IndoSC) 2016

[20] Arifin, A.H.R.Z., Mubarok, M.S. and Adiwijaya, A., 2016, September. Learning Struktur Bayesian Networks menggunakan Novel Modified Binary Differential Evolution pada Klasifikasi Data. In Indonesia Symposium on Computing (IndoSC) 2016.

[21] Yulietha, I. M., Faraby, S. A., & Adiwijaya. (2017). Klasifikasi Sentiment Review Film Menggunakan Support Vector Machine. EProceeding of Engineering 4 (3)