PENGKLASIFIKASIAN TEMA PENELITIAN BERDASARKAN ABSTRAK MENGGUNAKAN VECTOR SPACE MODEL (VSM) DAN K-NEAREST NEIGHBOR (K-NN)

Taufik Ramdani[1], Yulison Herry Chrisnanto[2], Asri Maspupah[3]
[1]Jurusan Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirma Kota Cimahi Telp 022 6656190
Email: taufikramdani32@yahoo.com, y.chrisnanto@gmail.com, asri.maspupah89@gmail.com

ABSTRAK

Pada sebuah karya ilmiah yang akan dipublikasikan, tentunya didahului dengan suatu rangkuman kegiatan penelitian yang dikenal sebagai abstrak. Abstrak secara singkat memuat tujuan penelitian, metoda yang digunakan serta hasil yang dicapai pada penelitian. Pengklasifikasian tema – tema penelitian berdasarkan abstrak dapat memudahkan pengelola dokumen hasil penelitian dalam mengelompokan penelitian berdasarkan tema yang spesifik. Pengelompokan tema dengan jumlah abstrak yang besar memiliki kesulitan tersendiri apabila dikerjakan secara konvensional. Penelitian ini menggunakan metoda Vector Space Model (VSM) dan K-Nearest Neighbor (K-NN) dalam melakukan proses pengklasifikasian tema penelitian berdasarkan abstrak. VSM digunakan untuk praproses abstrak berupa teks dan dilanjutkan menggunakan KNN untuk mengklasifikasi tema penelitian. Jumlah abstrak yang digunakan dalam penelitian ini sebanyak 75 data abstrak sebagai data training dan 25 data sebagai data uji. Jumlah kelas pada proses KNN ditentukan sebanyak 15 kelas yang merepresentasikan jumlah tema penelitian, dengan nilai K ideal ditentukan sebanyak 1, 3, 5, 7 dan 9. Berdasarkan hasil penelitian diperoleh bahwa pengklasifikasian tema penelitian dapat dilakukan dengan akurasi yang baik, dimana algoritma KNN mampu mengelompokan data abstrak ke dalam 15 kelas.

Kata kunci: ABSTRAK; Tema Penelitian; Klasifikasi; Vector Space Model; K-Nearest Neighbor.

PENDAHULUAN

Penelitian adalah pengolahan, pencarian ataupun menganalisa sebuah objek yang hendak dilakukan dengan berdasarkan teori maupun cara sistematis guna mendapatkan jawaban dari suatu masalah yang ada. Penelitian biasanya menghasilkan suatu produk atau kajian ilmu yang dapat berguna untuk masyarakat. Seseorang yang akan melakukan suatu penelitian diharuskan menentukan tema penelitian yang akan diangkat. Tema penelitian adalah pokok pokok pikiran dari penelitian. Tema dari suatu penelitian dapat beraneka ragam sesuai dengan kasus yang diangkat ke dalam sebuh penelitian. Pada suatu dokumen penelitian selalu tercantum abstrak. Abstrak adalah sebuah ringkasan isi dari sebuah karya tulis ilmiah yang ditujukan untuk membantu seorang pembaca agar dapat dengan mudah dan cepat untuk melihat tujuan dari penulisannya.

Penelitian biasanya dilakukan oleh banyak program studi di berbagai perguruan tinggi di seluruh dunia. Program studi di dalam institusi perguruan tinggi memiliki judul – judul penelitian yang bermacam – macam sesuai dengan tema penelitian dari mahasiswanya yang telah lulus. Banyaknya judul penelitian, membuat mahasiswa yang mencari judul penelitian, sesuai dengan tema yang diinginkan membuat pencarian tidak mudah, karena judul – judul penelitian belum dikelompokkan dengan temanya masing – masing. Pengklasifikasian dokumen menjadi hal yang penting untuk mengorganisasikan dokumen sehingga dapat memudahkan pencarian (Pratiwi and Widodo, 2017). *Text mining* dapat diusulkan sebagai salah satu pendekatan yang dapat dilakukan untuk pra-proses klasifikasi dokumen penelitian, karena t*ext mining* umumnya mengacu pada proses penggalian informasi dan pengetahuan yang menarik dari teks yang tidak terstruktur (Gupta and Lehal, 2009).

Klasifikasi adalah salah satu metode dalam *data mining* yang bertujuan untuk mendefinisikan kelas dari sebuah objek yang belum diketahui kelasnya (Imandoust and Bolandraftar, 2013). Masalah klasifikasi bertujuan untuk mengidentifikasi karakteristik yang menunjukkan kelompok tempat masing-masing kasus berada. Salah satu metode yang dapat digunakan dalam pengklasifikasian adalah algoritma *K-Nearest Neighbor.*

K-Nearest Neighbor adalah algoritma yang populer digunakan dalam proses pengklasifikasian teks (Yusra dkk, 2016). Beberapa penelitian menggunakan Algoritma *K-Nearest Neighbor* mengenai klasifikasi dokumen temu kembali informasi (Purwati, 2015), sentiment analysis pada teks bahasa Indonesia (Lidya dkk, 2015), dan efek penggunaan keterkaitan kata pada algoritma similaritas semantik terhadap kinerja proses klasifikasi teks (Thamrin and Sabardila, 2014).

Rumusan Masalah.

Adapun permasalahan yang muncul dari latar belakang tersebut adalah untuk menentukan tema penelitian menggunakan abstrak membutuhkan pemetaan kata, dimana perlu dilakukan proses pengamatan setipa kata dengan jumlah sebanyak 250 kata dengan jumlah abstrak sebanyak 100, serta dibandinkan dengan kata – kata penting yang berkorelasi dengan 15 kelas yang menyatakan kelompok tema penelitian dibidang informatika. Proses pemetaan kata tersebut menjadi faktor – faktor yang sulit untuk dikerjakan secara konvensional, oleh karena itu diperlukan sistem yang dapat mengotomasi proses pengelompokkan abstrak berdasarkan tema.

Tujuan Penelitian.

Tujuan penelitian ini adalah mengklasifikasikan tema penelitian berdasarkan abstrak yang belum diketahui dengan pasti kelasnya yang akan dibandingkan dengan data latih yang sudah diketahui kebenaran kelasnya menggunakan metode *K-Nearest Neighbor*. Tujuan lainnya adalah untuk mengkomputasikan penggunaan *text mining* pada abstrak, penggunaan *CF-IDF* pada tahap pembobotan dokumen abstrak dan penggunaan metode *K-Nearest Neighbor* dalam proses klasifikasi dokumen abstrak.

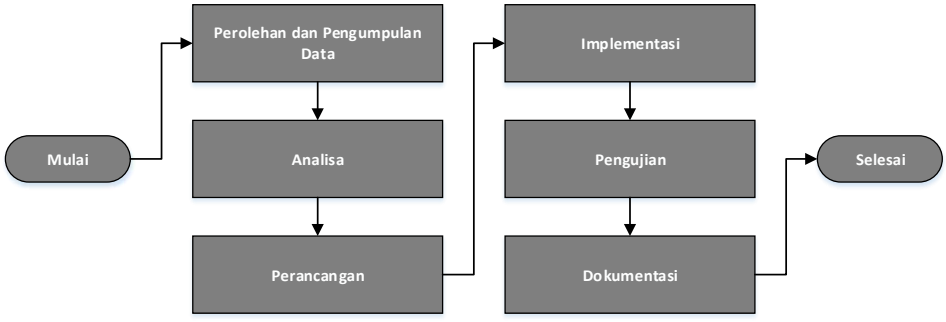
Penelitian Terdahulu.

Penggunaan abstrak untuk mengklasifikasikan dokumen karya ilmiah atau tugas akhir menjadi hal yang sangat penting untuk mengelompokkan dokumen sehingga dapat memudahkan dalam pencarian (Pratiwi and Widodo, 2017). Untuk mendapatkan informasi yang terdapat dalam dokumen terdapat suatu teknik yaitu *text mining* (Purwanti, 2015). Text mining banyak digunakan untuk menggali informasi dari sekumpulan teks menjadi suatu angka – angka statisktik yang dapat dikomputasikan (Lidya dkk, 2015). Terdapat beberapa teknik dan pengaplikasian *text mining* yang dapat dilakukan (Gupta and Lehal, 2009), salah satunya adalah pengkategorisasian atau pengklasifikasian. Klasifikasi adalah salah satu cara utntuk mengorganisasikan teks sehingga teks dengan isi atau topik yang sama akan dikelompokkan ke dalam kelas yang sama (Yusra dkk, 2016). Salah satu metode yang dapat digunakan dalam mengklasifikasikan teks atau dokumen adalah algoritma K-Nearest Neighbor (Imandoust and Bolandraftar, 2013), tetapi sebelum masuk ke dalam proses algoritma K-Nearest Neighbor terdapat pra – proses. Pra – proses tersebut antara lain *case folding, tokenizing, filtering*, dan *stemming* (Thamrin and Sabardila, 2014). Setelah melalui pra-proses selanjutnya masuk ke dalah pembobotan kata (Aditya, 2016) dan pada akhirnya masuk dalam proses K-Nearest Neigbor untuk mendapatkan hasil akhir dari dokumen yag diuji coba.

Pembahasan

METODE PENELITIAN.

Dalam melakukan sebuah penelitian, dibutuhkan suatu acuan pelaksanaan yang dinamakan metodologi penelitian. Metodologi penelitian terdiri dari beberapa tahapan kerangka kerja penelitian secara terstruktur / sistematis mulai dari tahap awal penelitian hingga mendapatkan hasil yang ingin dicapai. Berikut adalah gambaran tahapan yang dilakukan dalam penelitian yang akan disajikan pada Gambar 1 berikut.



Gambar 2. METODE PENELITIAN.

Perolehan dan Pegumpulan Data.

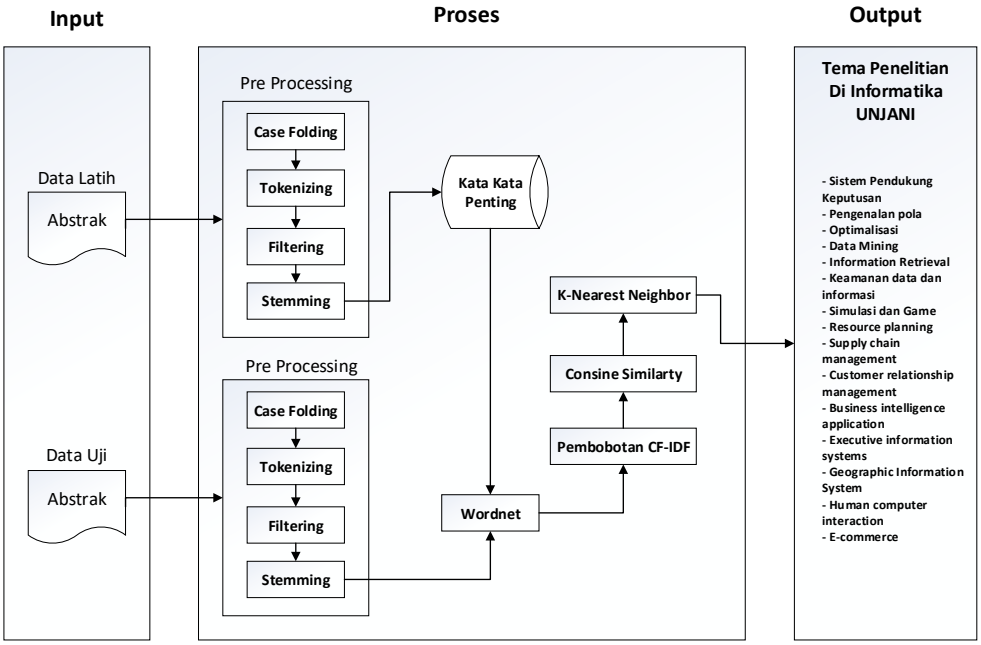
Data yang digunakan berasala dari data abstrak laporan tugas akhir mahasiswa jurusan Informatika Universitas Jenderal Achmad Yani, dan beberapa data abstrak didapatkan dari Google Scholar. Data yang digunakan sebanyak 100 data abstrak untuk dijadikan sebagai data training dan data testing.

Tabel 30. Sample Data Yang Digunakan.

No	ABSTRAK	Kelas/Tema
1.	Penelitian ini bertujuan membangun sistem informasi pengujian kendaraan bermotor menggunakan Business Intelligence untuk memudahkan pegawai dan KA UPTD di Dishub Kabupaten Bandung Barat khususnya pada data analisis kendaraan dan data analisis pemohon dengan menggunakan kategori untuk membuat hasil analisis dalam mengimplementasikan Dashboard Business Intelligence.	Business Intelligence
2.	Konsep CRM (Customer Relationship Management) diterapkan untuk meningkatkan layanan pelanggan, dimana pelanggan ditempatkan sebagai pusat proses. Di bidang pendidikan CRM dapat diterapkan seperti sebuah perusahaan. Misalnya dalam pendidikan tinggi di mana siswa adalah pusat dari proses CRM itu sendiri. Perguruan tinggi dapat mengambil manfaat dari CRM dengan meningkatkan layanan mahasiswa, personalisasi komunikasi dengan mahasiswa, berbagi informasi antar departemen, dan meningkatkan retensi dan kepuasan siswa	Customers Relationship Management
…	….	….
99.	Electronic Commerce (e-commerce) adalah proses pembelian, penjualan atau pertukaran Barang, jasa dan informasi melalui jaringan komputer tingkat penggunaan teknologi semakin lama semakin meningkat. Penerapan media e-commerce dilakukan untuk akomodasi penjualan dan mempromosikan barang dapat lebih menguntungkan, tipe e-commerce ini yaitu lebih ke marketplace	E-Commerce
100.	Sistem informasi eksekutif ini digunakan untuk membantu jajaran eksekutif dalam melihat potensi wilayah mana saja yang banyak menggunakan produk dari BTN serta menghasilkan laporan yang dapat dijadikan sebagai bahan kajian kepala cabang dalam pengambilan keputusan. Sistem informasi eksekutif ini dibangun dengan menggunakan metode pengembangan perangkat lunak waterfall dan keluaran berupa laporan yang dibutuhkan dalam bentuk grafik agar membantu kepala cabang dalam pengambilan keputusan.	Sistem Informasi Eksekutif

Analisa dan Perancangan Sistem.

Sistem ini merupakann sistem yang dapat mengklasifikasikan sebuah dokumen abstrak. Terdapat dua proses pada sistem ini yaitu proses data latih dan data latih. Dimana nantinya akan data uji dengan data latih yang ada, sehingga menghasilkan rekomendasi informasi. Gambaran umum dari sistem yang akan dibuat dapat dilihat pada Gambar 2.



Gambar 3 Perancangan Sistem.

Keterangan :

Dalam perancangan sistem ini terdapat 3 tahapan, yaitu :

1. Input. Terdapat dua data masukan yaitu data latih dan data uji. Data tersebut selanjutnya diproses pada tahap *preprocessing*.
2. Proses. Dalam tahap ini, terdapat beberapa langkah yang dilakukan, yaitu sebagai berikut :
   a. Tahap *Pre-processing.* Pada tahap ini dokumen melalui tahap *text mining* yaitu proses *case folding, tokenizing, filtering, dan stemming*.
   - *Case folding* adalah tahap dimana mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.
   - *Tokenizing* adalah tahap pemotongan kalimat inputan menjadi kata perkata pada setiap kata yang menyusunnya.
   - *Filtering* adalah tahap mengambil kata – kata penting dari hasil token menggunakan algoritma *stoplist* (membuang kata yang kurang penting).
   - *Stemming* adalah tahap untuk mendapatkan kata dasar dari kata yang telah mendapatkan imbuhan atau keterangan lainnya. Hasil dari *stemming* akan disimpan pada *database* kata-kata penting yang mewakili setiap dokumen data latih setiap kelas.

   b. Tahap *Processing.* Setelah melalui proses pada *preprocessing* yang memfilter dokumen berupa *text*, selanjutnya adalah tahap *processing*, pada tahap ini menghitung CF, menghitung DF dan meghitung hasil bobot dari suatu dokumen. Kemudian setelah mendapatkan bobot setiap dokumen data uji dan data latih masuk proses *vector space model* mencari sudut antara dua *vector* dengan menghitung dua sudut antara bobot suatu dokumen uji dan dokumen banding / bobot kata kunci menggunakan VSM dengan pendekatan *cosine similarty*. Kemudian setelah mendapatkan hasil dari proses cosine similarity, masuk pada proses k-nearest neighbor untuk mendapatkan hasil akhir dengan mencari nilai k optimal sebanyak 1, 3,5, 7 dan 9.
3. Output. Dalam tahap ini hasil akhir sudah diketahui, bahwa dari dokumen *input* mendekati kelas yang mana dari 15 kelas yang ada.

Implementasi.

Text Mining.

Text mining adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah. Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi dan asosiasi. Tahapan *preprocessingi* (Priliyanti dan Wijaya, 2014) dalam *text mining* secara umum adalah *case folding*, *tokenizing*, *filtering*, dan *stemming.*

Concept Frequency-Inverse Document Frequency (CF-IDF).

Algoritma CF-IDF adalah algoritma perhitungan bobot kesesuaian dokumen Perhitungan pada metode ini tidak melakukan perhitungan terhadap *term* (seperti pada TF-IDF) tetapi dengan menghitung *key concept* yang ditemukan didalam pesan. Pada CF-IDF, dilakukan pendekatan representasi isi dokumen dengan menggunakan jaringan semantik yang disebut dokumen inti semantik. Dokumen tersebut kemudian dipetakan dalam jaringan semantik yang disebut *WordNet* dan dikonversikan dari sekumpulan *terms* menjadi sekumpulan konsep (Aditya, 2016).

Algoritma CF-IDF dipilih pada penelitian ini karena CF-IDF lebih kepada pendekatan representasi isi dari dokumen. Berbeda dengan algoritma TF-IDF yang hanya memberikan bobot hubungan suatu kata terhadap dokumen.

Pada tahap pertama dalam metode CF-IDF yaitu melakukan pembobotan dengan menghitung CF (Concept Frequency):

$$cf_{i,j} = \frac{n_{i,j}}{\sum n_{i,j}} \qquad (1)$$

Dimana :

$cf_{i,j}$ = rasio frekuensi concept pada dokumen
$n_{i,j}$ = jumlah kemunculan concept dalam dokumen
$\sum n_{i,j}$ = total kemunculan seluruh concept dalam dokumen
i = menunjukkan concept yang keberapa
j = menunjukkan jumlah dokumen

Setelah itu, dilakukan perhitungan nilai IDF dengan membagi jumlah total dokumen dengan jumlah dokumen yang terdapat kemunculan konsep (Ci).

$$idf_i = log \frac{[D]}{[\{d:ci \in d\}]} \qquad (2)$$

Dimana :

idf = rasio frekuensi dokumen
$[D]$ = jumlah total dokumen
$[\{d:ci \in d\}]$ = jumlah dokumen yang terdapat kemunculan concept

Pada tahap terakhir nilai CF dikalikan dengan IDF.

$$W = cf_i * idf_i \qquad (3)$$

Dimana :

W = bobot CF-IDF
cf_i = rasio frekuensi concept pada dokumen
idf_i = rasio frekuensi dokumen

Cosine Similarity.

Cosine similarity adalah perhitungan kesamaan antara dua vektor n dimensi dengan mencari kosinus dari sudut diantara keduanya dan sering digunakan untuk membandingkan dokumen dalam text mining. Rumus Cosine similarity dapat dilihat pada persamaan 4 :

$$similarity\ (x, y) = \cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (4)$$

Dimana :

$\sum_{i=1}^{n} x_i y_i$ = jumlah bobot kata dokumen x terhadap dokumen y.
$\sqrt{\sum_{i=1}^{n} x_i^2}$ = akar dari jumlah bobot dokumen x.
$\sqrt{\sum_{i=1}^{n} y_i^2}$ = akar dari jumlah bobot dokumen y.

K - Nearest Neighbor.

K-Nearest Neighbor adalah salah satu algoritma yang paling populer untuk kategorisasi teks. Untuk mengklasifikasikan dokumen baru, sistem menemukan tetangga terdekat "k" di antara dokumen pelatihan, dan menggunakan kategori tetangga terdekat untuk memberi bobot pada kandidat kategori. Salah satu kekurangan algoritma K-NN adalah efisiensinya, karena perlu membandingkan dokumen uji dengan semua sampel dalam rangkaian pelatihan. Selain itu, kinerja algoritma ini sangat bergantung pada dua faktor, yaitu fungsi kesamaan yang sesuai dan nilai parameter "k" yang sesuai (Baoli dkk, 2003). Prinsip kerja dari *K-Nearest Neighbor* adalah mencari jarak antara dua titik yaitu titik *training* dan titik *testing*. Ada banyak cara untuk mengukur jarak kedekatan antara data *testing* dengan data *training*, diantaranya *euclidean distance* (Leidiyana), *manhattan distance* (*city block distance*), dan *cosine similarity.* Dan pada penelitian ini digunakan pendekatan *cosine similarity.*\

Pra-Proses.

Sebagai contoh data uji terdapat suatu teks abstrak yang belum diketahui kelasnya dengan kode "**DU**", yaitu: *"Data Mining merupakan proses ekstraksi data menjadi informasi yang sebelumnya belum tersampaikan, dengan teknik yang tepat proses data mining akan memberikan hasil yang optimal. Data mining lebih tepat disebut sebagai penambangan pengetahuan dari data. Clustering merupakan salah satu teknik dalam data mining untuk mengelompokkan data berdasarkan kriteria."*

Case Folding.

Data uji tersebut akan masuk dalam tahap pra-proses *case folding*, dan hasilnya yaitu: "*data mining merupakan proses ekstraksi data menjadi informasi yang sebelumnya belum tersampaikan dengan teknik yang tepat proses data mining akan memberikan hasil yang optimal data mining lebih tepat disebut sebagai penambangan pengetahuan dari data clustering merupakan salah satu teknik dalam data mining untuk mengelompokkan data berdasarkan kriteria*".

Tokenizing.

Selanjutnya masuk tahap *tokenzing* untuk memisahkan kata demi kata, hasilnya menjadi seperti: *"-data-mining-merupakan-proses-ekstraksi-data-menjadi-informasi-yang-sebelumnya-belum-tersampaikan-dengan-teknik-yang-tepat-proses-data-mining-akan-memberikan-hasil-yang-optimal-data-mining-lebih-*

tepat-disebut-sebagai-penambangan-pengetahuan-dari-data-clustering-merupakan-salah-satu-teknik-dalam-data-mining-untuk-mengelompokkan-data-berdasarkan-kriteria".

Filtering.
Selanjutnya masuk tahap *filtering* untuk membuang kata yang kurang penting, hasilnya menjadi seperti: *"-data-mining-merupakan-proses-ekstraksi-data-menjadi-informasi-belum-tersampaikan-teknik-tepat-proses-data-mining-memberikan-hasil-optimal-data-mining-penambangan-pengetahuan-data-clustering-merupakan-teknik-dalam-data-mining-mengelompokkan-data-berdasarkan-kriteria*".

Stemming.
Selanjutnya masuk tahap *stemming* untuk mendapatkan kata dasar dari kata yang berimbuhan: *"-data-mining-rupa-proses-ekstraksi-data-jadi-informasi-belum-sampai-teknik-tepat-proses-data-mining-beri-hasil-optimal-data-mining-tambang-pengetahuan-data-clustering-rupa-teknik-dalam-data-mining-kelompok-data-dasar-kriteria*".

Pemetaan Data Uji Dalam Wordnet Data Latih.
Selanjutnya hasil pra-proses data uji tersebut akan dipetakan ke dalam kumpulan wordnet data latih. Pada uji coba ini digunakan dokumen data latih sebanyak 20 dokumen dengan kelas yang berbeda beda. Karena keterbatasan halaman maka hasil dari pemetaan tesebut akan ditulis secara singkat, berikut hasilnya pada Tabel 2.

Tabel 31. Pemetaan Wordnet.

Concept	Frekuensi												DF
	DU	D1	D2	D3	D4	D5	….	D16	D17	D18	D19	D20	
data	7	2		1			….	5					7
mining	4			2			….			1			2
rupa	2					1	….	1					7
proses	2		2	6	1		….		1	1		2	7
ekstraksi	1						….						0
jadi	1						….						0
informasi	1	1	1		1	2	….	1					7
belum	1						….						2
….	….	….	….	….	….	….	….	….	….	….	….	….	….
kasus							….					2	1
terdahulu							….					1	1
gejala							….					2	1
mirip							….					1	1
retrive							….					1	1
Total Jumlah Concept per Dokumen	33	28	42	38	27	39	….	38	38	35	45	37	

Pembobotan CF dan IDF.
Setelah kata dipetakan dalam wordnet selanjutnya dihitung bobot setiap kata, dengan menggunakan persamaan 1 untuk mencari bobot CF dan persamaa 2 untuk mencari bobot IDF, berikut hasilnya pada Tabel 3.

Tabel 32. Pembobotan CF dan IDF.

Concept	Concept Frekuensi												IDF
	DU	D1	D2	D3	D4	D5	….	D16	D17	D18	D19	D20	

data	0.212	0.071	0.000	0.026	0.000	0.000	….	0.132	0.000	0.000	0.000	0.000	0.456
mining	0.121	0.000	0.000	0.053	0.000	0.000	….	0.000	0.000	0.029	0.000	0.000	1.000
rupa	0.061	0.000	0.000	0.000	0.000	0.026	….	0.026	0.000	0.000	0.000	0.000	0.456
proses	0.061	0.000	0.048	0.158	0.037	0.000	….	0.000	0.026	0.029	0.000	0.054	0.456
ekstraksi	0.030	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.000	0.000
jadi	0.030	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.000	0.000
informasi	0.030	0.036	0.024	0.000	0.037	0.051	….	0.026	0.000	0.000	0.000	0.000	0.456
belum	0.030	0.000	0.000	0.000	0.000	0.000	….		0.000	0.000	0.000	0.000	1.000
….	….	….	….	….	….	….	….	….	….	….	….	….	….
kasus	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.054	1.301
terdahulu	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.027	1.301
gejala	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.054	1.301
mirip	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.027	1.301
retrive	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.027	1.301

Pembobotan CF-IDF.

Setelah mendapatkan bobot cf dan idf, maka selanjutnya mencari bobot cf-idf dengan menggunakan persamaan 3, berikut hasilnya pada Tabel 4.

Tabel 33 Pembobotan CF-IDF.

Concept	CF-IDF											
	DU	**D1**	**D2**	**D3**	**D4**	**D5**	**….**	**D16**	**D17**	**D18**	**D19**	**D20**
data	0.096713	0.032567	0	0.011998	0	0	….	0.059991	0	0	0	0
mining	0.121212	0	0	0.052632	0	0	….	0	0	0.028571	0	0
rupa	0.027632	0	0	0	0	0.011691	….	0.011998	0	0	0	0
proses	0.027632	0	0.021711	0.072	0.016886	0	….	0	0.011998	0.013027	0	0.024645
ekstraksi	0	0	0	0	0	0	….	0	0	0	0	0
jadi	0	0	0	0	0	0	….	0	0	0	0	0
informasi	0.013816	0.016283	0.010856	0	0.016886	0.023381	….	0.011998	0	0	0	0
belum	0.030303	0	0	0	0	0	….	0	0	0	0	0
….	….	….	….	….	….	….	….	….	….	….	….	….
kasus	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.070326
terdahulu	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.035163
gejala	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.070326
mirip	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.035163
retrive	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.035163

Vector Space Model.

Setelah mendapatkan bobot cf-idf, selanjutnya masuk dalam proses Vector Space Model untuk mencari sudut antara dua *vector* dengan menghitung dua sudut antara bobot suatu dokumen uji dan dokumen banding. Berikut hasilnya pada Tabel 5 dan Tabel 6.

Tabel 34. Vector Space Model.

Concept	Bobot CF-IDF ^ 2

	DU	D1	D2	D3	D4	D5	….	D16	D17	D18	D19	D20
data	0.00935	0.0011	0	0.00014	0	0	….	0.00360	0	0	0	0
mining	0.01469	0	0	0.00277	0	0	….	0	0	0.00082	0	0
rupa	0.00076	0	0	0	0	0.00014	….	0.00014	0	0	0	0
proses	0.00076	0	0.000471	0.00518	0.00029	0	….	0	0.00014	0.00017	0	0.00061
ekstraksi	0	0	0	0	0	0	….	0	0	0	0	0
jadi	0	0	0	0	0	0	….	0	0	0	0	0
informasi	0.00019	0.0003	0.000118	0	0.00029	0.00055	….	0.00014	0	0	0	0
belum	0.00092	0	0	0	0	0	….	0	0	0	0	0
….	….	….	….	….	….	….	….	….	….	….	….	….
kasus	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.00495
terdahulu	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.00124
gejala	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.00495
mirip	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.00124
retrive	0.000	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.00124
Akar Jumlah Bobot Per Dokumen $\sqrt{\sum_{i=1}^{n} x_i^2}$	0.198619	0.256624	0.228223	0.203192	0.243684	0.207869	….	0.217535	0.218135	0.243946	0.233627	0.213093

Tabel 35. Vector Space Model (2).

Concept	DU*D1										
	DU*D 1	DU*D 2	DU*D 3	DU*D 4	DU*D 5	….	DU*D 16	DU*D 17	DU*D 18	DU*D 19	DU*D 20
data	0.00000992	0	0.00000135	0	0	….	0.00003366	0	0	0	0
mining	0	0	0.00004070	0	0	….	0	0	0.00001199	0	0
rupa	0	0	0	0	0.00000010	….	0.00000011	0	0	0	0
proses	0	0.00000036	0.00000396	0.00000022	0	….	0	0.00000011	0.00000013	0	0.00000046
ekstraksi	0	0	0	0	0	….	0	0	0	0	0
jadi	0	0	0	0	0	….	0	0	0	0	0
informasi	0.00000005	0.00000002	0	0.00000005	0.00000010	….	0.00000003	0	0	0	0
belum	0	0	0	0	0	….	0	0	0	0	0
….	….	….	….	….	….	….	….	….	….	….	….
kasus	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.000
terdahulu	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.000
gejala	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.000
mirip	0.000	0.000	0.000	0.000	0.000	….	0.000	0.000	0.000	0.000	0.000

retrive	0.000	0.000	0.000	0.000	0.000	⋯	0.000	0.000	0.000	0.000	0.000
Jumlah Bobot Per Dokumen $\sum_{i=1}^{n} x_i y_i$	0.0000 1012	0.0000 0038	0.0000 4600	0.0000 0027	0.0000 0029	⋯	0.0000 3562	0.0000 0011	0.0000 2081	0.0000 0000	0.0000 0046

K – Nearest Neighbor Menggunakan Cosine Similarity.

Selanjutnya adalah mencari nilai kedekatan antar dokumen uji dengan dokumen data latih dengan menggunakan rumus *cosine similarity* (persamaan 4), berikut hasilnya pada tabel 7.

Tabel 36. Hasil Cosine Similarity.

No	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Dikalikan 100%
1.	D1	0.00000784	0.000784%
2.	D2	0.00000033	0.000033%
3.	D3	0.00004497	0.004497%
4.	D4	0.00000022	0.000022%
5.	D5	0.00000028	0.000028%
6.	D6	0.00000150	0.000150%
7.	D7	0.00000023	0.000023%
8.	D8	0.00000020	0.000020%
9.	D9	0.00002904	0.002904%
10.	D10	0.00000221	0.000221%
11.	D11	0.00000002	0.000002%
12.	D12	0.00000167	0.000167%
13.	D13	0.00000550	0.000550%
14.	D14	0.00000472	0.000472%
15.	D15	0.00000229	0.000229%
16.	D16	0.00003252	0.003252%
17.	D17	0.00000010	0.000010%
18.	D18	0.00001695	0.001695%
19.	D19	0.00000000	0.000000%
20.	D20	0.00000043	0.000043%

Perangkingan Dengan Nilai K 1, 3, 5, 7, dan 9.

Untuk mengetahui hasil akhir dari dokumen uji tersebut selanjutnya adalah mencari dokumen yang paling dekat dengan data uji dengan nilai K 1, 3, dan 5, berikut hasilnya pada Tabel 8.

Tabel 37. Perangkingan.

Ranking	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Kelas
1.	D3	0.004497%	Data Mining
2.	D16	0.003252%	Sistem Informasi Geografi
3.	D9	0.002904%	Kemanan Data dan Informasi
4.	D18	0.001695%	Data Mining
5.	D1	0.000784%	Business Intelligence
6.	D13	0.000550%	Simulasi dan Game
7.	D14	0.000472%	Sistem Pendukung Keputusan
8.	D15	0.000229%	Supply Chain Management
9.	D10	0.000221%	Optimalisasi
10.	D12	0.000167%	Enterprise Resource Planning
11.	D6	0.000150%	Sistem Informasi Geografi
12.	D20	0.000043%	Information Retrieval
13.	D2	0.000033%	Customers Relationship Management
14.	D5	0.000028%	Sistem Informasi Eksekutif
15.	D7	0.000023%	Human Computer Interface
16.	D4	0.000022%	E-Commerce
17.	D8	0.000020%	Information Retrieval
18.	D17	0.000010%	E-Commerce
19.	D11	0.000002%	Pengenalan Pola
20.	D19	0.000000%	Pengenalan Pola

Tabel 38. Nilai K = 1

Ranking	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Kelas
1.	**D3**	**0.004497%**	**Data Mining**

Tabel 39. Nilai K = 3

Ranking	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Kelas
1.	D3	0.004497%	Data Mining
2.	D16	0.003252%	Sistem Informasi Geografi
3.	D9	0.002904%	Kemanan Data dan Informasi

Tabel 40. Nilai K = 5

Ranking	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Kelas
1.	**D3**	**0.004497%**	**Data Mining**
2.	D16	0.003252%	Sistem Informasi Geografi
3.	D9	0.002904%	Kemanan Data dan Informasi
4.	**D18**	**0.001695%**	**Data Mining**
5.	D1	0.000784%	Business Intelligence

Tabel 41. Nilai K = 7

Ranking	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Kelas
1.	**D3**	**0.004497%**	**Data Mining**
2.	D16	0.003252%	Sistem Informasi Geografi
3.	D9	0.002904%	Kemanan Data dan Informasi
4.	**D18**	**0.001695%**	**Data Mining**
5.	D1	0.000784%	Business Intelligence
6.	D13	0.000550%	Simulasi dan Game
7.	D14	0.000472%	Sistem Pendukung Keputusan

Tabel 42. Nilai K = 9

Ranking	Dokumen Latih	Nilai Kedekatan Dengan Dokumen Uji	Kelas
1.	**D3**	**0.004497%**	**Data Mining**
2.	D16	0.003252%	Sistem Informasi Geografi
3.	D9	0.002904%	Kemanan Data dan Informasi
4.	**D18**	**0.001695%**	**Data Mining**
5.	D1	0.000784%	Business Intelligence
6.	D13	0.000550%	Simulasi dan Game
7.	D14	0.000472%	Sistem Pendukung Keputusan
8.	D15	0.000229%	Supply Chain Management
9.	D10	0.000221%	Optimalisasi

KESIMPULAN.

Berdasarkan data uji berupa abstrak, dengan menerapkan teknik VSM dimana abstrak berupa teks diubah bentuk menjadi sekumpulan angka lalu dihitung nilai kedekatan setiap kelas dengan menggunakan teknik K-NN serta menguji jumlah kedekatan dengan nilai K yang beragam diperoleh kelas terbanyak pada setiap pengujian menunjukkan bahwa abstrak yang diuji dapat diidentifikasi sebagai kelompok kelas tertentu dalam hal ini abstrak yang digunakan sebagai data uji pada penelitian ini mengarah pada kelas Data Mining.

DAFTAR PUSTAKA.

N. I. Pratiwi and Widodo, "Klasifikasi Dokumen Karya Akhir Mahasiswa Menggunakan Naïve Bayes Classifier (NBC) Berdasarkan Abstrak Karya Akhir Di Jurusan Teknik Elektro Universitas Negeri Jakarta," *Jurnal Pinter*, vol. I, no. 1, pp. 33-40, 2017.

V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal Of Emerging Technologies In Web Intelligence*, vol. I, no. 1, pp. 60-76, 2009.

S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," *Journal of Engineering Research and Applications*, vol. III, no. 5, pp. 605-610, 2013.

Yusra, D. Olivita and Y. Fitriani, "Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor," *Jurnal Sains, Teknologi dan Industri*, vol. XIV, no. 1, pp. 79-85, 2016.

E. Purwanti, "Klasifikafi Dokumen Temu Kembali Informasi dengan K-Nearest Neghbour," *Record and Library*, vol. I, no. 2, pp. 129-138, 2015.

S. K. Lidya, O. S. Sitompul and S. Efendi, "Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN)," *Seminar Nasional Teknologi Informasi dan Komunikasi*, pp. 1-8, 2015.

H. Thamrin and A. Sabardila, "Efek Penggunaan Keterkaitan Kata pada Algoritma Similaritas Semantik Terhadap Kinerja Proses Klasifikasi Teks dengan K-Nearest Neighbour," *Komuniti*, vol. VI, no. 2, pp. 104-110, 2014.

O. Sumantri, S. Wiyono and Dairoh, "Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM)," *Scientific Journal of Informatics*, vol. III, no. 1, pp. 34-45, 2016.

K. R. Prilianti and H. Wijaya, "Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering," *Cybermatika*, vol. II, no. 1, pp. 1-6, 2014.

Z. Yong, L. Youwen and X. Shixiong, "An Improved KNN Text Classification Algorithm Based on Clustering," *Journal Of Computers*, vol. IV, no. 3, pp. 230-237, 2009.

C. Darujati and A. B. Gumelar, "Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia," *Jurnal Link*, vol. VI, no. 1, pp. 51-58, 2012.

A. Rakhmatullah, "Penerapan Knowledge Management System Di Dinas Pertanian Cianjur Menggunakan CF-IDF Dan Vector Space Model," *Seminar Nasional Teknologi Informasi dan Komunikasi*, pp. 624-631, 2016.

L. Baoli, Y. Shiwen and L. Qin, "An Improved k-Nearest Neighbor Algorithm for Text Categorization," *International Conference on Computer Processing of Oriental Languages*, 2003.

H. Leidiyana, "Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor," *Jurnal Penelitian Ilmu Komputer*, vol. I, no. 1, pp. 65-76, 2013.

K. Khamar, "Short Text Classification Using kNN Based on Distance Function," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. II, no. 4, pp. 1916-1919, 2013.