Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

PENGKLASIFIKASIAN DOKUMEN PDF MENGGUNAKAN FUNGSI COSINE UNTUK PENCARIAN INFORMASI

Classification of PDF Document using Cosine Functions for Information Search

Jeremy Jonathan[1)]

[1)]Magister Ilmu Komputer/Program Pascasarjana, Budi Luhur University

Diterima 14 December 2017 / Disetujui 26 January 2018

ABSTRACT

In the classification of document information we can use information retrieval system (Information Retrieval). Information Retrieval (IR) is the process by which data collection is represented, stored, and searches for the purpose of knowledge discovery in response to query requests. Pouring information in text form for a review material or review can be poured in different styles of review, due to the different styles of review for a review or topic on the document it takes longer to understand the material and classify it. This makes it difficult for system developers to create a system that is able to properly understand the content or topic of a document. To solve the problem, the writer uses one of the information retrieval model that is vector space model with K-NN algorithm and Stemming nazief adriani with weighting of word frequency () in document and cosine equality function. The results of effectiveness obtained in this study in terms of reviewing the effectiveness of the use of word frequency weighting () in the document on the vector space model with similarity cosine function. From the test results obtained value of K = 5 is 93.33% for precision and 54.67% for recall. And the value of K = 3 is 96.33% for precision and 66.67% for recall.

Keywords: *Information Retrieval System, Vector Space Model, Stemming Nazief Andriani, Fungsi Cosine, K-Nearest Neighbor, Precision, Recall.*

ABSTRAK

Dalam pengklasifikasian informasi dokumen kita dapat menggunakan sistem temu kembali informasi *(Information Retrieval). Information Retrieval (IR)* adalah proses dimana pengumpulan data diwakili, disimpan, dan mencari tujuan penemuan pengetahuan sebagai tanggapan atas permintaan pengguna (*query*). Penuangan informasi dalam bentuk teks untuk sebuah materi atau ulasan yang dibahas dapat dituangkan dalam berbagai bentuk gaya ulasan yang berbeda, dikarenakan adanya perbedaan gaya mengulas untuk suatu ulasan atau topik pada dokumen maka membutuhkan waktu relatif lebih lama untuk memahami materi dokumen tersebut serta mengklasifikannya. Hal ini menyebabkan sulitnya bagi pengembang sistem untuk membuat sebuah sistem yang mampu memahami isi materi atau topik dari sebuah dokumen secara tepat. Untuk memecahkan masalah tersebut penulis menggunakan salah satu model *information retrieval* yaitu *vector space model* dengan algoritma *K-NN* dan Stemming nazief adriani dengan pembobotan frekuensi kata ($log f_{in}$) dalam dokumen serta fungsi kesamaan *cosine*. Hasil efektifitas yang didapat pada penelitian ini dalam hal meninjau efektifitas penggunaan pembobotan frekuensi kata ($log f_{in}$) dalam dokumen pada *vector space model* dengan fungsi *similarity cosine.* Dari hasil pengujian tersebut didapat nilai K=5 adalah 93.33% untuk *precision* dan 54.67% untuk *recall*. Dan nilai K=3 adalah 96.33% untuk *precision* dan 66.67% untuk *recall*.

Kata Kunci: *Information Retrieval System, Vector Space Model, Stemming Nazief Andriani, Fungsi Cosine, K-Nearest Neighbor, Precision, Recall.*

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

PENDAHULUAN

Beberapa tahun belakangan ini teknologi informasi berkembang dengan pesatnya, yang dapat memudahkan masyarakat banyak dalam hal mendapatkan segala informasi dari waktu yang lama sampai dengan waktu yang terbaru dengan estimasi waktu mendapatkannya yang lebih cepat daripada sebelumnya. Perkembangan teknologi tersebut memicu terjadinya perubahan suatu kebiasaan masyarakat dalam mengakses informasi yang sebelumnya mendapatkan informasi hanya melalui sesuatu yang didengar dari orang lain dan juga melalui media cetak seperti koran dan majalah, kini masyarakat dapat mendapatkan informasi tersebut melalui situs portal berita yang dapat diakses melalui *personal computer* (PC) atau melalui gadget (*Smartphone* atau tablet). Informasi yang beredar saat ini berjumlah sangat banyak dan sangat berguna bagi setiap individu masyarakat di dalamnya, seperti informasi pendidikan, olahraga, dan keperluan lainnya. Sebagai contohnya di dalam dunia jurnalis setiap informasi yang didapat dapat simpan dalam sebuah media dalam bentuk teks. Dokumen dalam teks merupakan bentukan data yang tidak terstruktur. Penuangan informasi dalam bentuk teks untuk sebuah materi pembahasan (topik) dapat dituangkan dalam berbagai bentuk gaya ulasan yang berbeda dalam suatu bahasa tertentu. Gaya mengulas yang menuangkan dalam sekumpulan kalimat-kalimat dalam sebuah teks sangat dipengaruhi oleh pengetahuan dari orang tersebut. Dengan adanya gaya mengulas suatu topik ulasan atau topik pada sejumlah dokumen menyebabkan sulitnya bagi seorang pengembang sistem untuk membangun sebuah sistem yang dapat mampu memahami isi materi dari sebuah topik atau ulasan secara tepat. Dalam pengklasifikan informasi dokumen kita dapat menggunakan sistem temu kembali informasi (*Information Retrieval). Information Retrieval* (IR) adalah proses dimana pengumpulan data diwakili, disimpan, dan mencari tujuan penemuan pengetahuan sebagai tanggapan atas permintaan pengguna (*query*). Dalam *Information Retrieval System* (IRS), terdapat berbagai model yang dapat digunakan untuk mengukur kemiripan (pembobotan) dari suatu pencarian, diantaranya model *boolean*, model ruang vektor (*Vector Space Model*) dan model probabilistik. Pada penelitian ini akan digunakan pembobotan model ruang vektor (*Vector Space Model*) yang menekankan kepada teknik pembobotan berdasarkan kata-kata (*term*). Menurut Zumhur (Alamin, 2015) baik *query* maupun dokumen-dokumen yang disimpan, dinyatakan dalam bentuk vektor. Elemen pembobotan *Vector Space Model* yaitu pembobotan *term* yang terdapat pada dokumen dan *query*. Proses ini melibatkan berbagai macam proses mulai dari proses konversi dokumen ke teks, proses tokenisasi, proses *stopword removal*, proses *stemming*, proses pembobotan, dan proses perangkingan. Pembobotan yang didasarkan pada kata selalu dikaitkan pada teknik *stemming* untuk mendapatkan bentuk kata dasar dari kata yang bersangkutan. *Stemming* adalah proses pencari kata dasar pada suatu kata. Pada proses pencarian, Imbuhan merupakan bagian dari informasi yang perlu dihilangkan untuk mencapai efektifitas pencarian. Selain itu, perlu dilakukan proses menghilangkan *term* yang tidak bermakna dari informasi yang merupakan proses filtrasi dikenal dengan istilah *stopword removal*. *Stemming* merupakan suatu proses yang terdapat dalam sistem *Information Retrieval* yang mentransformasikan kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (*root word*) dengan menggunakan aturan-aturan tertentu. *Stemming* disebut juga teknik pengolahan bahasa alami inti untuk efesiensi dan efektifitas pada *Information Retrieval*. Salah satu *stemming* ialah nazief dan andriani. Algoritma Nazief dan Adriani merupakan algoritma *stemming* yang dikembangkan oleh Bobby Nazief dan Mirna Adriani pada tahun 1996 sebagai hasil penelitian internal Universitas Indonesia.

Banyak algoritma yang dapat digunakan dalam menentukan suatu jenis informasi atau pengklasifikasian informasi

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

salah satunya adalah *K-Nearest Neighbor* (KNN). KNN merupakan salah satu algoritma pembelajaran mesin sederhana. Algoritma *K-Nearest Neighbor* (KNN) Adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Hal ini hanya didasarkan pada gagasan bahwa "benda-benda yang 'dekat' satu sama lain juga akan memiliki karakteristik yang mirip. Jadi jika Anda tahu ciri-ciri dari salah satu objek, Anda juga dapat memprediksi untuk terdekatnya.

Didalam penelitian ini akan menggunakan teknik yang pengklasifikasian dokumen menggunakan pembobotan frekuensi kata dalam dokumen serta pengukuran nilai relevansinya pada pembobotan *query* yang ditentukan dengan fungsi kesamaan *cosine*.

Masalah yang bisa diidentifikasi adalah Dengan adanya perbedaan gaya mengulas untuk suatu ulasan topik pada sejumlah dokumen menyebabkan sulitnya dalam membangun sebuah sistem yang mampu memahami isi dari sebuah dokumen dengan banyaknya jumlah dokumen yang ada, agar dapat diklasifikasikan sesuai dengan kategori dari isi masing-masing dokumen tersebut.

Penelitian ini dibatasi bagaimana merancang dan membangun sistem pendukung keputusan penilaian kinerja karyawan. Adapun beberapa batasan yang dapat ditentukan dalam melakukan penelitian ini adalah sebagai berikut:

1. Menggunakan prototipe berbasis *web* untuk melakukan pengujian proses pengklasifikasian pada dokumen-dokumen pengujian.
2. Obyek yang digunakan sebagai media pengujian dalam penelitian ini adalah dokumen-dokumen teks berbahasa Indonesia dari beberapa media sumber yang disimpan ke dalam dokumen PDF.
3. Menggunakan pembobotan frekuensi kata ($\log_{fin}$) yang ada didalam dokumen dan menggunakan fungsi persamaan *cosine*.

STUDI PUSTAKA

Information Retrieval

Information Retrieval adalah studi atau ilmu tentang sistem pengindeksan, pencarian, dan mengingat kembali data yang dapat berupa teks atau bentuk tidak terstruktur lainnya. Ilmu ini dikenalkan oleh Vannevar Bush pada tahun 1945 dan mulai diimplementasikan pada tahun 1950-an. Mulai dari tahun 1990-an sampai sekarang, banyak teknik dan metode dari information retrieval yang dikembangkan dan digunakan. *Information Retrieval* sering disebut temu kembali informasi yang artinya suatu proses untuk menemukan kembali informasi yang tersimpan dari berbagai sumber (*resources*) yang relevan dengan cara pengindeksan (*indexing*), pencarian (*searching*), temu kembali (*recalling*) (Kent, 1971).

Konsep sederhana Information retrieval adalah proses pencarian dengan menghasilkan sesuatu yang dicari. Jika dititik-beratkan pada prosesnya akan terungkap bagaimana perjalanan informasi yang dicari, menjadi informasi yang ditemukan.

Information Retrieval System

Adalah proses dimana pengumpulan data diwakili, disimpan, dan mencari tujuan penemuan pengetahuan sebagai tanggapan atas permintaan pengguna (*query*). Proses ini melibatkan berbagai tahapan mulai dengan mewakili data dan diakhiri dengan kembali informasi yang relevan kepada pengguna. tahap peralihan meliputi penyaringan, pencarian, pencocokan dan operasi peringkat. *Information Retrieval System* (IRS) juga dapat diartikan suatu sistem untuk menemukan kembali dokumen dalam bentuk teks berbasis komputer. Oleh karena itu maka sistem ini sangat dibutuhkan, terutama dalam pencarian dokumen baik online maupun offline. (Mandala & Setiawan, 2002) Sistem temu kembali informasi (*Information Retrieval System*) digunakan untuk menemukan kembali (*retrieve*) informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Sistem ini merupakan salah satu cara

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

pencarian informasi yang isinya tidak terstruktur yaitu berupa dokumen teks. Demikian pula dengan kata kunci (keyword) kebutuhan pengguna tidak memiliki struktur yang disebut query. Hal ini yang membedakan *Information Retrieval System* (IRS) dengan sistem basis data.

Morfologi Bahasa Indonesia

Morfologi adalah salah satu cabang ilmu bahasa yang mempelajari seluk- beluk bentuk kata serta fungsi perubahan-perubahan bentuk kata itu, baik fungsi gramatik maupun fungsi semantik.

Dalam definisi lain di katakan bahwa morfologi merupakan bagian dari ilmu bahasa yang membicarakan atau mempelajari seluk- beluk bentuk kata serta pengaruh perubahan bentuk kata terhadap golongan dan arti kata (Turney & Pantel, 2010).

Fungsi Morfologi: (Ramlan, 1983)

1. Mengetahui bagaimana perubahan-perubahan bentuk kata, baik dari fungsi gramatik maupun sematik.
2. Mengetahui bagaimana seluk-beluk kata.
3. Mengetahui bagaimana suatu arti yang timbul akibat peristiwa gramatik.
4. Mempelajari peristiwa-peristiwa umum, peristiwa yang berturut-turut terjadi atau dengan kata-kata lain sebagai system dalam bahasa.

Vector Space Model

Vector space model merupakan representasi dokumen dan *query* sebagai vektor dalam ruang multidimensi, yang dimensinya adalah istilah yang digunakan untuk membangun indeks untuk mewakili dokumen. Hal ini digunakan dalam pencarian informasi, pengindeksan dan relevansi peringkat dan dapat berhasil digunakan dalam evaluasi mesin pencari *web*.

Menurut Turne, *Vector Space Model* adalah suatu model dalam sistem temu kembali informasi yang digunakan untuk mengukur kemiripan antara suatu dokumen dan suatu *query* dengan mewakili setiap dokumen dalam sebuah koleksi sebagai sebuah titik dalam ruang (vektor dalam ruang vektor) (Turney & Pantel, 2010).

Prosedur vektor model ruang dapat dibagi dalam tiga tahap: (Jitendra & Sanjay, 2012)

1. Pengindeksan dokumen yang berhubungan dengan istilah yang telah diambil dari dokument *text*.
2. Bobot istilah diindeks untuk meningkatkan pengambilan dokumen yang relevan kepada pengguna.
3. Peringkat dokumen sehubungan dengan permintaan sesuai dengan pengukuran kesamaan.

K-Nearest Neighbor

Dalam pengenalan pola, algoritma KNN adalah metode untuk mengklasifikasikan objek berdasarkan contoh pelatihan terdekat di ruang fitur. KNN adalah jenis pembelajaran berbasis pada contoh, atau *Lazy Learning* di mana fungsi ini hanya didekati secara lokal dan semua perhitungan ditangguhkan sampai klasifikasi selesai.

Algoritma *K-Nearest Neighbor* (NN atau KNN) Adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Teknik ini sangat sederhana dan mudah diimplementasikan. Dalam hal ini, jumlah data atau tetangga terdekat ditentukan oleh user yang dinyatakan dengan k. Misal ditentukan k=5, maka setiap data *testing* dihitung jaraknya terhadap data *training* dan dipilih 5 (lima) data *training* yang jaraknya paling dekat dengan data *testing*. Lalu, periksa output atau *label*-nya masing-masing. Kemudian, tentukan output mana yang frekwensinya paling banyak. Lalu masukan suatu data testing ke kelompok dengan *output* paling banyak. Misal dalam kasus klasifikasi dengan 3 (tiga) kelas, lima data tadi terbagi atas tiga data dengan *output* kelas 3. Dapat disimpulkan bahwa *output* dengan label kelas 1 adalah yang paling banyak. Maka, data baru tadi dapat dikelompokkan ke dalam kelas 1. Prosedur ini dilakukan untuk

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

semua data *testing* (Lancaster & Wilfrid, 1979).

Algoritma Nazief Andriani

Perubahan kata berimbuhan ke kata dasar disebut dengan *Stemming*. Proses *stemming* adalah bagian dari proses *preprocesing* yang berfungsi untuk mengembalikan kata dalam bentuk kata dasarnya. *Stemming* adalah suatu proses untuk mengurangi varian morfologi kata ke bentuk kata dasar pada suatu kata (Asian & Jelita, 2007). *Stemming* merupakan suatu proses dalam IRS (*Information Retrieval System*) berfungsi mentransformasikan kata dalam suatu dokumen ke kata dasarnya dengan menggunakan aturan tertentu.

Stemming disebut juga teknik pengolahan bahasa alami inti untuk efesiensi dan efektifitas pada *Information Retrieval*, Salah satu Stemming bahasa Indonesia ialah Nazief Andriani dimana pada proses *Stemming* bahasa Indonesia dihilangkan bagian sufiks, prefiks, dan konfiks.

Tokenisasi

Tokenisasi adalah proses membagi teks berupa kalimat atau paragraph dalam dokumen menjadi token-token tertentu. Tokenisasi seringkali digunakan dalam ilmu linguistic dan hasil tokenisasi berguna untuk analisis teks lebih lanjut. Proses tokenisasi adalah proses pemotongan string input berdasarkan tiap kata yang menyusunnya. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata lain oleh karakter spasi dan tab, sehingga proses tokenisasi mengandalkan karakter spasi dan tab pada dokumen untuk melakukan pemisahan kata. Di dalam pembuatan sebuah indeks istilah, dokumen dipecah menjadi unit yang lebih kecil misalnya berupa kata, frasa atau kalimat. Unit tersebut biasanya disebut sebagai token. Sedangkan algoritma untuk memecahkan kumpulan kalimat atau frasa menjadi token disebut tokenizer. Sebagai contoh, tokenisasi dari kalimat "Ali mempelajari bahasa Indonesia" menghasilkan empat token, yakni: "Ali", "mempelajari", "bahasa", "Indonesia".

Proses tokenisasi dipengaruhi oleh pengetahuan bahasa yang digunakan dalam sebuah dokumen untuk menangani karakter khusus, serta memberikan batasan token dalam sebuah dokumen. Tokenisasi menghasilkan daftar istilah beserta informasi tambahan seperti frekuensi kemunculan istilah dan posisi istilah itu muncul dalam sebuah dokumen yang digunakan pada pemrosesan selanjutnya yaitu proses pembuangan kata-kata yang tidak perlu atau filtrasi, juga dikenal dengan proses *stopword removal* (Asian & Jelita, 2007).

Filtrasi

Tahap filtrasi adalah tahap pengambilan kata penting dari hasil tokenisasi. Tahap filtrasi ini menggunakan daftar *stopword*. Penyaringan (filtrasi) terhadap kata yang tidak layak untuk dijadikan sebagai kata kunci (*keyword*) dalam pencarian dokumen sehingga kata-kata tersebut dapat dihilangkan dari dokumen. *Stopword* adalah daftar kata yang mungkin digunakan sebagai kata kunci dalam pencarian dokumen.

Pada proses pengindeksan dokumen dan *query* perlu dilakukan *stopword removal* agar dapat meningkatkan kinerja mesin pencari. Jika *stopword* terdapat pada query yang diinputkan pengguna tidak dihilangkan, hal ini dapat menyebabkan hampir semua dokumen dalam koleksi pengujian akan dikembalikan (*retrieve*). Dengan demikian akan semakin jauh dari fungsi utama suatu model mesin pencari karena tidak dapat memberikan dokumen yang relevan dengan permintaan pengguna. Karena dalam dokumen akan banyak ditemukan kata hubung, kata bantu, maupun kata ganti, yang merupakan bagian dari *stopword*.

Proses pembuangan *stopword* dimaksudkan untuk mengetahui suatu kata masuk ke dalam *stopword* atau tidak. Pembuangan *stopword* adalah proses pembuangan *term* yang tidak memiliki arti atau tidak relevan. *Term* yang diperoleh dari tahap tokenisasi dicek dalam suatu daftar *stopword*, apabila sebuah kata termasuk dalam daftar *stopword* maka kata tersebut tidak akan diproses lebih lanjut. Sebaliknya

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

apabila sebuah kata tidak termasuk di dalam daftar *stopword* maka kata tersebut akan masuk ke proses berikutnya. Penghilangan *stopword* setelah proses tokenisasi pada pengindeksan dokumen akan dapat mempercepat proses model mesin pencari karena dapat mengurangi jumlah *term* yang akan dibandingkan kemiripannya antara dokumen dan *query* serta yang akan dicari bobotnya dalam proses perankingan dokumen (Asian & Jelita, 2007)

Metode Evaluasi Keefektifan

a. Pengujian ketepatan (*Precision*)
   Pengujian ketetapan (*precision*) adalah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen yang terklafisikasi oleh sistem baik relevan maupun tidak relevan. Hasil *precision* didapat dari pembagian antara jumlah dokumen yang terklasfikasi dengan seluruh dokumen yang terklasifikasi (Alamin, 2015).
b. Pengujian Kelengkapan (*Recall*)
   Pengujian kelengkapan (*recall*) adalah perbandingan jumlah dokumen relevan yang ditetapkan sistem dengan jumlah seluruh dokumen relevan yang seharusnya terklafisikasi. Hasil *recall* didapat dari pembagian antara jumlah dokumen yang terklasifikasi dengan jumlah seluruh dokumen yang seharusnya terklasifikasi (Alamin, 2015).

Rumusan Pembobotan

Untuk mengetahui kategori suatu dokumen, diperlukan nilai dari dokumen itu sendiri, maka diperlukan sebuah cara untuk mengukur pembobotan nilai dokumen. Pada penelitian ini penulis menggunakan pembobotan frekuensi kemunculan kata didalam dokumen untuk mengukur nilai dari dokumen itu sendiri, didefinisikan dengan: (Manish & Rahul, 2013)

$W_{in} = \log f_{in}$

Ket :
i = kata
W_{in} = Pembobotan i pada dokumen n
f_{in} = Frekuensi kata i pada dokumen n
log = Proses penghilangan.

Similarity Cosine

Cosine Similarity ialah menentukan sudut antara vektor dokumen dan vektor *query*. Sudut antara dua *vector* dianggap sebagai ukuran perbedaan antara *vector*, sudut *cosine* ini dipakai untuk menghitung angka kesamaan. Menentukan sudut antara dokumen vektor dan *query* vektor ketika mereka terwakili di V-dimensi ruang *Euclidian* yang mana V adalah ukuran.

Fungsi *similarity* antara dokumen *vector* dan *query* adalah:

$$Cosine\theta = \frac{\sum_{j=1}^{V} W_{Q,j} \times W_{i,j}}{\sqrt{\sum_{j=1}^{V} W_{Q,j}^2} \times \sqrt{\sum_{j=1}^{V} W_{i,j}^2}}$$

ket:
$W_{Q,j}$ = nilai bobot *query vektor* Q pada document j
$W_{i,j}$ = nilai bobot *vektor* i pada dokumen j.

METODOLOGI PENELITIAN

Dalam penelitian ini menggunakan pendekatan metode penelitian experimental dengan membangun prototipe system yang diuji keefektifannya dengan basis ketepatan (*precision*) dan kelengkapan (*recall*) dari hasil proses yang dituangkan dengan menggunakan *model confusion matrix* sebagai alat untuk menganalisa/mengevaluasi hasil eksperimen tersebut.

Metode Pemilihan Sampel

Teknik sampling probabilitas (*Random Sampling*) merupakan teknik sampling yang dilakukan dengan memberikan peluang atau kesempatan kepada seluruh anggota populasi untuk menjadi sampel. Dengan demikian sampel yang diperoleh diharapkan menjadi sampel yang representatif. Salah satu bagian dari teknik ini adalah teknik sampling secara kluster (*Cluster Sampling*). Ada kalanya peneliti tidak tahu persis karakteristik populasi yang ingin dijadikan subjek penelitian karena populasi tersebar di wilayah yang amat luas. Untuk itu peneliti hanya dapat menentukan sampel wilayah,

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

berupa kelompok klaster yang ditentukan secara bertahap.

Pengambilan sampel dengan jumlah besar dalam sebuah populasi dapat menghasilkan penelitian yang lebih akurat. Namun karena keterbatasan peneliti, maka sampel yang digunakan hanya beberapa karena luasnya penyebaran populasi sampel. Untuk menguji tingkat efektifitas hasil pencarian dari sebuah model mesin klasifikasi dokumen, maka metode pengambilan sampel dalam penelitian ini menggunakan metode *Cluster Sampling*. Sampel dalam penelitian ini adalah 15 dokumen berbahasa indonesia yang diperoleh dari media *online* dan diubah kedalam dokumen PDF. Artikel yang digunakan terbatas pada isi artikel berbahasa Indonesia berbentuk teks tanpa disertakan gambar. Pemilihan sampel melalui pertimbangan berdasarkan keterlibatan di dalam sistem sehingga pemilihan sample menjadi lebih tepat sasaran.

Instrumentasi

Instrumen yang digunakan peneliti dalam penelitian ini, berupa sejumlah dokumen penunjang seperti jurnal, skripsi, tesis dan buku tentang penelitian terdahulu yang didapat dari media cetak maupun media *online* dan sebuah model aplikasi mesin klasifikasi yang menggunakan pembobotan jumlah kata dalam dokumen ($logf_{tn}$) dengan fungsi kesamaan *cosine.*

Teknik Analisis Data

Pengklasifikasi dokumen merupakan sebuah model untuk menyusun dokumen denan ketentuan-ketentuan yang telah disepakati. Pengklasifikasi dokumen secara manual memiliki beberapa kendala, contohnya seperti hasil yang menjadi tidak akurat dan memakan waktu yang lebih relatif lama untuk mengklasifikasikan dokumen-dokumen tersebut secara satu persatu dikarenakan harus membaca terlebih dahulu dokumen-dokumen tersebut untuk dapat mengetahui jenis kategori dokumen tersebut.

Pada klasifikasi dokumen bisa terjadi kondisi dimana terdapatnya lebih dari satu *user* yang bertugas untuk melakukan klasifikasi terhadap dokumen-dokumen itu dan pasti akan terjadi kondisi dimana *user* pertama mengetahui kategori dokumen tersebut namun *user* yang lain tidak mengetahuinya atau sependapat dikarenakan berbedanya pola pemikiran serta cara dalam mengklasifikasikannya. Tentunya ada penilaian secara subjektif yang mungkin tidak sengaja terjadi dari masing-masing individu *user* untuk memutuskan sebuah dokumen layak atau tidak dikelompokkan ke dalam kategori yang dimaksud.

Oleh karena itu diperlukan sebuah *rule* yang akan menilai sebuah dokumen apakah layak atau tidak untuk dimasukkan ke dalam suatu kategori dengan melihat aspek di dalam dokumen tersebut.

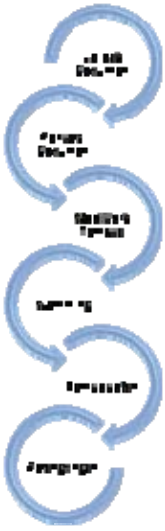
Di dalam Teknik analisis ini penulis melakukan dalam beberapa proses. Pelaksanaan analisis dimulai dari awal mula dilakukan pengumpulan data yang dilakukan secara intensif, kemudian dilanjutkan dengan proses penerapan pembobotan frekuensi jumlah kata ($logf_{tn}$) pada dokumen dengan menggunakan fungsi kesamaan *Cosine* pada model sistem aplikasi pengklasifikasian. Kemudian melakukan pemasukan semua data yang akan diuji kedalam sistem *database*, dilanjutkan proses pengujian sampel yang didapat sebelumnya. Proses selanjutnya adalah melakukan analisis tingkat efektifitas dari hasil pengujian sebelumnya.

Prototipe Model

Model sistem pengklasifikasian pada penelitian ini menggunakan prototipe model, pendekatan ini dipilih karena mempunyai struktur yang sesuai dalam mengembangkan simulasi model sistem pengklasifikasian. Dalam melakukan analisis tingkat efektifitas hasil pengklasifikasian dokumen, diperlukan sebuah sistem model pengklasifikasian yang menerapkan *Vector Space Model* menggunakan algoritma *K-nearest Neighbor* dan *Stemming* nazief adriani bahasa Indonesia dengan pembobotan frekuansi kata dalam dokumen dan fungsi kesmaan *cosine.* Proses tersebut seperti

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

proses pemecahan kalimat bahasa Indonesia menjadi sebuah kata baik dalam dokumen maupun *query*, proses *Stopword removal* yaitu mengilangkan kata yang tidak berkaitan seperti kata penghubung, proses *Stemming* kata dengan algoritma *Stemming* Nazief Adriani berbahasa Indonesia, proses penghitungan pembobotan frekuensi kata ($logf_{tn}$) dalam dokumen kemudian dihitung dengan fungsi kesamaan *cosine.* Berikut penjelasan proses yang menjadi komponen utama dalam sebuah sistem model pengklasifikasian:



Gambar 1. Proses Pengklasifikasian

1. *Upload* Dokumen
   Adalah proses dimana dokumen dimasukkan ke dalam aplikasi.
2. *Convert* Dokumen
   Adalah proses dimana dokumen PDF yang diuplod diubah formatnya menjadi dokumen teks, yang kemudian isi dokumen teks tersebut akan diproses lebih lanjut untuk mencari nilai token di dokumen tersebut
3. Proses *Stopword Removal*
   Stopword adalah kumpulan kata yang sering muncul pada dokumen dan dianggap tidak memiliki arti. *Stopword Removal* merupakan proses yang dilakukan untuk menghilangkan kata yang sering ditampilkan dalam berbagai kategori dokumen.
4. Proses *Stemming*
   Merupakan proses mengubah sebuah kata menjadi kata dasar dengan menghilangkan imbuhan yang dapat berupa awalan dan akhiran. Proses ini dilakukan setelah melalui proses filtrasi / *stopword*. Dalam penelitian ini menggunakan algoritma *stemming* nazief andriani.
5. Proses Pembobotan *term* (*Term Weighting*)
   Merupakan proses dimana setiap *term* yang ada dalam dokumen diberikan nilai bobot. Proses pembobotan kata yaitu proses pemberian bobot tiap kata yang dihitung dari jumlah kemunculan kata dalam sebuah *query* maupun isi dari sebuah dokumen.
6. Proses penghitungan nilai *similarity* (*Similarity Measurement*)
   Merupakan proses pengukuran kemiripan dokumen yang dimiliki dengan *query* yang dimasukan. *Similarity measurement* ini digunakan setelah nilai *term* ditemukan. Algoritma yang digunakan pada aplikasi ini adalah algoritma fungsi *cosine*.
7. Proses Perangkingan Dokumen
   Proses perangkingan dokumen menggunakan Algoritma *K-Nearest Neighbor. K-Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data *learning* yang jaraknya paling dekat dengan objek tersebut. Algoritma KNN termasuk algoritma *supervised learning* dimana *query instanc*e yang baru diklasifikasikan berdasarkan mayoritas dari kategori KNN. Kelas yang paling banyak yang muncul yang akan menjadi hasil klasifikasi. Proses ini mengurutkan dokumen dari yang tertinggi ke rendah berdasarkan dengan nilai *similarity*.

Pengujian Data

Teknik pengujian data yang penulis gunakan pada penelitian ini adalah sebagai berikut:

a. Pengujian Ketepatan (*Precision*)
   Pengujian ketetapan (*precision*) adalah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen yang terklafisikasi

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

oleh sistem baik relevan maupun tidak relevan.

$$\text{Precision} = \frac{\text{jumlah dokumen relevan yang terklafisikasi}}{\text{jumlah seluruh dokumen yang terklafisikasi}}$$

b. Pengujian Kelengkapan (*Recall*)
Pengujian kelengkapan (*Recall*) adalah perbandingan jumlah dokumen relevan yang ditetapkan sistem dengan jumlah seluruh dokumen relevan yang seharusnya terklafisikasi.

$$\text{Recall} = \frac{\text{Jumlah dokumen relevan yang terklafisikasi}}{\text{Jumlah seluruh dokumen yang seharusnya terklafisikasi}}$$

HASIL DAN PEMBAHASAN

Dengan menggunakan K-NN dengan k=5 maka, didapatkan nilai precision dan recall sebagai berikut:

Tabel 1. Hasil *Precision* dan *Recall* untuk K=5

Kategori	*Query*	*Precision*	*Recall*
Otomotif	Q1	100%	60%
Otomotif	Q2	100%	40%
Otomotif	Q3	80%	80%
Otomotif	Q4	100%	40%
Otomotif	Q5	80%	80%
Olahraga	Q6	100%	20%
Olahraga	Q7	100%	20%
Olahraga	Q8	40%	40%
Olahraga	Q9	100%	20%
Olahraga	Q10	100%	20%
Teknologi	Q11	100%	100%
Teknologi	Q12	100%	80%
Teknologi	Q13	100%	60%
Teknologi	Q14	100%	80%
Teknologi	Q15	100%	80%

Tabel 2. Hasil *Precision* dan *Recall* untuk K=3

Kategori	*Query*	*Precision*	*Recall*
Otomotif	Q1	100%	33%
Otomotif	Q2	100%	67%
Otomotif	Q3	100%	100%
Olahraga	Q4	100%	33%
Olahraga	Q5	100%	33%
Olahraga	Q6	100%	67%
Teknologi	Q7	100%	100%
Teknologi	Q8	67%	67%
Teknologi	Q9	100%	100%

KESIMPULAN DAN SARAN

Kesimpulan

Pada bab ini berisikan hasil dari penelitian yang telah dilakukan, maka didapat kesimpulan sebagai berikut:

1. Dengan menggunakan *k*=3 mendapatkan hasil nilai *precision* dan *recall* lebih baik daripada menggunakan *k=5*. Dengan data *document learning* untuk masing-masing kategori berjumlah lima dokumen.
2. Dari hasil pengujian tersebut didapat nilai k=5 adalah 93.33% untuk *precision* dan 54.67% untuk *recall*. Dan nilai K=3 adalah 96.33% untuk *precision* dan 66.67% untuk *recall*. Hasil nilai *precesion* dan *recall* pada proses pengklasifikasian dokumen sangat dipengaruhi oleh frekuensi kata yang ada pada dokumen, dimana proses *stemming* dan *stopword removal* sangat mempengaruhi nilai pembobotan pada setiap dokumen tersebut.

Saran

Adapun saran untuk penelitian ini berdasarkan hasil dan kesimpulan, adalah sebagai berikut:

1. Dapat menggunakan data pengujian yang berjumlah lebih banyak lagi pada proses *document learning*, agar mampu mengenali dokumen dengan beberapa macam kategori, yang tidak hanya terfokus pada beberapa kategori saja.
2. Membuat kategori yang lebih spesifik untuk dapat menghasilkan nilai pengklasifikasian yang lebih akurat lagi.
3. Mengembangkan kembali teknik *stemming* nazief andriani Bahasa Indonesia supaya dapat melakukan *stemming* untuk format teks dokumen yang menggunakan Bahasa Inggris.
4. Menggunakan fungsi kesamaan yang ada selain fungsi *cosine*, misalnya fungsi *jaccard*, fungsi *euclidean*, dan fungsi-fungsi lainnya.
5. Menggunakan teknik pembobotan lainnya untuk mendapatkan hasil pengklasifikasian yang lebih akurat.
6. Aplikasi dikembangkan supaya dapat menggunakan berbagai macam jenis

Journal of Business and Audit Information Systems
Vol. 1 (No. 1): no 9 – no 18. Th. 2018
p-ISSN 2615-6431

format dan tipe dokumen dikarenakan untuk saat ini model aplikasi hanya dapat mengenali dokumen yang berformat PDF.

DAFTAR PUSTAKA

Alamin, Z. 2015, Mesin Pencari Dokumen Teks Bahasa Indonesia, Studi Efektifitas Pencarian Pada Vector Space Model, Algoritma Stemming Porter, Pembobotan Frekuensi Term Serta Fungsi Jaccard, *Tesis*, M.KOM, Universitas Budi Luhur, Jakarta.

Asian dan Jelita. 2007, *Effective Techniques for Indonesian Text Retrieval*, Ph.D. diss., School of Computer Science and Information Technology, RMIT University, Australia.

Jitendra and Sanjay. 2012, *Analysis of Vector Space Model in information retrieval*, IJCA.

Kent, A. 1971, *Information Analysis and Retrieval*, 3 rd Edition, Becker and Heys, New York.

Lancaster, F. dan Wilfrid. 1979, *Information Retrieval Systems: Characteristics, Testing and Evaluation*, 2nd ed., New York: Jon Wiley & Sons.

Mandala, Rila dan Setiawan, Hendra. 2002, *Peningkatan Performansi Sistem Temu-Kembali Informasi dengan Perluasan Query Secara Otomatis*, Departemen Teknik Informatika, Institut Teknologi Bandung, Bandung.

Manish dan Rahul. 2013, *A Survey On Information Retrieval Models, Technique And Application*, IJETAE.

Ramlan, M. 1983, *Morfologi Suatu Tinjauan Deskriptif*,CV.Karyono, Yogyakarta.

Turney, P.D. & Pantel, P. 2010, *From Frequency to Meaning: Vector Space Models of Semantics*, Journal of Artificial Intelligence Research. 37:141-188.