STUDI KOMPARATIF PEMANFAATAN *VECTOR SPACE MODEL* PADA PENERAPAN ALGORITMA NAZIEF ADRIANI, *K-NEAREST NEIGHBOR* DAN FUNGSI *JACCARD*: KASUS PROTOTIPE APLIKASI KATAGORISASI TEKS BERBAHASA INDONESIA

Sukisno
Dosen Jurusan Teknik Informatika, Universitas Islam Syekh Yusuf Tangerang
Jl. Maulana Yusuf, Babakan Kota Tangerang, Banten, Telp 021-5527061
E-mail: sukisnonhp@gmail.com

ABSTRAK

Kemajuan teknologi yang sangat cepat mendorong manusia dalam memanfaatkan tumbuh kembangnya teknologi tersebut untuk melakukan pekerjaan yang dahulu dikerjakan secara manual. Teknik yang digunakan untuk memecahkan masalah adalah dengan menggunakan teknik *text mining* untuk pengkategorian dokumen penulisan ilmiah. Sedangkan untuk mencari nilai similaritas suatu dokumen dengan dokumen lainnya menggunakan kata kunci yang didapat dari hasil pengkategorian dokumen dan algoritma yang digunakan adalah algoritma TF-IDF *(Term Frequency – Inversed Document Frequency),* WIDF *(Weighted Inverse Document Frequency)*. Untuk pengujian sistem adalah dengan *Recall & Precision*. Tujuan dari penelitian ini adalah untuk implementasi sistem klasifikasi dokumen teks berbahasa Indonesia dengan banyak kategori dan mengetahui tingkat akurasi hasil klasifikasi dengan metode TF-IDF dan WIDF dalam mengklasifikasikan dokumen teks berbahasa Indonesia. Penelitian ini menghasilkan nilai *precision* dan *recall* rata-rata sebesar 87.50% dengan pembobotan TF-IDF dan WIDF sebesar 83.33%. Dengan ini diambil kesimpulan bahwa TF-IDF menghasilkan *precision* & *recall* tertinggi sebesar 87.50% dibandingkan dengan WIDF yaitu 83.33% menggunakan similaritas fungsi *jaccard.* Hasil terbaik adalah pembobotan kata dengan TF-IDF.

Kata Kunci: Pengkategorian Dokumen, Similaritas, *Text Mining,* TF-IDF, WIDF, *K-Nearest Neighbor,* Fungsi *Jaccard*

1. PENDAHULUAN

Adanya pertumbuhan yang pesat dari informasi pada masa kini, menjadikan katagorisasi teks *(text categorization)* yang merupakan suatu proses pengklasifikasian dokumen ke dalam satu atau lebih kategori yang telah didefinisikan sebelumnya atau ke dalam kelas-kelas dari dokumen dokumen yang sama, sebagai suatu teknik kunci dalam penanganan dan pengorganisasian data yang berupa teks.

Pada kategorisasi teks, representasi suatu dokumen adalah kata, dimana tiap kata memiliki ciri khas yang berbeda. Oleh karena itu, pada sebagian besar proses kategorisasi teks, terdapat banyak ciri khas yang mungkin terjadi, baik ciri khas yang relevan dengan tema dokumen maupun yang tidak relevan dari proses kategorisasi. Adapun metode yang mengelompokkan semua ciri khas tersebut cenderung lebih baik daripada metode yang hanya mengelompokkan ciri khas yang relevan.[4]

Pada era globalisasi seperti saat ini dunia teknologi dan informasi perkembangannya sangat cepat, sehingga mendorong timbulnya kebutuhan manusia akan sebuah layanan informasi. Kemajuan teknologi yang sangat cepat mendorong manusia dalam memanfaatkan tumbuh kembangnya teknologi tersebut untuk melakukan pekerjaan yang dahulu dikerjakan secara manual. Terlebih lagi didorong dengan perkembangan ilmu pengetahuan yang sangat cepat. Sebagai contoh, dengan adanya teknologi komputer segala kegiatan dapat dilakukan dengan cepat dan meminimalkan resiko kesalahan. Perkembangan dokumen berbasis teks menyebabkan jumlah dokumen menjadi sangat besar dan menyebabkan pencarian dan klasifikasi didalam dokumen berbasis teks menjadi sebuah pekerjaan yang tidak mudah. Perkembangan tersebut direspon dengan penelitian di bidang informatika khususnya di bidang pemprosesan dokumen teks berbahasa Indonesia.

Dalam penelitian ini teknik yang digunakan untuk memecahkan masalah diatas adalah dengan menggunakan teknik *text mining* untuk pengkategorian dokumen penulisan ilmiah. Sedangkan untuk mencari nilai similaritas suatu dokumen dengan dokumen lainnya menggunakan *learning document* yang didapat dari hasil pengkategorian dokumen dan algoritma yang digunakan adalah algoritma TF-IDF *(Term Frequency – Inversed Document Frequency),* WIDF *(Weighted Inverse Document Frequency)* dan algoritma *Vector*

Space Model. Selain itu salah satu metode kategorisasi teks yang lainnya adalah algoritma *K-Nearest Neighbour* (KNN).

2. PENELITIAN TERKAIT

2.1. Tinjauan Studi

Tinjauan studi yang penulis lampirkan ialah berupa hasil dari beberapa studi atau penulisan sebelumnya yang mempunyai keterkaitan untuk mengkaji kemampuan dan mengukur kinerja sistem informasi yang akan dibangun.

Berdasarkan informasi yang diperoleh oleh penulis, ditemukan beberapa penulisan yang relevan dengan penelitian penulis dengan beberapa metode, yaitu diantaranya sebagai berikut:

1. Penelitian yang dilakukan oleh Amir Hamzah dengan judul Klasifikasi Teks Dengan *Naïve Bayes Classifier* (NBC) untuk Pengelompokan Teks Berita dan *Abstract* Akademis.[1]

   Tujuan dari penelitian ini adalah untuk meneliti sejauh mana kinerja algoritma NBC dalam kategorisasi teks yang berupa teks berita dan teks akademis berupa abstrak akademis dari berbagai disiplin ilmu.

   Hasil penelitian ini adalah Algoritma NBC memiliki kinerja yang cukup tinggi untuk klasifikasi dokumen teks, baik dokumen berita maupun dokumen akademik. Pada klasifikasi dokumen berita didapatkan akurasi yang lebih tinggi (maksimal 91%) dibandingkan dengan dokumen akademik (maksimal 82%). Baik pada dokumen berita maupun dokumen akademik, penggunaan 50% dokumen sebagai dokumen pelatihan memberikan kinerja akurasi diatas 75%.
2. Penelitian yang dilakukan oleh F.Soesianto dan Adhi Susanto dengan judul Studi Kinerja Fungsi-fungsi Jarak dan Similaritas Dalam *Clustering* Dokumen Teks Berbahasa Indonesia.[3]

   Tujuan dari penelitian ini adalah melakukan kajian tentang unjuk kerja fungsi jarak *euclidean* dengan empat fungsi similaritas *jaccard, dice, cosine, euclidean* dan *pearson* jika diterapkan untuk melakukan *clustering* dokumen teks berbahasa Indonesia.

   Hasil penelitian ini adalah Fungsi yang kinerjanya terburuk adalah fungsi similaritas korelasi *pearson*. Diduga kinerja komputasi yang buruk karena kompleksitas formulanya. Fungsi similaritas yang lain, yaitu *Dice* dan *Jaccard* kinerjanya mendekati fungsi similaritas *cosine*, meskipun dari sisi efisiensi komputasinya masih kalah dengan *cosine*. Fungsi jarak *euclidean* memiliki kinerja yang buruk meskipun tidak yang terburuk, baik dari sisi hasil *clustering* yang dihasilkan maupun dari sisi efisiensi komputasinya.
3. Penelitian yang dilakukan oleh Diah Pudi Langgeni, ZK. Abdurahman Baizal dan Yanuar Firdaus A.W dengan judul *Clustering* Artikel Berita Berbahasa Indonesia Menggunakan *Unsupervised Feature Selection*.[2]

   Tujuan dari penelitian ini adalah untuk mengimplementasikan *unsupervised feature selection* yaitu *Document Frequency (DF)* dan *Term Contribution (TC)* pada clustering berita berbahasa Indonesia.

   Hasil dari penelitian adalah *Term Contribution* lebih baik daripada *Document Frequency* yaitu dapat menghasilkan nilai *precision* dan *entropy* lebih baik dengan fitur yang lebih sedikit. Hal ini dikarenakan *Term Contribution* mempertimbangkan frekuensi kemunculan *term* dan frekuensi dokumen sebuah *term*, sehingga *term* yang tetap dipertahankan adalah *term* yang khas atau bersifat diskriminator, berbeda halnya dengan *Document Frequency* yang hanya mempertahankan *term – term* yang bersifat umum.
4. Penelitian yang dilakukan oleh Listiyanti Musa, Angelina Prima Kurniati dan Moch Arif Bijaksana dengan judul Analisis dan Perbandingan Penggunaan Metode *Distributional Feature* Dengan TF-IDF dan ITF pada *Text Categorization*.[5]

   Tujuan dari penelitian ini menganalisis pengaruh penggunaan *Distributional Feature* pada performansi *text categorization* dengan membandingkan performansi antara *text categorization* yang hanya menggunakan fitur TF-IDF, TF-IDF dengan *Distributional Feature* serta *text categorization* yang menggunakan fitur ITF yang digabungkan dengan *Distributional Feature*. Performansi dalam hal ini berupa akurasi dari pengklasifikasian dokumen.

   Hasil penelitian ini adalah penggunaan *Distributional Feature* dengan skema pembobotan TF-IDF (TFIDF-DF) dan *Distributional Feature* dengan ITF (ITF-DF) mampu menghasilkan performansi yang cukup baik. Hal ini dibuktikan dengan rata-rata *precision*, *recall* dan *F1 measure* yang mendekati angka 1.

Berdasarkan pada penelitian diatas dapat dilihat belum adanya penelitian yang dilakukan untuk membuat sebuah model kategorisasi dokumen bahasa Indonesia dengan menggunakan pembobotan *WIDF (Weighted Inverse Document Frequency)*. Dalam proses ini akan dilakukan dengan menggunakan efektifitas *Vector Space Model* dan *Stemming* Algoritma Nazief Adriani serta pembobotannya menggunakan frekuensi *term*

TF-IDF dan *term weight* WIDF, untuk fungsi kesamaan menggunakan fungsi *jaccard* dan dalam proses pengukuran menggunakan *precision* dan *recall* untuk mengukur ketepatan dan kelengkapan dokumen berdasarkan 3 bentuk *query* sehingga didapat bentuk *query* yang tepat dalam proses katagorisasi dokumen.

2.2. Tinjauan Obyek Penelitian

Penelitian ini sesuai dengan judulnya yaitu studi komparatif pemanfaatan *vector space model* pada penerapan algoritma nazief adriani , *K-nearest neighbor* dan fungsi *jaccard* kasus prototipe aplikasi katagorisasi teks berbahasa Indonesia, dimana dalam penelitian ini tidak menggunakan obyek penelitian disalah satu perusahaan akan tetapi penelitian ini lebih kepada mengklasifikasikan dokumen-dokumen dalam bentuk pdf yang didapatkan dari media *online* yang mempunyai kategori pada masing-masing dokumen uji. Obyek penelitian ini adalah teks dalam bentuk bahasa Indonesia.

2.3. Hipotesis

Berdasarkan kerangka konsep yang telah dikemukakan, maka hipotesis dari penelitian ini dapat dirumuskan sebagai berikut:

1. Diduga aplikasi katagorisasi dokumen/teks dapat mengkatagorisasi secara otomatis dokumen-dokumen elektronik kedalam sebuah kategori yang telah ditentukan sebelumnya berdasarkan *query*.
2. Diduga bahwa untuk meningkatkan performansi katagorisasi dokumen/teks dibutuhkan bentuk *query* yang tepat yang dapat di ukur dengan teknik pengukuran *Precision* dan *Recall* serta didukung dengan menggunakan studi efektifitas penerapkan *Vector space model*, algoritma *stemming* nazief adriani, pembobotan *Term Frequency* dan fungsi kesamaan (fungsi *jaccard*).
3. Diduga bahwa dengan menggunakan pembobotan WIDF *(Weighted Inverse Document Frequency)* menghasilkan performansi yang lebih baik untuk kategorisasi teks berbahasa Indonesia.

3. PEMBAHASAN HASIL PENELITIAN

3.1. Data Set

Pada penelitian ini digunakan dokumen berformat *pdf* sebagai dokumen eksperimen / *data set*. Dokumen tersebut terdiri dari 12 dokumen yang dibagi menjadi 3 kategori, dari masing-masing kategori terdapat 4 sampel dokumen. Dokumen tersebut diperoleh dari media *online* yang diubah menjadi dokumen *pdf*. Kategori dan dokumen dari tiap kategori yang digunakan sebagai data *set* adalah sebagai berikut:

1. Kategori Politik
   - D1 = JK Aburizal dan Agung Laksono Sepakat Damai
   - D2 = Tetap Gelar Rapimnas Partai Golkar Ical Menghadap Jusuf Kalla
   - D3 = JK Senior Golkar Turun Tangan Satukan Partai
   - D4 = JK Indonesia Pernah Lebih Miskin dari Somalia dan Kamboja
2. Kategori Ekonomi
   - D5 = Jokowi Resmikan Proyek Kereta Cepat Jakarta-Bandung Pagi Ini
   - D6 = Jakarta-Bandung Cuma 35 Menit Berapa Harga Tiket Kereta Cepat
   - D7 = Tak Hadir di Acara Kereta Cepat Ini Alasan Menteri_Jonan
   - D8 = 3 Penjelasan JK Atas Pembelian Gerbong Kereta Bekas Jepang
3. Kategori Kesehatan
   - D9 = Iklan Rokok Elektronik Pengaruhi Anak
   - D10 = 3 Alasan Kenapa Rokok Elektrik Bisa Membahayakan
   - D11 = Usai Minuman Alkohol Mendag akan Larang Rokok Elektrik
   - D12 = Dampak Rokok Elektrik Masih Perlu Dikaji

Data set adalah data yang digunakan untuk *learning* dokumen ke *database* aplikasi *learning machine*. Hal tersebut mempunyai tujuan supaya mesin mengenal dokumen-dokumen sesuai dengan kategori yang diinput oleh *user*. Data *set* berfungsi sebagai data *learning* (D1 sampai dengan D12) dan sebagai data *query* (Q1 sampai dengan Q12).

3.2. Hasil Pengujian

Hasil pengujian adalah hasil *similarity* yang di hitung dengan fungsi *jaccard*, dimana *term* pada tiap dokumen sudah dihitung pembobotannya dengan pembobotan TF-IDF dan WIDF. Berikut ini hanya dicontohkan sampai dengan Q3.

3.2.1. Hasil *Similarity* Dokumen Dengan Pembobotan TF-IDF

3.2.1.1. Jika D1 Digunakan Sebagai Q1 *(Query 1)*

Tabel IV-1. *Similarity* Pembobotan TF-IDF Q1

Similarity (Q1, D1)	1
Similarity (Q1, D2)	0.17756262
Similarity (Q1, D3)	0.199310033
Similarity (Q1, D4)	0.067241682
Similarity (Q1, D5)	0
Similarity (Q1, D6)	0
Similarity (Q1, D7)	0
Similarity (Q1, D8)	0.016409496
Similarity (Q1, D9)	0
Similarity (Q1, D10)	0
Similarity (Q1, D11)	0
Similarity (Q1, D12)	0

3.2.1.2. Jika D2 Digunakan Sebagai Q2 *(Query 2)*

Tabel IV-2. *Similarity* Pembobotan TF-IDF Q2

Similarity (Q2, D1)	0.17756262
Similarity (Q2, D2)	1
Similarity (Q2, D3)	0.124791727
Similarity (Q2, D4)	0.041754356
Similarity (Q2, D5)	0
Similarity (Q2, D6)	0
Similarity (Q2, D7)	0
Similarity (Q2, D8)	0.013892419
Similarity (Q2, D9)	0
Similarity (Q2, D10)	0
Similarity (Q2, D11)	0
Similarity (Q2, D12)	0

3.2.1.3. Jika D3 Digunakan Sebagai Q3 *(Query 3)*

Tabel IV-3. *Similarity* Pembobotan TF-IDF Q3

Similarity (Q3, D1)	0.199310033
Similarity (Q3, D2)	0.124791727
Similarity (Q3, D3)	1
Similarity (Q3, D4)	0.087963932
Similarity (Q3, D5)	0
Similarity (Q3, D6)	0
Similarity (Q3, D7)	0
Similarity (Q3, D8)	0.017197688
Similarity (Q3, D9)	0
Similarity (Q3, D10)	0
Similarity (Q3, D11)	0
Similarity (Q3, D12)	0

3.2.2. Hasil *Similarity* Dokumen Dengan Pembobotan WIDF

3.2.2.1. Jika D1 Digunakan Sebagai Q1 *(Query 1)*

Tabel IV-4. *Similarity* Pembobotan WIDF Q1

Similarity (Q1, D1)	1
Similarity (Q1, D2)	0.219793643
Similarity (Q1, D3)	0.126641253
Similarity (Q1, D4)	0.036707496
Similarity (Q1, D5)	0.004933399
Similarity (Q1, D6)	0.007518797
Similarity (Q1, D7)	0.004319388
Similarity (Q1, D8)	0.037149706
Similarity (Q1, D9)	0.004859086
Similarity (Q1, D10)	0.006715917
Similarity (Q1, D11)	0.008368201
Similarity (Q1, D12)	0.005767013

3.2.2.2. Jika D2 Digunakan Sebagai Q2 *(Query 2)*

Tabel IV-5. *Similarity* Pembobotan WIDF Q2

Similarity (Q2, D1)	0.219793643
Similarity (Q2, D2)	1
Similarity (Q2, D3)	0.097296079
Similarity (Q2, D4)	0.028424644
Similarity (Q2, D5)	0.004341534
Similarity (Q2, D6)	0.006453695
Similarity (Q2, D7)	0.003774602
Similarity (Q2, D8)	0.027348119
Similarity (Q2, D9)	0.00477327
Similarity (Q2, D10)	0.006289308
Similarity (Q2, D11)	0.007575758
Similarity (Q2, D12)	0.005509642

3.2.2.3. Jika D3 Digunakan Sebagai Q3 *(Query 3)*

Tabel IV-6. *Similarity* Pembobotan WIDF Q3

Similarity (Q3, D1)	0.126641253
Similarity (Q3, D2)	0.097296079
Similarity (Q3, D3)	1
Similarity (Q3, D4)	0.024803418
Similarity (Q3, D5)	0.003051882
Similarity (Q3, D6)	0.004597701
Similarity (Q3, D7)	0.002561288
Similarity (Q3, D8)	0.024296331
Similarity (Q3, D9)	0.005649718
Similarity (Q3, D10)	0.006060606
Similarity (Q3, D11)	0.007092199
Similarity (Q3, D12)	0.005405405

3.3. Hasil Perangkingan KNN

Dari hasil perhitungan dengan menggunakan fungsi *jaccard* pada pembahasan diatas, maka selanjutnya di lakukan perangkingan berdasarkan metode *K-Nearest Neighbor* (KNN) yaitu dalam penelitian ini diatur dengan nilai K=4. Perangkingan tersebut diurutkan dari hasil *similarity* tertinggi ke *similarity* terendah.

3.3.1. Hasil Perangkingan KNN Dengan Pembobotan TF.IDF

Hasil perangkingan KNN=4 dengan pembobotan TF.IDF adalah seperti pada tabel dan gambar dibawah ini. Nilai *precision* diperoleh dari hasil bagi antara jumlah dokumen yang relevan dibagi dengan jumlah semua dokumen yang relevan dan tidak relevan pada tabel perangkingan, sedangkan nilai *recall* diperoleh dari hasil bagi antara jumlah dokumen yang relevan pada tabel perangkingan dibagi dengan jumlah dokumen masing-masing kategori. Dalam penelitian ini, jumlah masing-masing kategori adalah 4 dokumen.

3.3.1.1. Hasil Perangkingan TF.IDF Q1

Tabel IV-7. Hasil Perangkingan TF.IDF Q1 Dengan *Data Set*

Kategori *Query*	Dokumen *Query*	*Data Set*	Hasil *Similarity*	KNN (K=4)	Kategori	Keterangan
Politik	Q1	D1	1	1	Politik	Relevan
		D3	0.199310033	2	Politik	Relevan
		D2	0.17756262	3	Politik	Relevan
		D4	0.067241682	4	Politik	Relevan
P = 4/4 =	1.00	= Precision 100%				
R = 4/4 =	1.00	= Recall 100%				

3.3.1.2. Hasil Perangkingan TF.IDF Q2

Tabel IV-8. Hasil Perangkingan TF.IDF Q2 Dengan *Data Set*

Kategori *Query*	Dokumen *Query*	*Data Set*	Hasil *Similarity*	KNN (K=4)	Kategori	Keterangan
Politik	Q2	D2	1	1	Politik	Relevan
		D1	0.17756262	2	Politik	Relevan
		D3	0.124791727	3	Politik	Relevan
		D4	0.041754356	4	Politik	Relevan
P = 4/4 = 1.00 = Precision 100%						
R = 4/4 = 1.00 = Recall 100%						

3.3.1.3. Hasil Perangkingan TF.IDF Q3

Tabel IV-9. Hasil Perangkingan TF.IDF Q3 Dengan *Data Set*

Kategori *Query*	Dokumen *Query*	*Data Set*	Hasil *Similarity*	KNN (K=4)	Kategori	Keterangan
Politik	Q3	D3	1	1	Politik	Relevan
		D1	0.199310033	2	Politik	Relevan
		D2	0.124791727	3	Politik	Relevan
		D4	0.087963932	4	Politik	Relevan
P = 4/4 = 1.00 = Precision 100%						
R = 4/4 = 1.00 = Recall 100%						

3.3.2. Hasil Perangkingan KNN dengan pembobotan WIDF

Hasil perangkingan KNN=4 dengan pembobotan WIDF adalah seperti pada tabel dan gambar dibawah ini. Nilai *precision* diperoleh dari hasil bagi antara jumlah dokumen yang relevan dibagi dengan jumlah semua dokumen yang relevan dan tidak relevan pada tabel perangkingan, sedangkan nilai *recall* diperoleh dari hasil bagi antara jumlah dokumen yang relevan pada tabel perangkingan dibagi dengan jumlah dokumen masing-masing kategori. Dalam penelitian ini, jumlah masing-masing kategori adalah 4 dokumen.

3.3.2.1. Hasil Perangkingan WIDF Q1

Tabel IV-10. Hasil Perangkingan WIDF Q1 Dengan *Data Set*

Kategori *Query*	Dokumen *Query*	*Data Set*	Hasil *Similarity*	KNN (K=4)	Kategori	Keterangan
Politik	Q1	D1	1	1	Politik	Relevan
		D2	0.219793643	2	Politik	Relevan
		D3	0.126641253	3	Politik	Relevan
		D8	0.037149706	4	Ekonomi	Tidak Relevan
P = 3/4 = 0.75 = Precision 75%						
R = 3/4 = 0.75 = Recall 75%						

3.3.2.2. Hasil Perangkingan WIDF Q2

Tabel IV-11. Hasil Perangkingan WIDF Q2 Dengan *Data Set*

Kategori *Query*	Dokumen *Query*	*Data Set*	Hasil *Similarity*	KNN (K=4)	Kategori	Keterangan
Politik	Q2	D2	1	1	Politik	Relevan
		D1	0.219793643	2	Politik	Relevan
		D3	0.097296079	3	Politik	Relevan
		D4	0.028424644	4	Politik	Relevan
P = 4/4 = 1.00 = Precision 100%						
R = 4/4 = 1.00 = Recall 100%						

3.3.2.3. Hasil Perangkingan WIDF Q3

Tabel IV-12. Hasil Perangkingan WIDF Q3 Dengan *Data Set*

Kategori *Query*	Dokumen *Query*	*Data Set*	Hasil *Similarity*	KNN (K=4)	Kategori	Keterangan
Politik	Q3	D3	1	1	Politik	Relevan
		D1	0.126641253	2	Politik	Relevan
		D2	0.097296079	3	Politik	Relevan
		D4	0.024803418	4	Politik	Relevan
P = 4/4 = 1.00 = Precision 100%						
R = 4/4 = 1.00 = Recall 100%						

3.4. Hasil *Precision* dan *Recall*

Dengan menggunakan nilai KNN = 4, maka diperoleh hasil *precision* dan *recall* sebagai berikut :

Tabel IV-13. Hasil Pengujian *Precision* dan *Recall* (K=4)

Kategori	Query	*Precision*		*Recall*	
		TF-IDF	WIDF	TF-IDF	WIDF
Politik	Q1	100%	75%	100%	75%
Politik	Q2	100%	100%	100%	100%
Politik	Q3	100%	100%	100%	100%
Politik	Q4	100%	75%	100%	75%
Ekonomi	Q5	100%	100%	100%	100%
Ekonomi	Q6	75%	100%	75%	100%
Ekonomi	Q7	100%	75%	100%	75%
Ekonomi	Q8	50%	25%	50%	25%
Kesehatan	Q9	75%	75%	75%	75%
Kesehatan	Q10	75%	75%	75%	75%
Kesehatan	Q11	75%	100%	75%	100%
Kesehatan	Q12	100%	100%	100%	100%

3.5. Hasil dan Pembahasan

Hasil rata-rata *precision* dan *recall* dari perhitungan menggunakan TF.IDF dan WIDF adalah sebagai berikut:

Tabel IV-14. Hasil Perhitungan Rata-rata *Precision* dan *Recall*

Measure	TF-IDF	WIDF
Precision	87.50%	83.33%
Recall	87.50%	83.33%

Dengan menggunakan data *set* sebagai dokumen *learning* sebanyak 12 *data set* dengan 3 kategori menghasilkan nilai *precision* dan *recall* rata-rata

sebesar 87.50% dengan pembobotan TF-IDF dan WIDF sebesar 83.33%. Dengan ini diambil kesimpulan bahwa TF-IDF menghasilkan *precision* & *recall* tertinggi sebesar 87.50% dibandingkan dengan WIDF yaitu 83.33% menggunakan similaritas fungsi *jaccard.*

Hasil diatas tidak sesuai dengan dugaan pada saat hipotesa bahwa pembobotan WIDF lebih baik untuk mengklasifikasi dokumen *teks* berbahasa Indonesia. Dugaan bahwa pembobotan term dengan WIDF lebih baik adalah salah, karena pembobotan dengan menggunakan TF-IDF menghasilkan nilai *precision* dan *recall* yang tinggi. Pembobotan TF-IDF lebih baik dibandingkan dengan pembobotan WIDF.

3.6. Implikasi Penelitian

Berdasarkan hasil dalam penelitian ini, maka dapat disusun implikasi penelitian yang ditinjau dari aspek sistem, manajerial dan aspek penelitian lanjut. Implikasi dari aspek sistem terkait dengan konsep strategik, taktis sampai dengan teknis operasional, desain *hardware*, *software*, dan infrastruktur yang diperlukan. Implikasi dari aspek manajerial berkaitan dengan terkait organisasi yang mungkin perlu disempurnakan, sumber daya manusia yang perlu ditingkatkan kompetensinya, strategi atau kebijakan serta aturan-aturan yang perlu dibuat untuk mengatasi masalah atau meningkatkan pengelolaan obyek penelitian berdasarkan temuan – temuan dan interpretasi hasil penelitian. Dan implikasi dari aspek penelitian lanjut berkaitan dengan penelitian lanjutan yang diperlukan untuk meningkatkan kualitas penelitian sebelumnya.

Langkah berikutnya setelah prototipe katagorisasi teks berbahasa Indonesia diimplementasikan dalam sebuah sistem aplikasi adalah dengan menjabarkan implikasi dari aplikasi yang sudah diimplementasikan. Berikut adalah aspek-aspek implikasi yang dijelaskan dalam tiga bagian:

3.6.1. Aspek Sistem

Dari segi sistem, implikasi penelitian yang ditimbulkan dengan adanya pengembangan sistem aplikasi ini adalah memudahkan *staff* atau *user* dalam mengklasifikasikan dokumen berformat *pdf*, aplikasi tersebut mudah dan nyaman dan tidak membosankan dalam mengelola dokumen karena tampilannya yang cukup menarik dan sederhana. Dengan adanya sistem aplikasi ini tentunya kegiatan kinerja *staff* dan *user* lebih efektif dan efisen baik dari segi waktu dan biaya.

Adapun kebutuhan teknis untuk tahap implementasi sistem untuk perusahaan terdiri dari perangkat keras dan perangkat lunak yaitu :

1. Perangkat Keras *(Hardware)*
   Diperlukan pemanfaatan *server* sebagai sumber *resources* untuk pengembangan aplikasi katagorisasi teks. Dari arsitektur yang telah dibuat, maka perangkat keras yang dibutuhkan dapat memenuhi standar minimal kebutuhan sistem. Standar minimal spesifikasi perangkat keras PC yang dibutuhkan seperti berikut:

Tabel IV-15. Standar Minimal Spesifikasi

No	Jenis *Hardware*	Spesifikasi
1	*Processor*	Intel ® Core ™ i3CPU 2.27 GHz
2	*Memory (RAM)*	2.00 GB
3	*Harddisk*	500 GB
4	*Monitor*	Resolusi 1024 x 768
5	Jaringan	*Lan, Modem* (Koneksi Internet)
6	Perangkat	*Keyboard, Mouse & Monitor*

2. Perangkat Lunak *(Software)*
   Implikasi dari sisi perangkat lunak tidak terlalu signifikan karena perangkat lunak yang digunakan adalah *browser* yang digunakan untuk menampilkan sistem, *webserver, PHP* dan *MySQL database*. *Web server*, program *PHP* dan *database MySQL* telah terinstall sebelumnya di *server*. *Browser* di PC *client* umumnya sudah terinstall bersamaan dengan sistem operasi dan sistem sistem operasi yang digunakan adalah sistem operasi yang sudah biasa digunakan karena sistem ini berbasis web jadi tidak berpengaruh dengan sistem operasi yang digunakan. Standar minimal spesifikasi perangkat lunak yang dibutuhkan dikelompokan dalam dua bagian yaitu di sisi *server* dan di sisi *client* seperti berikut:
   a. *Server (Hosting)*
      i. *Apache web server* versi 2.2.9
      ii. *MySQL client* versi 5.0.67
      iii. PHP versi 5.2.6
   b. *Client*
      i. Sistem operasi *windows* 7
      ii. *Browser* yang mendukung *javascript* dan *flash player*, seperti : *Internet explorer* versi 8.0, *Mozilla Firefox* versi 4.01, *Google Chrome* versi 11.0

3.6.2. Aspek Manajerial

Informasi menjadi sangat penting pada zaman modern sekarang ini, oleh karena tidak jarang informasi bisa menjadi sesuatu yang sangat mahal untuk diketahui terutama bisa menyangkut masalah informasi di kalangan perusahaan ataupun individu dimana informasi bisa mempengaruhi aspek manajerial. Semakin rahasia suatu informasi semakin tinggi pesan rahasia dapat mempengaruhi aspek manajerial.

Oleh karena itulah, penerapan yang diusulkan dalam penelitian ini yaitu sistem aplikasi katagorisasi teks

berbasis *file* PDF yang telah dibuat melalui penelitian ini. Dari aspek manajerial, implikasi hasil penelitian dapat dikelompokkan dalam empat kategori, yaitu:

1. Organisasi
   Aplikasi katagorisasi teks berbahasa Indonesia jika di terapkan didalam lingkungan organisasi maka akan memberikan dampak yang signifikan. Hal tersebut akan menciptakan budaya organisasi yang baru, dimana semula belum ada pengelolaan pengetahuan dengan pemanfaatan teknologi informasi, sekarang dapat dilakukan melalui sistem yang memanfaatkan Teknologi Informasi dan Komunikasi (TIK), sehingga dapat meningkatkan penyerapan pengetahuan secara efisien.
2. Sumber Daya Manusia
   Aspek manajerial yaitu diperlukan sumberdaya manusia yang mampu menjalankan manajemen secara efisien. Aspek manajerial sumber daya manusia meliputi peningkatan kompetensi. Aspek manajerial terdiri dari perkembangan realisasi penerapan aplikasi, pencapaian target penggunaan aplikasi dan kendala serta tindak lanjut dari hasil evaluasi.
   Untuk tim yang melibatkan sumber daya manusia adalah administrator dan *user* yang memakai aplikasi katagorisasi teks berbahasa Indonesia. Administrator sendiri yang ditunjuk adalah dari bagian staff IT yang ada pada sebuah instansi yang menggunakan aplikasi katagorisasi tersebut.
3. Pelatihan
   Untuk memperlancar dalam penggunaan aplikasi katagorisasi dokumen teks berbahasa Indonesia, diperlukan pelatihan atau *training* terlebih dahulu kepada *user* agar pengoperasian program berjalan dengan baik dan benar. Hal ini dilakukan supaya sumber daya manusia mengerti tentang dunia teknologi informasi dan komunikasi. Selain training, perlu diadakan sosialisasi dan workshop tentang tatacara penggunaan aplikasi ini.

4. Aturan – aturan
   Implementasi sistem yang telah dibuat perlu dibuat aturan-aturan dalam pelaksanaannya supaya dapat memacu untuk mengembangkan budaya organisasi yang berbasis pembelajaran *(learning organization)* dan mengembangkan budaya *knowledge sharing* dalam organisasi.

3.6.3. Aspek Penelitian Lanjut

Dari hasil penelitian yang telah dilakukan masih memiliki kekurangan dan memerlukan penelitian lanjutan untuk menyempurnakannya. Beberapa hal yang perlu di perhatikan dalam penelitian lanjutan adalah:

1. Memperluas ruang lingkup dimana dalam penelitian ini hanya mengklasifikasi dokumen berbahasa Indonesia dan berformat PDF, maka untuk mengembangkan penelitian berikutnya dapat menggunakan klasifikasi dalam bahasa Indonesia dan bahasa Inggris dengan proses *stemming* yang lebih maksimal. Penelitian lanjut sangat memerlukan algoritma *stemming* dan algoritma similaritas yang lain untuk membandingkan hasil antara beberapa similaritas yang ada pada beberapa teori. Perlu adanya pengembangan dokumen *input* dimana dalam penelitian ini hanya menggunakan dokumen berformat PDF, maka selanjutnya dapat mengembangkan penelitian ini dengan dokumen *input* selain PDF atau kombinasi antara PDF dan *file* yang berformat lain.
2. Dibuat tampilan khusus untuk *smartphone* atau aplikasi *native* untuk *smartphone* berbasis sistem operasi *mobile device* seperti IOS, *Android* ataupun *Windows Phone*. Dengan memanfaatkan teknologi *cloud* untuk penyimpanan dokumen berformat PDF seperti *onedrive, dropbox* maupun *google drive*.
3. Semakin banyak penelitian yang dilakukan, terutama yang berhubungan dengan klasifikasi dokumen atau data yang di lakukan, diharapkan dapat terus meningkatkan variasi-variasi teknik keamanan yang ada sekarang ini dan meningkatkan serta memberikan rasa aman terhadap pengguna ke depannya.

3.7. Rencana Implementasi Sistem

Rencana implementasi sistem merupakan tahap awal dari penerapan sistem dan tujuan dari kegiatan implementasinya adalah agar sistem yang baru dapat beroperasi sesuai dengan yang diharapkan. Dalam proses implementasi sistem aplikasi katagorisasi teks ini diperlukan beberapa tahapan perencanaan untuk implementasi sistem. Tahapan tersebut adalah sebagai berikut:

Tabel IV-16. Rencana Implementasi Sistem

No	Kegiatan	Bulan 1				Bulan 2				Bulan 3				Bulan 4			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Pengadaan *Hardware* dan *Software*	■															
2	Instalasi *Hardware* dan *Software*		■	■													
3	Pemilihan *Operator*				■												
4	Pelatihan Pengguna					■	■										
5	Sosialisasi Kepada Pengguna							■	■								

6	Ujicoba Sistem Baru																
7	Evaluasi dan Perbaikan Sistem Baru																

Langkah – langkah kegiatan implementasi sistem aplikasi katagorisasi teks berbahasa Indonesia adalah sebagai berikut:

1. Pengadaan *Hardware* dan *Software*.
   Hardware dan *software* yang diperlukan untuk menunjang sistem ini adalah:
   a. *Hardware* berupa PC *Server*, yang akan diinstal sistem operasi *server*, *web server* dan database *server*.
   b. *Software* sistem operasi *Server*, termasuk didalamnya adalah aplikasi *Web Server* dan *Database Server*. Sistem operasi yang digunakan adalah Windows 7, *Web Server* menggunakan *Apache versi* 2.2.9, dan *Database Server* menggunakan MySQL versi 5.0.67
2. Instalasi *Hardware* dan *Software*.
   Setelah *hardware* untuk *server* dan *software*-nya tersedia maka selanjutnya adalah proses instalasi operasi *server*, instalasi *web server* dan *database server* serta instalasi aplikasi katagorisasi teks yang sudah dibangun.
3. Pemilihan Operator.
   Salah satu faktor yang mempunyai peranan penting dalam penerapan sistem baru berupa aplikasi sistem aplikasi katagorisasi teks, yaitu operator atau user yang selama ini bertugas untuk melakukan manajemen klasifikasi dokumen berbasis PDF. Untuk *administrator* terbagi menjadi 2 bagian, yaitu *administrator* sistem dan a*dministrator security*.
4. Pelatihan Pengguna.
   Pengguna atau *user* yang telah dipilih akan mengikuti pelatihan agar memahami sistem yang baru tersebut dan dapat menjalankan dengan baik dan tepat. Pelatihan tersebut dilakukan untuk mengoperasikan sistem, merawat dan menjaga sistem. pelatihan ini tidak hanya meningkatkan keahlian/keterampilan pengguna, namun juga memudahkan penerimaan pengguna terhadap sistem yang baru.
5. Sosialisasi Kepada Pengguna.
   Kegiatan sosialisasi dilakukan untuk memberikan informasi tentang keberadaan sistem aplikasi katagorisasi teks dokumen kepada semua pihak yang terkait.
6. Ujicoba Sistem Baru.
   Pengujian sistem baru harus dilakukan untuk mengetahui apakah sistem yang baru tersebut dapat berjalan sesuai dengan yang direncanakan atau tidak.
7. Evaluasi dan Perbaikan Sistem Baru.
   Kegiatan evaluasi ini bertujuan untuk mengetahui hasil dari sistem yang baru dibandingkan dengan hasil yang dicapai oleh sistem lama.

4. KESIMPULAN

Dari hasil penelitian yang telah dilaksanakan, maka dapat disimpulkan bahwa:

1. Aplikasi yang dibuat dapat mengklasifikasikan dokumen teks berbahasa Indonesia dengan pembobotan TF-IDF dan WIDF dengan *similarity* fungsi *jaccard.*
2. Bila dilihat dari hasil *precision* dan *recall*, maka dalam klasifikasi dokumen dapat menggunakan pembobotan *term* TF-IDF yang lebih baik dibandingkan dengan pembobotan *term* WIDF.
3. Dengan KNN=4 akan menghasilkan *precision* dan *recall* yang besar dibanding KNN lebih dari 4, karena data pada masing-masing kategori adalah berjumlah 4 file.
4. Dari hasil pengujian *precision* dan *recall* didapat nilai rata-rata tertinggi 87.50% pada pembobotan TF-IDF, hal ini berarti dengan menggunakan pembobotan TF-IDF, aplikasi mampu melakukan ketepatan dan kelengkapan dokumen saat proses klasifikasi lebih besar dibanding dengan pembobotan WIDF sebesar 83.33%.
5. Untuk menghasilkan klasifikasi dengan nilai *precision* dan *recall* yang lebih baik, sangat dipengaruhi oleh jumlah *term* yang ada pada dokumen, dimana proses *stemming* sangat berpengaruh dalam membentuk kata dasar yang mewakili *term* tiap dokumen.

REFERENSI:

[1] Amir Hamzah. *Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan Abstract Akademis.* Seminar Nasional Aplikasi Sains & Teknologi. Yogyakarta. 2012.

[2] Diah Pudi Langgeni, ZK.Abdurahman Baizal dan Yanuar Firdaus A.W. *Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection.* Institut Teknologi Telkom: Bandung.2010.

[3] F.Soesianto, Amir Hamzah, Adhi Susanto & Jazi Eko Istiyanto. *Studi Kinerja Fungsi-fungsi Jarak dan Similaritas Dalam Clustering Dokumen Teks Berbahasa Indonesia.* Seminar UPN Veteran. Yogyakarta.2008.

[4] Mooney, R., *Intelligent Information Retrieval and Web Search,* Austin: Texas University Pr., 2001.

[5] Musa, Listiyanti, Prima Kurniati, Angelina & Arif Bijaksana, Moch. *Analisis dan Perbandingan Penggunaan Metode Distributional Feature dengan TFIDF dan ITF pada Text Categorization.* Universitas Telkom.2012.