PEMBUATAN JUDUL OTOMATIS DOKUMEN BERITA BERBAHASA INDONESIA MENGGUNAKAN METODE *KNN*

Achmad Ridok[1]
[1]*Program Studi Ilmu Komputer, Jurusan Matematika, Fakultas MIPA*
Universitas Brawijaya Malang
Jl.MT. Haryono 169 Malang
E-mail: acridokb@ub.ac.id

ABSTRAK
Pembuatan judul otomatis adalah proses untuk menghasilkan judul dari suatu artikel dokumen secara otomatis. Dengan dibangkitkannya judul secara otomatis pembaca dapat menangkap ide utama dari sebuah dokumen tanpa harus membaca keseluruhan dokumen. Pada penelitian ini dilakukan penerapan metode K-Nearest Neighbor (KNN) untuk pembuatan judul otomatis dokumen berita berbahasa Indonesia. Hasil pengujian dan evaluasi menunjukkan bahwa sistem pembuatan judul ini menghasilkan kinerja terbaik pada katagori politik dengan nilai rata-rata pesisi sebesar 0.319, nilai rata-rata recall sebesar 0.321 dan nilai rata-rata F-measure sebesar 0.311. Sistem dengan metode KNN ini memiliki kelemahan, yaitu sangat bergantung pada data latih dan tidak dapat membuat judul baru, sehingga terdapat judul bentukan sistem yang kurang mencerminkan isi dari dokumen yang diujikan

Kata kunci: KNN, Pemrosesan Teks, Judul Otomatis

1. PENDAHULUAN

Dengan semakin berkembangnya teknologi komputer yang semakin cepat, mampu membawa perubahan yang besar dalam kehidupan manusia. Salah satunya adalah kemudahan dalam mendapatkan informasi baik dalam bentuk artikel, teks, gambar maupun suara. Sering kali teks-teks yang didapatkan terdiri atas rangkaian kalimat yang cukup panjang dan belum tentu juga informasi yang terkandung di dalamnya bermanfaat atau sesuai dengan yang dibutuhkan. Oleh karena itu, sebuah judul sangatlah penting bagi pembaca untuk memahami informasi penting yang ada dalam sebuah dokumen dengan cepat.

Judul merupakan representasi yang dapat membantu pembaca untuk menangkap ide utama dari sebuah dokumen tanpa harus membaca keseluruhan dokumen tersebut (Jin, 2003). Menurut Jin dan Hauptmann (2001), dalam membuat judul dari sebuah dokumen melibatkan beberapa tugas yang kompleks, pertama harus memahami isi dari dokumen, kedua harus mengetahui karakteristik dokumen dalam kaitannya dengan dokumen lain, ketiga harus mengetahui bagaimana membuat judul yang mempunyai bunyi yang baik agar dapat menarik pembaca dan bagaimana menyaring intisari dokumen ke dalam judul dalam beberapa kata

Penelitian mengenai pembuatan judul otomatis untuk bahasa Indonesia pernah dilakukan dengan teknik pendekatan statistik menggunakan metode *Naïve Bayes* oleh Ridok (2009). Dalam penelitian tersebut, sebuah kerangka statistik digunakan untuk membangkitkan sebuah judul dari suatu dokumen yang dibagi menjadi dua fase, yaitu fase *title word selection* (pemilihan kata judul) dan fase *title word ordering* (pengurutan kata judul). Pada fase *title word selection*, setiap kata di dalam judul akan diberi skor berdasarkan ada tidaknya kata tersebut dalam dokumen dengan menggunakan metode *Naïve Bayes*. Sedangkan pada fase *title word ordering*, kecocokan dari urutan kata dalam judul dihitung menggunakan model statistik *n-gram*. Urutan kata hasil dengan skor tertinggi pada fase *title word selection* dan *title word ordering* akan dipilih sebagai judul dari dokumen tersebut.

Pada penelitian kali ini, akan dilakukan penerapan metode *K-Nearest Neighbor* (*KNN*) dalam menyelesaikan permasalahan pembuatan judul otomatis dengan pendekatan statistik. *KNN* merupakan salah satu metode pembelajaran dan mempunyai performansi yang sangat baik untuk pengkategorian teks otomatis (*automatic text categorization*) (Yang dan Chute, 1994). Untuk merepresentasikan kata unik (*term*) dan keseluruhan dokumen digunakan *Vector Space Model* (*VSM*). Pada proses klasifikasi menggunakan metode *K-Nearest Neighbor* (*KNN*), dokumen-dokumen dikelompokkan ke dalam kategori yang sesuai dengan isi dokumen yang didasarkan pada perhitungan bobot *silmilarity* (kemiripan) antar dokumen. Bobot kemiripan tersebut dihitung dengan metode *cosine similarity*. Untuk membangkitkan judul pada dokumen uji, bobot *similarity* yang tertinggi antara data latih dan dokumen uji tersebut yang akan dipilih menjadi *output* judul untuk dokumen uji.

2. TINJAUAN PUSTAKA

2.1 Pembutan Judul Otomatis

Judul merupakan representasi yang dapat membantu pembaca untuk menangkap ide utama dari sebuah dokumen tanpa harus membaca keseluruhan dokumen tersebut (Jin, 2003).

Menurut Jin dan Hauptmann (2001), dalam membuat judul dari sebuah dokumen melibatkan beberapa tugas yang kompleks, pertama harus memahami isi dari dokumen, kedua harus mengetahui karakteristik dokumen dalam kaitannya dengan dokumen lain, ketiga harus mengetahui bagaimana membuat judul yang mempunyai bunyi yang baik agar dapat menarik pembaca dan bagaimana menyaring intisari dokumen ke dalam judul dalam beberapa kata.

Pembentukan judul secara otomatis (*Automatic Title Generation*) dapat dikategorikan ke dalam dua teknik, yaitu teknik peringkasan teks otomatis (*automatic text summarization*) dan teknik pendekatan statistik. Pada teknik peringkasan teks otomatis, judul dianggap sebagai ringkasan yang sangat pendek dan menggunakan teknik peringkasan teks secara langsung untuk membentuk judul. Sedangkan pada pendekatan statistik menekankan ide pembelajaran korelasi antara kata-kata dalam judul (*title words*) dan kata-kata yang menyusun dokumen (*document words*) dari *training corpus* (data pelatihan) dan menerapkan model tersebut untuk membuat judul pada data pengujian (Jin, 2003).

Terdapat beberapa metode statistik yang dapat digunakan untuk pembuatan judul, di antaranya adalah metode *A Nearest Neighbor* (*NN*), metode *K-Nearest Neighbor* (*KNN*), metode *Decision Tree*, metode *Statistical Translation*, metode *Reverse Information Retrieval*, metode *Naïve Bayes Approach with Limited Vocabulary* (*NBL*), dan metode *Naïve Bayes Approach with Full Vocabulary* (*NBF*) (Jin, 2003).

2.2 Text Preprocessing

Preprocessing adalah proses yang dilakukan untuk mempersiapkan dokumen baik dokumen latih maupun dokumen uji sebelum siap diolah. Guo (2004) menyatakan bahwa *preprocessing* data mengimplementasikan fungsi untuk memindahkan dokumen awal ke dalam representasi yang dapat diimplementasikan pada dokumen latih dan dokumen uji.

Menurut Sebastiani (2002), alasan dilakukannya *preprocessing* adalah karena dokumen teks tidak dapat diterjemahkan secara langsung oleh algoritma pembentuk hasil pencarian. Penerjemahan ini bertujuan untuk menghasilkan data numerik yang mudah diakses, karena data numerik tersebut yang dapat digunakan untuk perhitungan lebih lanjut.

Secara umum, langkah-langkah dalam *text preprocessing* meliputi beberapa tahap, yaitu:

1. *Case Folding*

Case folding adalah proses pengubahan semua huruf dalam dokumen menjadi huruf kecil. Sedangkan karakter selain huruf, seperti tanda baca dan angka dihilangkan dan dianggap sebagai *delimiter*.

2. *Tokenizing*

Tokenizing adalah proses untuk pemotongan (*parsing*) tiap-tiap kata yang menyusun dokumen menjadi kata tunggal.

3. *Filtering*

Filtering adalah proses mengambil kata-kata yang penting dan penghilangan *stopword* yang terdapat pada dokumen. Daftar *stopwords* yang digunakan diambil dari hasil penelitian yang dilakukan oleh Tala (2003).

4. *Term Weighting* (Pembobotan)

Metode yang digunakan untuk melakukan pembobotan terhadap *term* adalah pembobotan *TFIDF*. *TF* (*Term Frequency*) adalah pembobotan kata (*term*) yang yang didasarkan pada perhitungan jumlah kata yang muncul pada suatu dokumen. *IDF* (*Inverse Document Frequency*) adalah pembobotan kata (*term*) yang didasarkan pada perhitungan jumlah kata yang muncul pada seluruh dokumen. *TFIDF* merupakan perkalian dari hasil perhitungan *TF* dengan hasil perhitungan *IDF*. Persamaan untuk melakukan perhitungan bobot *TFIDF* adalah :

$$W = TFxIDF \quad (1)$$

$$w(t,d) = FT(t,d)xlog\frac{D}{DF_t} \quad (2)$$

dimana :

w(t,d) : bobot *term t* pada dokumen *d*
TF(t,d) : jumlah kemunculan *term t* dalam dokumen *d*
D : jumlah seluruh dokumen
DFt : jumlah dokumen yang memiliki *term t*

5. *Vector Space Model* (*VSM*)

Pada *vector space model*, setiap dokumen direpresentasikan sebagai sebuah vektor di mana tiap komponennya diasosiasikan dengan kata tertentu pada kumpulan kosakata yang dimiliki oleh dokumen tersebut (Soucy dan Mineau, 2003).

2.3 Algoritma *KNN* untuk Pembentukan Judul Otomatis

Algoritma *KNN*

K-Nearest Neighbor, yang disingkat *KNN*, adalah sebuah algoritma yang cukup populer untuk melakukan pengkategorian teks dan merupakan salah satu metode terbaik dalam bidang tersebut (Bergo, 2001).

KNN adalah sebuah metode untuk melakukan klasifikasi terhadap obyek berdasarkan data latih yang jaraknya paling dekat dengan obyek tersebut (Kozma, 2008).

Langkah awal yang dilakukan dalam *KNN* adalah menghitung nilai kemiripan antara dokumen tes dengan semua data latih menggunakan *cosine similarity* dengan rumus persamaan (Garcia, 2005):

$$Sim\,(Q,D) = \cos\,(Q,D) = \frac{Q \bullet D}{|Q|,\ |D|} \qquad (3)$$

dimana :

Q.D :hasil perkalian dalam (*inner product*/*scalar product*/*dot product*) kedua vektor

|Q|.|D| : panjang vektor (jarak *euclidean*) suatu vektor dengan titik nol

KNN Decision Rule

KNN Decision Rule adalah proses pengambilan keputusan yang digunakan pada *KNN*. Pada proses pengambilan keputusan diperlukan suatu nilai *k* yang akan digunakan untuk memilih kategori yang sesuai.

Untuk *k* = 1, dipilih nilai *similarity* pada urutan paling atas, dengan persamaan :

$$SIM_{\max}(X) = \max_{d \in T} SIM(X, d_i) \qquad (4)$$

Untuk *k* > 1, penentuan kategorinya adalah dengan menjumlahkan semua nilai kemiripan $SIM(X,d_i)$ yang termasuk dalam suatu kategori, dengan persamaan :

$$P(X, C_j) = \sum_{d_i \in KNNofX} SIM(X, d_i) \cdot y(d_i, C_j) \qquad (5)$$

dimana :

$P(X,C_j)$: probabilitas dokumen *X* menjadi anggota kategori *Cj*.

$SIM(X,d_i)$: kemiripan antara dokumen *X* dengan dokumen latih *d*.

$y(d_i,C_j)$: fungsi atribut dari sebuah kategori yang memenuhi

$$y(d_i, C_j) = \begin{cases} 1, & d_i \in C_j \\ 0, & d_i \notin C_j \end{cases} \qquad (6)$$

Proses Algoritma *KNN*

1. Mengolah dokumen *X* sehingga menjadi bentuk vektor (*X1*, *X2*,.., *Xm*) seperti semua data latih.
2. Mengitung kemiripan antara dokumen *X* dengan seluruh data latih.
3. Memilih *k* sampel dengan nilai-nilai $SIM(X,d_i)$ tertinggi, dan dianggap sebagai anggota himpunan *KNN* dari *X*. Kemudian, menghitung probabilitas *X* menjadi anggota tiap kategori secara berturut-turut.
4. Dokumen *X* masuk ke dalam kategori yang memiliki nilai $P(X,C_j)$ paling besar.

Proses Pembentukan Judul

1. Mengklasifikasi dokumen *X* menggunakan algoritma *KNN* untuk mengetahui dokumen tersebut masuk ke dalam kategori *Cj*.
2. Mencari nilai kemiripan antara dokumen *X* dengan seluruh data latih di dalam kategori *Cj* yang sudah dihitung saat klasifikasi.
3. Judul pada data latih dengan nilai tertinggi akan menjadi judul untuk dokumen *X*.

2.4 Tipe Evaluasi

Untuk pembentukan judul otomatis, digunakan perhitungan *precision* dan *recall. Precision* adalah nilai yang menunjukkan tingkat ketelitian sistem dalam membuat judul yang sesuai.

$$precision = \frac{kata\ judul\ sistem \cap kata\ judul\ senarnya}{\sum kata\ judul\ sistem} \qquad (7)$$

Sedangkan *recall* adalah nilai yang menunjukkan tingkat ketepatan sistem dalam membuat judul yang sesuai.

$$recall = \frac{kata\ judul\ sistem \cap kata\ judul\ sebenarnya}{\sum kata\ judul\ sebenarnya} \qquad (8)$$

Kombinasi antara nilai *precision* dan *recall* menghasilkan *F-measure* yang dapat didefinisikan dengan persamaan :

$$F_{measure} = \frac{2 \times recall \times precision}{recall + precision} \qquad (9)$$

F-measure memberikan penekanan yang sama atau pengaruh yang relatif untuk *precision* dan *recall*. Ketika *precision* dan *recall* bernilai kecil, maka nilai *F-measure* akan menjadi kecil. Begitu juga sebaliknya, nilai *F-measure* tinggi ketika *precision* dan *recall* bernilai besar, dan akan mencapai nilai maksimum 1 hanya ketika *precision* dan *recall* juga mencapai nilai maksimum, yaitu 1 (Jin, 2003).

3. METODOLOGI

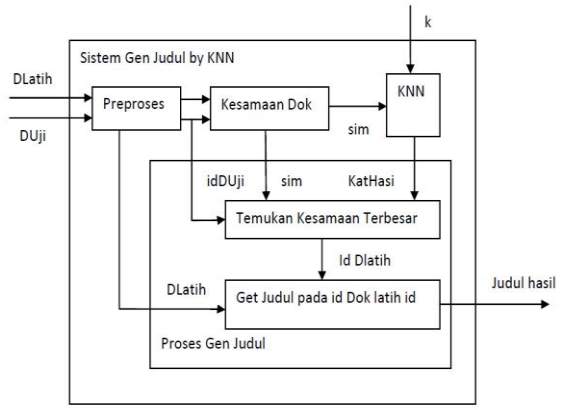
Dalam penelitian ini, akan dibangun suatu perangkat lunak yang dapat digunakan untuk membuat judul secara otomatis pada teks dokumen yang berupa file berformat teks (**.txt*) sebagai masukan. Metode yang digunakan dalam sistem ini adalah *K-Nearest Neighbor* (*KNN*).

Dalam sistem ini dibutuhkan dua jenis data, yaitu data latih dan data uji. Data uji merupakan teks dokumen yang akan dibuat judulnya. Sedangkan data latih merupakan teks dokumen lengkap dengan judulnya yang akan dijadikan pembanding data uji sehingga dapat dibuat judulnya.

Langkah-langkah proses sistem pembuatan judul otomatis ditunjukkan pada Gambar 1.

Perancangan Uji Coba

Pada pengujian sistem, sekumpulan dokumen akan dibagi menjadi dokumen latih dan dokumen uji. Selain itu juga dilakukan uji coba untuk semua dokumen yang isinya (*content*) berbeda, baik untuk dokumen latih maupun dokumen uji.



Gambar 1. Arsitektur Sistem Pembuatan Judul

4. PEMBAHASAN

Skenario Evaluasi

Evaluasi sistem dilakukan dengan menggunakan dokumen latih dan dokumen uji yang masing-masing telah diketahui judulnya dan katagorinya. Jumlah dokumen latih dan dokumen uji masing-masing sebanyak 178 dan 125 dokumen yang berasal dari 9 katagori yang berbeda. Seluruh data diambil dari www.kompas.com. Skenario evaluasi dilakukan dalam dua tahap, pertama evaluasi untuk proses klasifikasi dan kedua untuk proses pembentukan judul untuk setiap katagori.

Hasil Evaluasi

Hasil uji coba terhadap sistem dilakukan untuk mengetahui kinerja sistem yang dibangun. Evaluasi tersebut dilakukan dengan membandingkan hasil pembentukan judul yang dilakukan oleh sistem dengan judul asal dari sumber berita. Selanjutkan judul hasil dibandingkan dengan judul asal dari dokumen uji. Sebagian hasil perbandingan antara judul asal dengan judul bentukan system untuk katagori ekonomi ditunjukkan pada Tabel 1.

Tabel 1. Tabel hasil perbandingan judul

Judul Asal	**Judul hasil**
BI Inflasi Februari Membaik	Inflasi Belum Mengkhawatirkan
Bi setujui merger uob buana dan uob Indonesia	Bank mandiri incar kredit konsumer rp 2995 triliun di 2010
Bi siapkan aturan arus keluar masuk modal	Bi rate terus ditahan di 65
Cimb group finalisasi proses dual listing	Cimb niaga kucuran kredit sindikasi rp 15 triliun
Dana asing di surat utang ri kembali tembus rp 144 triliun	Pemerintah Turunkan Target Penerimaan Pajak
Ekspor tekstil tumbuh 10 di semester i 2010	Bi rate terus ditahan di 65
Ekspor China Melaju Yuan Akan Direvaluasi	Bi rate terus ditahan di 65
Hingga Hari Ini Penerimaan Pajak Capai Rp 109 Triliun	Pemerintah Turunkan Target Penerimaan Pajak
Oktober bea cukai jadikan satu 17 dokumen kepabeanan	Penerimaan Bea Cukai Capai Rp 18 301 Triliun
Penguatan ihsg sokong rupiah	Rupiah menawan di level rp9 200 an
Rupiah dibuka bergairah di level rp9 237	Rupiah awal pekan kian berjaya di rp9 175
Rupiah dan Saham Tertahan	Rupiah Perkasa IHSG Tetap Kepentok
Rupiah Masih di Atas 9 300	Harga Emas Akhirnya Turun di Bawah 1 100
Saham di Jalur Negatif	IHSG Melorot 16 Poin
Saham Langsung Merah	Tenggelam IHSG Rontok 74 Poin
Subsidi listrik bisa bengkak rp 5 triliun	Juli Tarif Listrik Bakal Naik Lagi
Tdl tak naik apbn terancam jebol	Juli Tarif Listrik Bakal Naik Lagi
Transaksi sepi ihsg mampu tembus 2 826	IHSG Ditutup Merosot 28 Persen
Turun Minyak di Asia Masih di Atas 82 Dollar AS	Harga Emas Akhirnya Turun di Bawah 1 100
Wall Street Menguat Saham Apple Sentuh Rekor	Rupiah dan Saham Malas malasan
Besok 17 PB Akan Tanda Tangani MoU	kurangi defisit ri belum bisa ketatkan anggaran
Dana masih Jadi Kendala Kontingen Indonesia	menkeu restui data wajib pajak terkait gayus dibuka

Keberhasilan sistem dalam membuat judul otomatis suatu dokumen dievaluasi menggunakan *precision* dan *recall* sebagaimana persamaan 7 dan 8. Sebelum evaluasi pembentukan judul, terlebih dahulu sistem dievaluasi pada nilai parameter k yang berbeda untuk menentukan presisi dan recall terbaik dalam hal melakukan klasifikasi. Dari uji coba yang dilakukan, hasil klasifikasi sistem dalam penentuan suatu katagori dapat diperoleh nilai kinerja terbaik pada k = 25 sebagaimana dapat dilihat pada table 2.

Selanjutnya berdasarkan hasil evaluasi klasifikasi pada k optimal, nilai katagori yang dihasilkan untuk masing-masing dokumen uji dijadikan dasar penentuan judul dengan mencai judul dari dokumen latih yang berkatagori hasil terbaik yang mempunyai kesamaan terbesar dengan dokumen uji.

Tabel 2. Tabel hasil evaluasi klasifikasi

Katagori	Presisi	Recall
edukasi	1.000	1.000
ekonomi	0.952	0.870
kesehatan	0.929	0.867
olahraga	0.895	1.000
otomotif	1.000	1.000
politik	0.909	0.952
sains	0.762	0.941
teknologi	1.000	0.900
travel	1.000	0.700

Hasil evaluasi rata-rata sistem dalam melakukan pembentukan judul untuk masing-masing katagori dapat dilihat pada tabel 3.

Tabel 3. Rata-rata Hasil Evaluasi Gen Judul

Katagori	Presisi	Recall	Fm
Edukasi	0.200	0.195	0.195
Ekonomi	0.212	0.199	0.199
Kesehatan	0.218	0.204	0.202
Olahraga	0.247	0.246	0.241
Otomotif	0.210	0.191	0.198
Politik	0.319	0.321	0.311
Sain	0.178	0.210	0.188
Teknologi	0.196	0.220	0.204
Travel	0.105	0.120	0.112

Berdasarkan tabel 3 di atas, keberhasilan sistem dalam pembentukan judul menggunakan metode KNN mempunyai presisi, recall dan Fm terbaik pada katagori politik yang masing-masing bernilai 0.319, 0.321 dan 0,311.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Sistem pembuatan judul otomatis dokumen berita berbahasa Indonesia dengan metode *K-Nearest Neighbor* (*KNN*) dapat melakukan pembentukan judul secara otomatis suatu dokumen dengan spesifikasi sebagai berikut :

1. Sistem ini menghasilkan kinerja terbaik pada katagori politk dengan nilai rata-rata *precision* sebesar 0.319, nilai rata-rata *recall* sebesar 0.321 dan nilai rata-rata F-measure sebesar 0.311.
2. Sistem tidak dapat menghasilkan judul baru dan sangat bergantung terhadap data latih. Hal ini merupakan kelemahan dari sistem.

5.2 Saran

Perlu dikembangkan efektifitas sistem dengan mengadopsi kemampuan metode Naïve Bayes yang tidak hanya mempertimbangkan isi dari judul tetapi juga memperhatikan isi dari dokumen.

DAFTAR PUSTAKA

Bergo, Alexander. 2001. *Text Categorization and Prototypes*.

Garcia, Dr. E. 2005. *The Classic Vector Space Model (Description, Advantages and Limitations of the Classic Vector Space Model)*.

Guo, G., Hui Wang, David Bell, Yaxin Bi, dan Kieran Greer. 2004. *An KNN Model Based Approach and Its Application in Text Categorization*. Northern Ireland, UK.

Jin, Rong. 2003. *Statistical Approachs Toward Title Generation*. A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy. Carnegie Mellon University.

Jin, Rong dan Hauptmann, A. G. 2001. *Learning to Select Good Title Word: A New Approach based on Reverse Information Retrieval*. Carnegie Mellon University, Pittsburgh, USA.

Kozma, Laszlo. 2008. *K Nearest Neighbors Algorithm (KNN)*. T-61.6020 Special Course in Computer and Information Science. Helsinki University of Technology.

Ridok, A. 2009. *Pembentukan Judul Suatu Dokumen Secara Otomatis Menggunakan Metode NBL*. Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Brawijaya, Malang.

Sebastiani, F. 2002. *Machine Learning in Automated Text Categorization*.

Tala, Fadillah Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Master of Logic Project Institute for Logic, Language and Computation. Universiteit van Amsterdam, The Netherlands.

Yang, Y. dan Chute, C. G. 1994. An example-based mapping method for text classification and retrieval, ACM Transactions on Information Systems (TOIS