PENGGUNAAN FITUR ABSTRAKSI DAN CATATAN PUBLIKASI PENULIS UNTUK KLASIFIKASI ARTIKEL ILMIAH DENGAN METADATA YANG TERBATAS

Halimatus Sa'dyah, Nurissaidah Ulinnuha
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya 60111
Email: aisyah2nd@gmail.com

ABSTRAK

Bertumbuhnya jumlah artikel ilmiah membuka ranah penelitian baru di bidang optimasi klasifikasi dokumen berbasis metadata. Dalam ranah ini, persoalan pokok yang harus dijawab adalah bagaimana cara memanfaatkan fitur metadata yang terbatas untuk menghasilkan nilai presisi dan recall yang tinggi dalam proses klasifikasi artikel ilmiah. Dalam makalah ini diusulkan sebuah metode klasifikasi artikel ilmiah dengan menggunakan atribut abstraksi dan catatan publikasi penulis pada metada data sebagai fitur. Hasil uji coba menunjukkan bahwa sistem klasifikasi yang menggunakan abstraksi dan catatan publikasi penulis sebagai fitur menghasilkan nilai presisi tertinggi sebesar 0.87 dan recall 0.59 sedangkan sistem klasifikasi yang menggunakan abstraksi sebagai fitur menghasilkan nilai presisi 0.75 dan recall 0.51. Hasil uji coba juga menunjukkan bahwa nilai presisi dan recall dari sistem klasifikasi stabil ketika nilai $\gamma\_1/\gamma\_2 = 1$. *Berdasarkan hasil uji coba ini, dapat disimpulkan bahwa sistem klasifikasi yang diusulkan lebih baik dalam hal presisi dan recall jika dibandingkan dengan sistem klasifikasi yang menggunakan abstraksi saja. Selain itu, juga dapat disimpulkan bahwa abstraksi dan catatan publikasi artikel ilmiah memiliki nilai signifikansi yang sama.*

***Kata Kunci**:* Artikel ilmiah, Klasifikasi, Metadata.

1. PENDAHULUAN

Saatini, kebutuhan terhadap artikel ilmiah semakin bertambah.Sementara itu, tidak semua jurnal ilmiah menyediakan dokumen artikelnya secara gratis sehingga kesesuaian antara dokumen jurnal ilmiah yang dicari oleh pengguna dengan dokumen yang dikembalikan oleh mesin pencari menjadi satu hal yang penting.

Klasifikasi dokumen artikel ilmiah berdasarkan bidang keilmuan dapat membantu pengguna untuk mengunduh dokumen artikel ilmiah secara tepat dan akurat.Sayangnya, informasi lengkap mengenai isi artikel ilmiah biasanya tidak tersedia dalam jurnal ilmiah online yang berbayar. Pada jurnal ilmiah online yang berbayar, informasi yang tersedia biasanya hanya sebatas metadata yang terdiri dariabstraksi, nama penulis artikel, referensi yang dirujuk di dalam artikel, tahun terbit artikel, nomor halaman artikel di dalam jurnal, serta nama jurnal yang mempublikasikan artikel ilmiah tersebut.

Dalam bidang sistem temu kembali informasi, tidak banyak penelitian yang fokus meneliti dokumen artikel ilmiah.Padahal, klasifikasi dokumen artikel ilmiah memiliki karakteristik yang berbeda dengan klasifikasi dokumen biasa. Pada klasifikasi dokumen biasa, semua informasi mengenai isi dokumen tersedia lengkap sehingga tantangan terbesar bagi peneliti di bidang ini selain penggalian topik dari dokumen adalah waktu komputasi klasifikasi.Pada dokumen artikel ilmiah, informasi dari isi dokumen diwakili oleh abstraksi danmetadata. Hal ini menyebabkan tantangan utama dari klasifikasi dokumen karya ilmiahadalah bagaimana memanfaatkan fitur metadata yang ada untuk menghasilkan presisi dan recall yang baik.

Penggunaan abstraksi dokumen artikel ilmiah sebagai fitur tunggal dalam klasifikasi biasanya menghasilkan presisi dan *recall* yang belum maksimal. Salah satu penyebabnya adalah banyaknya topik yang tersembunyi dari dokumen lengkap artikel ilmiah yang tidak bisa dideteksi melalui abstraksi.

Untuk mengurangi kesalahan klasifikasi akibat kurangnya informasi tersebut, dalam makalah ini diusulkan sistem klasifikasi dokumen artikel ilmiah yang mempertimbangkan catatan publikasi artikel ilmiah oleh penulis artikel dalam jurnal ilmiah.Hal ini dilakukan agar artikel tersebut dapat masuk ke dalam kategori yang benar sesuai dengan bidang keahlian penulisnya.

Makalah ini disusun dengan urutan sebagai berikut, bagian pertama menjelaskan pendahuluan, bagian kedua menjelaskan tentang metode yang diusulkan beserta latar belakangnya, bagian ketiga tentang skema pembobotan TF-IDF, bagian keempat menjelaskan tentang algoritma *Support Vector Machine*, bagian kelima menjelaskan hipotesa pembobotan bertingkat, bagian keenam menjelaskan

metodologi penelitian, bagian ketujuh menjelaskan hasil eksperimen serta evaluasinya dan bagian kedelapan menjelaskan kesimpulan.

2. KLASIFIKASI DOKUMEN ARTIKEL ILMIAH

Tidak banyak artikel ilmiah yang membahas sistem klasifikasi dokumen artikel ilmiah.Terdapat tiga macam metode yang pernah diusulkan untuk menyelesaikan permasalahan klasifikasi artikel ilmiah ini.Montejo dan timnya mengusulkan metode klasifikasi multilabel menggunakan fitur kata kunci yang diekstraksi dari abstraksi artikel ilmiah [1]. Setelah itu, Denecke dan timnya mengusulkan nama penulis serta kata kunci yang diekstraksi dari judul artikel ilmiah untuk digunakan sebagai fitur klasifikasi dokumen [2].

Montejo dan Denecke memfokuskan penelitiannya pada pemilihan fitur abstraksi artikel ilmiah.Penelitian dengan topik sejenis juga dilakukan oleh Martin dan timnya.Namun dalam hal ini, Martin berfokus pada metode penggalian topik [3].Dalam penelitiannya, Martin dan timnya mengusulkan metode *Latent Dirichlet Allocation* (LDA)untuk mengekstraksi topik dari abstraksi dokumen artikel ilmiah [3].

Metode yang diusulkan oleh Martin dan timnya menghasilkan presisi dan *recall* yang tinggi untuk dokumen yang bersifat heterogen (dokumen yang berbeda kelastidak banyak mengandung kata yang sama). Namun, nilai presisi dan *recall* ini akan turun ketikadokumen yang diekstraksi bersifat homogen (dokumen yang berbeda kelas banyak mengandung kata yang sama). Hal ini menunjukkan bahwa penggunaan abstraksi artikel ilmiah sebagai fitur tunggal pada sistem klasifikasi dokumen artikel ilmiah tidak dapat memberikan informasi yang cukup tentang topik artikel ilmiah.

Untuk menghasilkan presisi dan *recall*yang tinggi pada sistem klasifikasi dokumen artikel ilmiah yang bersifat homogen, dalam makalah ini diusulkan sistem klasifikasi dokumen artikel ilmiah dengan metodeperhitungan skor bobot yang mempertimbangkan abstraksi dan catatan publikasi penulis.Seorang peneliti yang telah mempublikasikan hasil penelitiannya ke dalam jurnal pada umumnya sudah fokus melakukan penelitian dalam bidang tertentu.Oleh karena itu, catatan publikasi penulis dapat memberikan informasi tambahan tentang topik dari artikel ilmiah yang sedang diklasifikasikan.

Adapun atribut dari catatan publikasi penulis yang digunakan untuk memperkaya informasi bagi sistem klasifikasi dokumen artikel ilmiah adalah atribut judul.Atribut judul digunakan karena kata-kata dalam judul cukup mewakili informasi tentang bidang penelitian yang sedang ditekuni oleh penulis artikel ilmiah.

Dalam penelitian ini, tidak semua judul artikel ilmiah yang terdapat dalam catatan publikasi penulis digunakan sebagai fitur klasifikasi.Pada penelitian ini, judul artikel ilmiah yang digunakan sebagai fitur klasifikasi dipilih berdasarkan tingkat kemiripannya dengan judul artikel ilmiah yang sedang diklasifikasikan. Kemiripan judul dalam penelitian ini diukur menggunakan koefisienJaccard yang didefinisikan pada persamaan 1 sebagai berikut:

$$Jaccard(A,B) = \frac{n\,(A \cap B)}{n\,(A \cup B)} \qquad (1)$$

dimana $Jaccard(A,B)$ adalah nilai koefisien *Jaccard* antara dokumen A dan dokumen B, $n\,(A \cap B)$ adalah jumlah katadalam dokumen A yang beririsan dengan katadalam dokumen B, sedangkan $n\,(A \cup B)$ adalah jumlah gabungan kata yang ada dalam dokumen A dengan kata yang ada dalam dokumen B [4].

Ketentuan pemilihan judul ini didasari oleh asumsi bahwa seorang peneliti bisa bekerja di dua bidang yang berbeda sehingga pemilihan judul tanpa mempertimbangkan kemiripan juga berpotensi menurunkan nilai presisi dan *recall* dalam proses klasifikasi.

3. PEMBOBOTAN TF-IDF

TF-IDF adalah singkatan dari *Term Frequency-Inverted Document Frequency*. *Term frequency* (TF) adalah frekuensi dari kemunculan sebuah kata dalam dokumen yang bersangkutan sedangkan *Inverse document frequency* (IDF) adalah sebuah pembobotan statistik global yang mengkarakteristikkan sebuah kata dalam keseluruhan koleksi dokumen.

IDF merupakan sebuah perhitungan dari bagaimana kata didistribusikan secara luas pada koleksi. Nilai IDF berbanding terbalik dengan tingkat kemunculan sebuah kata dalam koleksi. Semakin sering sebuah kata muncul dalam koleksi, maka nilai IDF dari kata tersebut semakin kecil. Begitu juga sebaliknya. Nilai bobot TF-IDF didefinisikan pada persamaan 2 sebagai berikut:

$$w(t_i, d_j) = \frac{count\ (t_i,d_j).log\ (|corpus\ |)}{count\ \_doc\ (t_i,corpus\ )} \qquad (2)$$

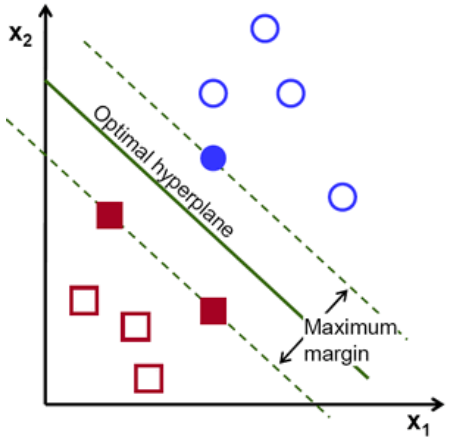
dimana$w(t_i, d_j)$ adalah nilai bobot TF-IDF dari kata *i* dalam dokumen *j*, $count(t_i, d_j)$ adalah jumlah kata *i* dalam dokumen *j*, $corpus$ adalah jumlah seluruh dokumen yang ada dalam koleksi dan$count_doc(t_i, corpus)$ adalah jumlah dokumen dalam koleksi yang mengandung kata*i*[5].

Penghitungan bobot dari kata tertentu dalam sebuah dokumen dengan menggunakan TF-IDF menunjukkan bahwa deskripsi terbaik dari dokumen adalah kata yang banyak muncul dalam dokumen

tersebut dan sangat sedikit muncul pada dokumen yang lain.

4. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) adalah sebuah algoritma klasifikasi yang bekerja dengan prinsip mencari *hyperplane* atau garis pemisah terbaik bagi dua buah kelas pada *input space*. Kriteria dari *hyperplane terbaik*untuk dua buah kelas adalah *hyperplane* yang mempunyaimarginterbesar. Marginadalah jarak antara *hyperplane* dengan *pattern* terdekat dari masing-masing kelas.*Pattern* yang paling dekat ini disebut sebagai supportvector.Ilustrasi dari prinsip kerja SVM dapat dilihat pada Gambar 1.



Gambar 1.Konsep Dasar SVM [6].

Untuk mencari *hyperplane*terbaik, algoritma SVM memiliki fungsi objektif yang dapat dilihat dalam persamaan 3 dimana *x*adalah data yang tersedia, *y* adalah kelas, α adalah *Lagrange Multiplier* untuk setiap data *i* dan $k(x_i,x_j)$ adalah fungsi kernel dalam algoritma SVM [4].

$$max_x \left(\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{n} y_i\, y_j\, \alpha_i \alpha_j\, k(x_i, x_j) \right) \quad (3).$$

5. HIPOTESA PEMBOBOTAN BERTINGKAT

Pada pembahasan sebelumnya, telah dijelaskan tentang fitur yang diusulkan dalam sistem klasifikasi pada makalah ini. Dalam mengusulkan metode ini, hipotesa yang akan dibuktikan adalah abstraksi artikel dan catatan publikasi dari penulis artikel memiliki tingkat signifikansi yang berbeda terhadap atribut kelas dalam proses klasifikasi artikel karya ilmiah.

Oleh karena itu, dalam makalah ini dilakukan uji coba klasifikasi artikel ilmiah dengan skema pembobotan bertingkat. Adapun rumus perhitungan bobot dalam skema pembobotan bertingkat adalah sebagai berikut:

$$w_{(term)} = \gamma_1 . w_{(term,abstrak)} + \gamma_2 . \sum_{i=0}^{i=max} w_{(term,judul)} \quad (4)$$

dimana $w_{(term)}$ adalah bobot TF-IDF total dari sebuah kata, $w_{(term,abstrak)}$adalah bobot TF-IDF sebuah katateрhadap abstrak, $w_{(term,judul)}$adalah bobot TF-IDF sebuah kata terhadap judul-judul artikel ilmiah yang didapatkan dari catatan publikasi penulis, sedangkan γ_1 dan $\gamma_2$adalah sebuah konstanta yang menyatakan bobot dari abstrak dan catatan publikasi ilmiah penulis.Jumlah konstanta $\gamma_1$dengan $\gamma_2$adalah satu.

6. METODOLOGI KLASIFIKASI

Pada bagian ini akan dijelaskan metodologi klasifikasi dokumen artikel dari jurnal ilmiah berdasarkan metadata. Metodologi tersebut meliputi data masukan, persiapan data hingga metode yang digunakan pada sistem klasifikasi.

6.1 Data masukan

Data masukan untuk sistem klasifikasi ini adalah abstraksi dan catatan publikasi penulis.Pada sistem ini, abstraksi dan catatan publikasi penulis diperlakukan sebagai data bertipe string.

6.2 Tahap persiapan data

Tahap persiapan data bertujuan untuk mempersiapkan data agar formatnya sesuai dengan kebutuhan sistem klasifikasi. Proses-proses yang berjalan dalam tahap ini adalahtokenisasi, penghapusan *stopword*, *stemming*, dan penghitungan bobot untuk setiap kata yang ada dalam koleksi dokumen (data masukan).

Tokenisasi adalah suatu proses yang bertujuan untuk membagi data masukan yang formatnya masih berupa teks panjang menjadiunit-unit kecil yang disebut token. Token dalam konteks dokumen dapat berupa suatu kata, angka, atau tanda baca.

Setelah proses tokenisasi dilakukan, proses selanjutnya adalah penghapusan *stop word*. Dalam sebuah dokumen terdapat banyak kata yang tidak memberikan makna, misalnya kata hubung, kata depan, dan lain-lain. Kata-kata jenis ini diistilahkan dengan *stopword*. Agar tidak menimbulkan kerancuan dalam proses pengolahan dokumen, *stop word* harus dibuang.

Setelah semua *stop word* dihapus, kata-katayang dihasilkan dari proses tokenisasi harus dimasukkan dalam proses *stemming*. *Stemming* adalah teknik yang digunakan untuk menemukan kata dasar dari sebuah kata berimbuhan dalam sebuah dokumen. Hal ini didasarkan pada fakta bahwa kata-kata yang memilikibentukdasar yang sama akan mendeskripsikan makna yang sama atau relatif dekat [7].Misalnya pada kata penggunaan, menggunakan, digunakan dan berguna memiliki bentuk dasar yang sama yaitu guna.

Setelah semua kata diubah ke dalam bentuk dasarnya, langkah selanjutnya adalah penghitungan bobot untuk setiap kata.Pada dasarnya, proses penghitungan bobot bertujuan untuk menilai relevansi kata yang terdapat pada sebuah dokumen dengan keseluruhan dokumen yang terdapat dalam koleksi. Setiap kata diberikan bobot sesuai dengan skema pembobotan yang dipilih, apakah pembobotan lokal, global atau kombinasi keduanya. Metode pembobotan yang digunakan adalah TF-IDF yang didefinisikan pada persamaan 2.

Luaran dari proses penghitungan bobot ini adalah matriks bobot. Baris dari matriks bobot menunjukkan dokumen sedangkan kolomnya menunjukkan kata. Contoh matriks bobot dapat dilihat pada Tabel 1.

Tabel 1. Contoh Matriks Bobot

	Kata 1	Kata 2	Kata 3	Kata 4	Kata 5
Doc1	0.3	0.1	...	...	....
Doc2	0.5	0.4	...	...	...
Doc3	0.7	0.9	...	...	...
Doc4	0.3	0.3	0.3	0.8	0.2

6.3 Proses Klasifikasi

Setelah semua kata diberi bobot, hal terakhir yang harus dilakukan adalah mengklasifikasikan dokumen menggunakan fitur bobot masing-masing kata di dalam dokumen.Adapun algoritma klasifikasi yang digunakan dalam sistem ini adalah algoritma *Support Vector Machine.*

7. UJI COBA DAN EVALUASI

Dalam uji coba pada penelitian ini, dokumen artikel ilmiah yang digunakan sebagai data uji coba diunduh dari perpustakaan digital Institut Teknologi Sepuluh Nopember[8].Jumlah dokumen yang digunakan untuk pengujian sebanyak 806 dan dokumen-dokumen tersebut terbagi dalam dua puluh tiga kategori berdasarkan bidang ilmu yang berkaitan dengan isi dokumen.

Dokumen yang digunakan untuk ujicoba bersifat homogen karena kata-kata yang ada di dalam satu dokumen memiliki kedekatan makna atau bahkan sama dengan kata-kata yang ada dalam dokumen lain. Misalnya, kata citra, algoritma, enkripsi dan deskripsi yang sering muncul dalam tiga kategori sekaligus yaitu kategori Pengolahan Citra Digital, Jaringan Nirkabel, dan Jaringan Multimedia.

Berdasarkan hipotesa pembobotan bertingkat yang dijelaskan pada sub bab sebelumnya, skenario uji coba dalam makalah ini dilakukan dengan menetapkan $\gamma_1$dan γ_2 sebagai parameter yang nilainya akan dirubah-rubah. Adapun hasil uji coba yang telah dilakukan dapat dibaca pada Tabel 2.

Dari Tabel 2 dapat dilihat bahwa nilai presisi dari sistem klasifikasi dokumen artikel ilmiah yang menggunakan fitur abstraksi dan catatan publikasi penulis lebih tinggi jika dibandingkan dengan sistem klasifikasi dokumen artikel ilmiah yang menggunakan fitur abstraksi saja. Sedangkan hasil *tunning* parameter menunjukkan bahwa sistem klasifikasi dokumen artikel ilmiah mencapai nilai presisi yang stabil ketika nilai $\frac{\gamma_1}{\gamma_2} \geq 1$.

Dari Tabel 2 dapat pula dilihat bahwa nilai *recall* dari sistem klasifikasi dokumen artikel ilmiah yang menggunakan fitur abstraksi dan catatan publikasi penulis lebih tinggi jika dibandingkan dengan sistem klasifikasi dokumen artikel ilmiah yang menggunakan fitur abstraksi saja. Sedangkan hasil *tunning* parameter menunjukkan bahwa proses klasifikasi mencapai nilai *recall* tertinggi ketika nilai $\frac{\gamma_1}{\gamma_2} = 1$. Namun ketika nilai $\frac{\gamma_1}{\gamma_2} \geq 1$, nilai *recall* turun sehingga dari sini dapat diketahui bahwa performa terbaik dari metode yang diusulkan dalam hal presisi dan *recall* dicapai ketika nilai $\frac{\gamma_1}{\gamma_2} = 1$.

Tabel 2. Nilai presisi dan *recall* yang dihasilkan pada proses uji coba

Fitur	γ_1	γ_2	Presisi	Recall
Abstraksi	-	-	0.75	0.51
Abstraksi. Catatan publikasi penulis	0.3	0.7	0.77	0.56
Abstraksi. Catatan publikasi penulis	0.4	0.6	0.84	0.59
Abstraksi. Catatan publikasi penulis	0.5	0.5	0.87	0.59
Abstraksi. Catatan publikasi penulis	0.6	0.4	0.87	0.57
Abstraksi. Catatan publikasi penulis	0.7	0.3	0.87	0.56

8. KESIMPULAN

Berdasarkan uji coba yang telah dilakukan, dapat disimpulkan bahwa sistem klasifikasi dokumen artikel ilmiah yang menggunakan fitur abstraksi dan catatan publikasi ilmiah penulis artikel menghasilkan presisi dan *recall* yang lebih tinggi dibandingkan dengan klasifikasi yang menggunakan fitur abstraksi saja. Namun nilai *recall* yang kurang dari 0.6 menunjukkan bahwa permasalahan seleksi fitur di bidang klasifikasi dokumen artikel ilmiah masih menjadi permasalahan yang terbuka untuk dicari solusinya.

Performa terbaik dalam hal presisi dan *recall* dari sistem klasifikasi dokumen artikel ilmiah yang diusulkan pada makalah ini dicapai ketika nilai $\frac{\gamma_1}{\gamma_2} = 1$. Berdasarkan hasil tersebut dapat disimpulkan bahwa baik abstraksi maupun catatan publikasi penulis artikel memiliki tingkat signifikansi yang sama dalam penentuan kelas pada proses klasifikasi sehingga dapat disimpulkan bahwa hipotesa pembobotan bertingkat tertolak atau tidak dapat dibuktikan kebenarannya.

9. DAFTAR PUSTAKA

[1] Montejo-Raez. 2005. "Text Categorization Using Bibliographic". Procesamiento del Lenguaje Natural 35. 119–262.

[2] Kerstin Denecke. 2009. "Topic Classification Using Limited Bibliographic". Digital Information Management. ICDIM 2009.

[3] G. H. Martin, S. Schokaert, C. Cornelis, and H. Naessens. 2013. "Using Semi-structured Data for Assessing Research Paper Similarity". Information Sciences 221. 245-261.

[4] P. N. Tan, M. Steinbach, and V. Kumar. 2005. Introduction to Data Mining. Addison Weasley.

[5] A. Rachmania. 2011. Klasifikasi Kategori dan Identifikasi Topik pada Artikel Berita Berbahasa Indonesia. Surabaya: ITS.

[6] OpenCV. September. 2012.Introduction to Support Vector Machines. <Available: http://docs.opencv.org/2.4.2/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html. >

[7] J. B. Lovins. 1968. "Development of a Stemming Algorithm". Mechanical Translation and Computational Linguisticsvol. 11. 22-31,

[8] digilib.its.ac.id (Surabaya). 2012. 20 Oktober.