KLASIFIKASI TEXT BERITA DENGAN WEIGHT ADJUSTED K-NEAREST NEIGHBOR

Hendrice Elisabeth[1], Adiwijawa[2], Angelina Prima Kurniati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Klasifikasi teks adalah salah satu permasalahan dalam text mining. Banyak metode yang dapat digunakan untuk menyelesaikan masalah tersebut. Salah satu metode klasifikasi adalah Weight Adjusted K-Nearest Neighbor(WAKNN). Metode ini adalah metode yang didasarkan pada K-Nearest Neighbor yang merupakan salah satu algoritma learning yang sangat efektif untuk berbagai domain permasalahan. Namun, K-Nearest Neighbor dianggap kurang efektif dalam pengukuran similarity/kemiripan sifat antara satu dokumen dengan dokumen yang lainnya karena memakai semua term yang terdapat dalam dokumen tersebut baik penting maupun tidak. Sedangkan pada Weigth Adjusted K-Nearest Neighbor akan menghitung serta mengevaluasi bobot kata dari setiap dokumen untuk menentukan kata-kata penting dari suatu kelas sehingga pada proses klasifikasi, setiap dokumen akan dibandingkan antara satu dengan yang lainnya sesuai dengan kata-kata penting yang dimilikinya. Pada tugas akhir ini, akan dicoba untuk mengklasifikasikan teks berita berbahasa Indonesia dengan menggunakan Weigth Adjusted K-Nearest Neighbor. Parameter yang akan diuji adalah precision, recall, dan f-measure. Berdasarkan hasil pengujian, WAKNN terbukti menghasilkan tingkat akurasi yang lebih baik daripada KNN

Kata Kunci : Klasifikasi teks berita, Weigth Adjusted K-Nearest Neighbor, K-Nearest Neighbor

Abstract

Text classification is one of problems in text mining. Many methods that can be used to solve this problem. One of those methods is Weight Adjusted K-Nearest Neighbor(WAKNN). This method is based on the K-Nearest Neighbor classification paradigm which is proved very effective for many problems. But, K-Nearest Neighbor seems to be less effective in similarity measurement because it uses all terms in a document without consider the importances of those terms. Whereas in Weight Adjusted K-Nearest Neighbor, it will count and evaluate the weight from each term in a document for choosing some important terms from each class so that in the classification process, one document will be compared to another document by using their important terms. In this final project, it will try to classify news text in Bahasa Indonesia by using Weight Adjusted K-Nearest Neighbor. Some parameters that will be tested are precision, recall, and f-measure. Refers to the result of the experiment, WAKNN is proved giving better accuracy than KNN.

Keywords : Text classification, Weigth Adjusted k-Nearest Neighbor, K-Nearest Neighbor

1. Pendahuluan

1.1 Latar belakang masalah

Semakin hari, informasi yang berupa dokumen, teks atau artikel semakin banyak dihasilkan dan dibutuhkan oleh seluruh kalangan masyarakat. Sebagai contoh, artikel berita. Setiap hari jumlah artikel berita makin meningkat karena selalu ada berita terbaru. Jika peningkatan ini tidak dikelola dengan baik maka masyarakat tidak akan mendapatkan informasi yang diperlukan dengan cepat dan mudah.

Untuk mengatasi masalah tersebut maka muncullah suatu bidang ilmu yang disebut *text mining*. *Text mining* adalah proses pencarian informasi yang didapat dengan memproses data teks dalam jumlah yang besar yang ditulis dengan bahasa manusia. Aktivitas-aktivitas yang terdapat pada teks mining adalah *Information Retrieval, Text Categorization, POS Tagging, Clustering, dll*. Pembahasan dalam Tugas Akhir ini lebih fokus pada tahap *categorization*. Aktivitas pada *categorization* lebih terfokus pada penentuan kelas-kelas dari beberapa dokumen ke dalam kelas-kelas yang sudah didefinisikan sebelumnya.

Pada tahap *text categorization* atau klasifikasi teks terdapat banyak metode yang dapat digunakan. Namun demikian, tidak semua metode cocok untuk diterapkan pada berbagai masalah klasifikasi terutama klasifikasi teks dalam jumlah besar. Oleh karena itu, pada Tugas Akhir ini digunakan metode *Weight Adjusted k-Nearest Neighbor* (WAKNN). Metode ini merupakan perbaikan dari metode *k-Nearest Neighbor* yang merupakan salah satu algoritma learning yang sangat efektif untuk berbagai domain permasalahan. Namun, *k-Nearest Neighbor* (KNN) dianggap kurang efektif dalam pengukuran *similarity*/kemiripan sifat antara satu dokumen dengan dokumen yang lainnya karena memakai semua kosakata yang terdapat dalam dokumen tersebut baik penting maupun tidak.

Dalam hal ini, WAKNN akan menghitung dan mengevaluasi bobot kata dari setiap dokumen untuk menentukan kata-kata penting dari suatu kelas sehingga pada proses klasifikasi, setiap dokumen akan dibandingkan antara satu dengan yang lainnya sesuai dengan kata penting yang dimilikinya.

1.2 Perumusan masalah

Beberapa permasalahan yang akan dibahas pada Tugas Akhir ini adalah:

1. Bagaimana mengubah suatu dokumen berita menjadi data yang siap untuk digunakan dalam proses klasifikasi melalui tahap *preprocessing* yang meliputi: proses penghilangan *stopword*, dan *feature selection*.
2. Bagaimana proses klasifikasi data yang berupa teks berita dengan menggunakan WAKNN sehingga diperoleh suatu model yang akurat dalam proses klasifikasi.
3. Bagaimana akurasi metode WAKNN dalam proses klasifikasi yang dapat dilihat dari jumlah dokumen yang diklasifikasikan dengan benar bila dibandingkan dengan metode KNN.

Batasan masalah yang digunakan pada Tugas Akhir ini adalah :

1. Artikel berita yang digunakan bukanlah berbentuk halaman web tapi dokumen yang bersifat offline. Sumber dokumen didapat dari www.kompas.com tanggal 1 April sampai dengan 31 April 2008.
2. Artikel berita yang digunakan hanya artikel yang berbahasa Indonesia.

Hipotesis sementara adalah metode WAKNN merupakan metode yang lebih akurat untuk proses klasifikasi teks bila dibandingkan dengan metode *k-Nearest Neighbor*.

1.3 Tujuan

Berdasarkan permasalahan di atas, tujuan dari Tugas Akhir ini adalah:

1. Membangun suatu aplikasi untuk proses klasifikasi teks dengan menggunakan metode WAKNN.
2. Membandingkan akurasi proses klasifikasi teks yang dilakukan dengan metode WAKNN dengan metode klasifikasi KNN yang belum dimodifikasi.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian yang akan dilakukan adalah:

1. Studi Literatur :
   a. Pencarian referensi yang berhubungan dengan Tugas Akhir
   b. Mempelajari dan lebih memperdalam materi yang berhubungan dengan Tugas Akhir.
2. Mengumpulkan artikel-artikel berita.
3. Melakukan *preprocessing* terhadap artikel-artikel berita yang telah dikumpulkan.
4. Mempelajari konsep dari metode WAKNN untuk diterapkan pada klasifikasi berita.
5. Melakukan analisis dan perancangan perangkat lunak.
6. Melakukan implementasi perangkat lunak.
7. Melakukan pengujian akurasi metode WAKNN dalam proses klasifikasi berita dengan melihat persentase dokumen yang diklasifikasikan dengan benar bila dibandingkan dengan metode klasifikasi KNN.
8. Menganalisis hasil pengujian.
9. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

5. Penutup

5.1 Kesimpulan

Berdasarkan percobaan dan analisis yang telah dibahas dan dilaksanakan pada bab 4, maka dapat disimpulkan beberapa hal sebagai berikut :

1. Metode *Weigth Adjusted K-Nearest Neighbor*(WAKNN) merupakan metode yang cukup akurat untuk digunakan di dalam klasifikasi berita berbahasa Indonesia dengan nilai *precision* sebesar 77,71%, *recall* sebesar 74,54%, dan *f-measure* sebesar 73,9% pada saat optimum yaitu pada saat nilai $k = 5$ dan minimum prosentase kemiripan = 50.
2. Metode WAKNN terbukti mempunyai tingkat akurasi yang lebih baik daripada KNN dengan perbedaan akurasi antara 2-9%. Hal-hal yang mempengaruhi perbedaan hasil klasifikasi antara metode KNN dan WAKNN adalah pemilihan *terms* yang akan digunakan untuk proses klasifikasi. Pada metode WAKNN, semua *terms* dipakai pada proses klasifikasinya namun terdapat inisialisasi bobot untuk pemilihan *terms* yang akan digunakan. Selain itu, pada WAKNN juga terdapat evaluasi bobot yang dapat meningkatkan syarat minimum prosentase kemiripan.
3. Penentuan jumlah nilai tetangga terdekat (k), minimum prosentase kemiripan (p), dan inisialisasi bobot sangat penting. Nilai – nilai tersebut tidak boleh terlalu kecil atau terlalu besar. Berdasarkan hasil pengujian, hasil klasifikasi yang optimum ada pada saat $k = 5$, prosentase kemiripan sebesar 50%, dan inisialisasi bobot pada rentang 0,01-0,09.

5.2 Saran

Sebagai acuan dalam melengkapi atau memperbaiki hasil analisis data yang dilakukan dalam Tugas Akhir ini. Ada beberapa saran yang dapat dijadikan pertimbangan bagi analisis data selanjutnya, diantaranya :

1. Aplikasi dapat dikembangkan untuk pengambilan berita yang dilakukan secara *on-line*.
2. Penambahan jumlah data berita yang digunakan untuk *training*.