ANALISIS PENGARUH FEATURE SELECTION MENGGUNAKAN INFORMATION GAIN DAN CHI-SQUARE UNTUK KATEGORISASI TEKS BERBAHASA INDONESIA

Ika Sofiana[1], Imelda Atastina[2], Arie Ardiyanti Suryani[3]

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

---

Abstrak

Salah satu permasalahan yang signifikan dalam kategorisasi teks adalah dimensionalitas data yang sangat tinggi yang menyebabkan waktu pemrosesan menjadi lebih lama. Salah satu cara untuk mengatasi hal tersebut adalah dengan melakukan feature selection. Feature selection dilakukan untuk memilih fitur-fitur penting dan relevan terhadap data dan membuang fitur-fitur yang tidak berpengaruh. Dengan demikian, dimensionalitas data dapat dikurangi.

Dalam tugas akhir ini, permasalahan yang diangkat berkaitan dengan feature selection menggunakan Information Gain dan Chi-Square, pada kategorisasi teks dengan classifier N-gram. Kategorisasi dilakukan dengan menghitung jarak profil kategori ke profil dokumen, di mana profil dibentuk dari fitur-fitur yang ada. Sehingga jumlah fitur sangat mempengaruhi waktu yang dibutuhkan dalam proses kategorisasi. Hasil F-measure yang didapatkan pada kategorisasi teks dengan N-gram tanpa feature selection adalah 0.89, di mana gram yang digunakan adalah 2-gram. Dan ketika mengalami feature selection dengan menggunakan Information Gain sebanyak 80 %, F-Measure meningkat menjadi 0.935, serta ketika mengalami feature selection sebanyak 20 % dengan Chi-Square, F-Measure meningkat menjadi 0.94.

Proses pemilihan fitur dengan menggunakan Information Gain lebih cepat dibandingkan dengan Chi-Square. Akan tetapi, secara keseluruhan performansi yang dihasilkan oleh fitur-fitur hasil pemilihan Chi-Square memberikan hasil yang lebih baik.

Kata Kunci : feature selection, 2-gram, Information Gain, Chi-Square, F-Measure.

Abstract

One of the significant problem in the text categorization is high dimentionality of data that cause a long processing time. One of the severals ways to overcome this problem is doing feature selection phase to the data before categorization process. The goal of feature selection is to produce important and relevant features. Therefore, the data dimensionality can be reduced.

In this final task, the research is about feature selection using Information Gain and Chi-Square in N-Gram text categorization. Categorization is done by counted the distance of category profile and the document profile, where the profiles is made from the features existed. Therefore, the number of the feature is have a high influence in the time needed for categorization process. In the text categorization using N-gram without feature selection, the result shows that F-measure give a value of 0.89, where 2-gram is used here. When feature selection is done by Information Gain to the number of 80 %, F-measure value increase up to 0.935. And When feature selection is done by Chi-Square to the number of 20 %, F-measure value increase up to 0.94.

Selecting feature using Information Gain feature selection is faster than selecting feature using Chi-Square feature selection. However, the performace of text categorization using features from the feature selection proses by Chi-square is better.

Keywords : feature selection, 2-gram, Information Gain, Chi-Square, F-Measure.

---

1. Pendahuluan

1.1. Latar belakang

Kategorisasi teks merupakan sebuah permasalahan pemilihan kategori untuk sebuah teks yang memiliki karakteristik atau tanda-tanda umum yang dimiliki oleh teks, artikel atau dokumen tersebut [13]. Suatu teks, artikel, atau dokumen berita dapat dikategorisasikan oleh beberapa, tepat satu, atau tidak termasuk dalam kategori manapun [8].

Permasalahan yang signifikan dalam kategorisasi teks adalah dimensionalitas data yang sangat tinggi. Dimensi data yang sangat tinggi tersebut menyebabkan waktu pemrosesan menjadi lebih lama. Salah satu cara untuk mengatasi hal tersebut adalah dengan melakukan *feature selection* terhadap data. *Feature selection* dilakukan untuk memilih fitur-fitur penting dan relevan terhadap data dan membuang fitur-fitur yang tidak berpengaruh terhadap proses kategorisasi teks. Dengan demikian, dimensionalitas data dapat dikurangi.

Dalam tugas akhir ini metode kategorisasi teks yang digunakan adalah N-Gram. Pemilihan metode tersebut didasarkan pada penelitian-penelitian sebelumnya yang menyebutkan bahwa metode tersebut memiliki performansi yang baik dalam permasalahan kategorisasi teks. Selain itu, pada N-Gram, jumlah fitur sangat mempengaruhi proses komputasinya, yang berpengaruh pula pada waktu yang dibutuhkan untuk kategorisasi itu sendiri.

Metode N-Gram menurut Trenkle & Cavnar (1994) adalah munculnya makna atau kata baru dari seperangkat karakter hasil pemotongan pada sebuah kata. Dari penelitian yang telah dilakukan, kategorisasi *Tri*-gram menunjukkan permformansi yang paling baik di antara yang lain, dengan persentase tertinggi 81,25% dan terendah 5,357% [16]. N-gram merupakan urutan karakter dengan panjang *n* yang di dapatkan dari ekstraksi dokumen atau teks. Dengan menggunakan N-gram sistem dapat mencapai *Language Independence* dimana kebanyakan *Information Rietrieval System* memiliki ketergantungan terhadap bahasa yang digunakan[1,10]. Selain itu, N-Gram dapat diterapkan untuk teks yang memiliki kesalahan penulisan, penggunaan karakter, dan tanda asing.

Dalam tugas akhir ini, penulis menganalisis pengaruh dari *feature selection* menggunakan *Information Gain* dan *Chi-Square* terhadap performansi kategorisasi teks yang dilakukan dengan metode Ngram. Analisis dilakukan terhadap waktu kategorisasi dan akurasi kategorisasi yang diukur dengan nilai *F-Measure* dari kategorisasi menggunakan N-Gram.

1.2. Perumusan masalah

Dalam tugas akhir ini permasalahan yang akan di bahas diantaranya adalah:

a. Bagaimana pengaruh *feature selection* terhadap performansi kategorisasi menggunakan klasifier N-gram?
b. Bagaimana pengaruh perubahan jumlah dokumen terhadap performansi *feature selection* pada kategorisasi teks?
c. Bagaimana pengaruh perubahan jumlah kategori terhadap performansi *feature selection* pada kategorisasi teks?

Hipotesis :

Waktu yang dibutuhkan untuk kategorisasi teks dengan *feature selection* lebih kecil dibandingkan kategorisasi teks tanpa *feature selection*. Nilai *F-Measure* yang dihasilkan dari kategorisasi teks dengan *feature selection* lebih besar dibandingkan kategorisasi teks tanpa *feature selection.*

1.3. Tujuan

Tujuan yang ingin dicapai dari Tugas Akhir ini adalah :

a. Mengimplementasikan dan menganalisis *feature selection* dengan *Information Gain* dan *Chi-Square* pada kategorisasi teks dengan metode N-Gram.
b. Menganalisis pengaruh *feature selection* terhadap performansi hasil kategorisasi berdasarkan *F-Measure* dan waktu kategorisasi.
c. Menganalisis pengaruh jumlah dokumen training dan variansi data terhadap nilai *F-Measure* dan waktu yang dibutuhkan untuk kategorisasi teks dengan *feature selection.*

1.4. Batasan Masalah

Permasalahan dalam Tugas Akhir ini dibatasi untuk hal-hal sebagai berikut.

a. Dokumen atau teks yang digunakan sebagai data latih dan data uji merupakan dokumen atau teks berita berbahasa Indonesia yang diambil dari situs berita **www.okezone.com** dan disimpan dalam file berekstensi *.txt.
b. Dokumen atau teks yang digunakan sebagai data latih dan data uji merupakan dokumen dengan kapasitas maksimal 4 kb.

1.5. Metodologi Penyelesaian Masalah

Metodologi yang digunakan adalah :

a. Studi literatur

Pada tahap ini, penulis mencari dan mempelajari konsep dan teori yang berhubungan dengan topik yaitu *feature selection*, *Information Gain*, *Chi-Square*, *Text Categorization* (TC), dan N-Gram.

b. Pengumpulan data

Pada tahap ini dilakukan pencarian data, data yang akan digunakan pada tugas akhir ini berupa dokumen atau teks-teks berbahasa Indonesia yang disimpan dalam file berekstensi *.txt dengan ukuran maksimal 4 kb.

c. Analisis dan perancangan sistem

Melakukan analisis dan perancangan terhadap sistem yang dibangun, menganalisis metode yang akan digunakan untuk menyelesaikan permasalahan, termasuk menentukan bahasa pemrograman yang digunakan, arsitektur, fungsionalitas, dan antarmuka sistem. Input dari sistem ini adalah dokumen atau teks berbahasa Indonesia yang akan ditentukan kategorinya. Sedangkan output dari sistem adalah kategori dari teks atau dokumen yang telah di input.

Sistem dalam tugas akhir ini terdiri dari *feature selection*, pembentukan model, dan analisis performansi untuk masing-masing metode.

1) *Feature selection*
   Pada proses ini dilakukan pemilihan fitur yang akan digunakan dalam proses kategorisasi. Fitur ini berupa *term* (token) yang diambil dari dokumen.
2) Proses *Learning*
   Dalam proses ini, akan dilakukan *learning* dengan metode yang telah ditentukan terhadap dokumen atau teks-teks yang telah dikumpulkan sebagai data latih.
3) Analisis performansi
   Analisis performansi dilihat untuk masing-masing metode dengan parameter *F-Measure* dan waktu kategorisasi.

d. Implementasi dan pembangunan sistem
   Sistem ini diimplementasikan dengan bahasa pemrograman *java*.
e. Pengujian dan analisis
   Pengujian dan analisis dilakukan dengan cara:
   1) Mengkategorisasikan dokumen atau teks yang digunakan untuk pengujian (data uji) terhadap sistem.
   2) Mengkalkulasikan akurasi kategorisasi.
   3) Mengkalkulasikan waktu yang dibutuhkan untuk melakukan kategorisasi.
f. Penyusunan laporan Tugas Akhir.
   Dalam tahap ini akan disusun buku laporan Tugas Akhir yang berisi segala implementasi dari sistem yang telah dikerjakan.

1.6. Sistematika Penulisan

Tugas akhir ini disusun berdasarkan sistematika berikut :

BAB 1 : Pendahuluan
Yang dipaparkan dalam bab ini adalah latar belakang penelitian, perumusan masalah, tujuan penelitian, batasan masalah, metodologi penelitian, dan sistematika penulisan tugas akhir.

BAB 2 : Landasan Teori
Pada bab ini dijelaskan mengenai teori-teori yang berhubungan dengan tugas akhir ini.

BAB 3 : Analisis dan Perancangan Sistem
Pada bab ini akan dijelaskan proses perancangan dan implementasi sistem yang digunakan dalam tugas akhir ini.

BAB 4 : Implementasi dan Evaluasi Hasil
Pada bab ini, dipaparkan mengenai pengimplementasian sistem dan evaluasi hasil pengujian sistem.

BAB 5 : Kesimpulan dan Saran
Pada bab ini dipaparkan beberapa kesimpulan mengenai permasalahan yang dibahas. Selain itu, pada bab ini, juga akan diberikan saran untuk pengembangan selanjutnya.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Dari hasil pengujian dan analisis yang telah dilakukan sebelumnya dalam tugas akhir ini, dapat diperoleh kesimpulan sebagai berikut.

1) Pada kategorisasi teks dengan menggunakan N-gram *classifier*, penggunaan *feature selection* meningkatkan nilai F-Measure dan menurunkan waktu yang dibutuhkan untuk kategorisasi.
2) Waktu yang dibutuhkan untuk melakukan *feature selection* dengan menggunakan *Information gain* lebih cepat dibandingkan dengan *Chi-Square.*
3) Performansi yang didapatkan oleh *Information Gain* tinggi pada jumlah fitur 80 % dari fitur awalnya, sedangkan pada *Chi-Square* performansi tinggi didapatkan pada jumlah fitur 20 % dari fitur awalnya.
4) Kategorisasi dengan *feature selection* membutuhkan waktu kategorisasi secara keseluruhan yang lebih kecil dari pada kategorisasi tanpa *feature selection* pada jumlah fitur terpilih 10-40 %, dan membutuhkan waktu yang lebih lama pada jumlah fitur terpilih 50-90%. Sehingga dari kesimpulan poin 3, dapat disimpulkan bahwa pada kategorisasi teks, *feature selection* dengan *Chi-Square* lebih baik dibandingkan *feature selection* dengan *Information Gain.*
5) Semakin banyak jumlah dokumen yang digunakan dalam kategorisasi meningkatkan nilai *F-Measure* kategorisasi teks dengan *Chi-Square feature selection*, dan menurunkan nilai *F-Measure* kategorisasi teks dengan *Information Gain feature selection.*
6) Salah satu yang mempengaruhi perubahan nilai *F-measure* pada kesimpulan poin 5 adalah adanya perubahan terhadap karakteristik data yang dihasilkan dari proses *feature selection*. Karakteristik data tersebut berupa perubahan variansi data, perubahan fitur, dan perubahan bobot dari fitur itu sendiri.

5.2. Saran

Saran yang dapat penulis sampaikan untuk pengembangan tugas akhir ini adalah sebagai berikut.

1) Dokumen yang digunakan dalam pembelajaran kategorisasi teks lebih banyak, sehingga hasil yang didapatkan lebih baik.
2) Sebelum melakukan klasifikasi atau kategorisasi, ada baiknya dilakukan *clustering* terlebih dahulu, sehingga didapatkan *cluster-cluster* dokumen yang benar-benar memiliki karakteristik yang sama.
3) Sebaiknya, pada saat memilih dokumen yang akan digunakan, diperhatikan juga mengenai karakteristik yang dimiliki oleh dokumen-dokumen tersebut pada masing-masing kategori.

Daftar Pustaka

[1] Babu, A Suresh and Kumar, PNVS Pavan. *Comparing Neural Network Approach with N-Gram Approach for Text Categorization*. International Journal on Computer Science and Engineering Vol 2(1) 80-83. 2010.

[2] Burges, Christopher JC. *A Tutorial on Support Vector Machines for Pattern Recognition*. Datamining and Knowledge Discovery Vol 2, 121-167. 1998.

[3] Cavnar, W.B and Trenkle, John M. *N-Gram Based Text Categorization*. Environmental Research Institut of Michigan.

[4] Damishu, Debby. *Wrapper Feature Selection pada Pengkategorisasian Artikel Berita Berbahasa Indonesia.* IT Telkom. Bandung. 2008.

[5] Forman, George. *Feature Selection for Text Categorization*. Information Services and Process Innovation Laboratory, HP Laboratory Palo Alto. May 3. 2007.

[6] Frey, Remo. *Text Categorization : Support Vector Machine*. ETH Zurich, Algorithm for Database Systems. 2007.

[7] Hamzah, Amir. *Deteksi Bahasa untuk Dokumen Berbahasa Indonesia.* Seminar Nasional Informatika, UPN "Veteran" Yogyakarta. Mei 2010

[8] Joachim, Thorsten. *Text Categorization with Support Vector Machines : Learning with Many Relevant Feature*. Universitat Dortmund Informatik LSVII, Germany. 1997.

[9] Ladha, L. dan Deepa, T. *Feature Selection Method and Algorithm*. International Journal on Computer Science and Engineering, Departement of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamilnadu, India Vol 3 No 5 May 2011.

[10] Mansur, Munirul. *Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus.* Center of Research on Bangla Language Processing, BRAC University Dhaka, Bangladesh. 2006.

[11] Morariu, Daniel I dan Vintan, Lucian N dan Tresp, Volker. *Evolutionary Feature Selection for Text Document Using the Support Vector Machines*. World Academy of Science, Engineering, and Technology. 2006.

[12] Mukras, Rahman dkk.. *Information Gain Feature Selection for Ordinal Text Categorization Using Probability Re-Distribution*. School of Computing, The Robert Gordon University.

[13] Nather, Peter. *N-Gram Based Text Categorization*. Faculty of Mathematics, Physics, and Informatics, Comenius University, Bratislava. 2005.

[14] Nugroho, Anto Satriyo. *Pengantar Support Vector Machines*. Disampaikan pada e-Tutorial SVM di Milis indo_dm@yahoogroups.com, 5-18 Februari 2007.

[15] Nugroho, Anto Satriyo dan Witarto, Arief Budi dan Handoko Dwi. *Support Vector Machine : Teori dan Aplikasinya dalam Bioinformatika*. Kuliah Umum IlmuComputer.com. 2003.

[16] Permadi, Yudha. *Kategorisasi Teks Menggunakan N-Gram untuk Dokumen Berbahasa Indonesia*. Departemen Ilmu Komputer Fakultas Matematika dan IPA Institut Pertanian Bogor. 2008.

[17] Putri, Yoshi Amelia. *Stemming untuk Teks Berbahasa Indonesia dan Pengaruhnya dalam Kategorisasi.* IT Telkom. Bandung. 2009.
[18] Said, Dina Adel. *Dimentionality Reduction Technique for Enhancing Automatic Text Categorization.* Faculty of Engineering, Cairo University, Giza, Egypt. 2007.
[19] Sari, Khrisna Dini Yunita . *Kategorisasi Teks dengan Metode Klasifikasi Support Vector Machine.* Departemen Teknik Informatika Sekolah Tinggi Teknologi Telkom. 2006.
[20] Sembiring, Krisantus. *Penerapan Teknik Support Vector Machines untuk Pendeteksian Intrusi pada Jaringan.* Program Studi Informatika, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung. 2007.
[21] Thabtah, Fadi dan Eljinini, Mohammad Ali H dan Zamzeer, Mannam dan Hadi, Wa'el Musa. *Naive Bayesian on Chi Square to Categorize Arabic Data.* Philadelphia University and Al Isra Private University, Jordan.
[22] Yang, Cheng-San. *A Hybrid Feature Selection Method for Microarray Classification.* IAENG International Journal of Computer Science, 35:3, IJCS_35_3_05. Agustus 2008.