METODE FEATURE SELECTION DALAM MENANGANI DATA IMBALANCE PADA KLASIFIKASI DOKUMEN MULTI-LABEL

Luthfia Rahmani[1], Moch Arif Bijaksana[2], Rimba Widhiana Ciptasari[3]

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

Abstrak

Pengkategorian dokumen merupakan salah satu permasalahan dalam text mining. Salah satu cara agar suatu dokumen dapat dikategorikan adalah dengan menggunakan teknik klasifikasi. Sekumpulan dokumen selain memiliki feature space yang berdimensi tinggi dapat juga memiliki sifat data yang imbalance. Sifat imbalance tersebut akan mengakibatkan klasifikasi data yang akan dibentuk kurang akurat. Untuk meningkatkan efisiensi dan keakuratan dalam klasifikasi dokumen, salah satunya dengan menggunakan teknik feature selection.

Pada tugas akhir ini dilakukan analisis perbandingan metode feature selection antara lain Odds Ratio (OR), GSS Coefficient, Information Gain (IG), improved OR (iOR), dan improved SIG (iSIG). Metode-metode feature selection tersebut diterapkan secara filter feature selection sedangkan pada wrapper feature selection menerapkan Odds Ratio (OR). Penerapan dilakukan menggunakan teknik klasifikasi multinomial naive bayes, metode tersebut menggunakan algoritma naive bayes dengan memperhitungkan jumlah kemunculan kata dalam dokumen. Selain menggunakan multinomial naive bayes, pada penerapan filter feature selection dilakukan juga proses pengklasifikasian dokumen menggunakan software Weka 3.5. Dengan melakukan analisis perbandingan metode feature selection diketahui metode mana yang paling handal dalam menangani data imbalance dengan menguji tingkat akurasi data setelah dilakukan klasifikasi dengan test set yang diberikan. Data yang digunakan berasal dari Reuters 21578 dengan dokumen bersifat multi-label.

Kata Kunci : Pada tugas akhir ini dilakukan analisis perbandingan metode feature

Abstract

Document categorization is one of problem in text mining. Classification technique is one of ways to categorize the document. Documents not only have high dimension of feature space, but also can have imbalance data characteristic. This imbalance will reduce the accuracy of data classification which is going to be built. One of solutions to increase the efficiency and the accuracy in classification document is by using feature selection technique.

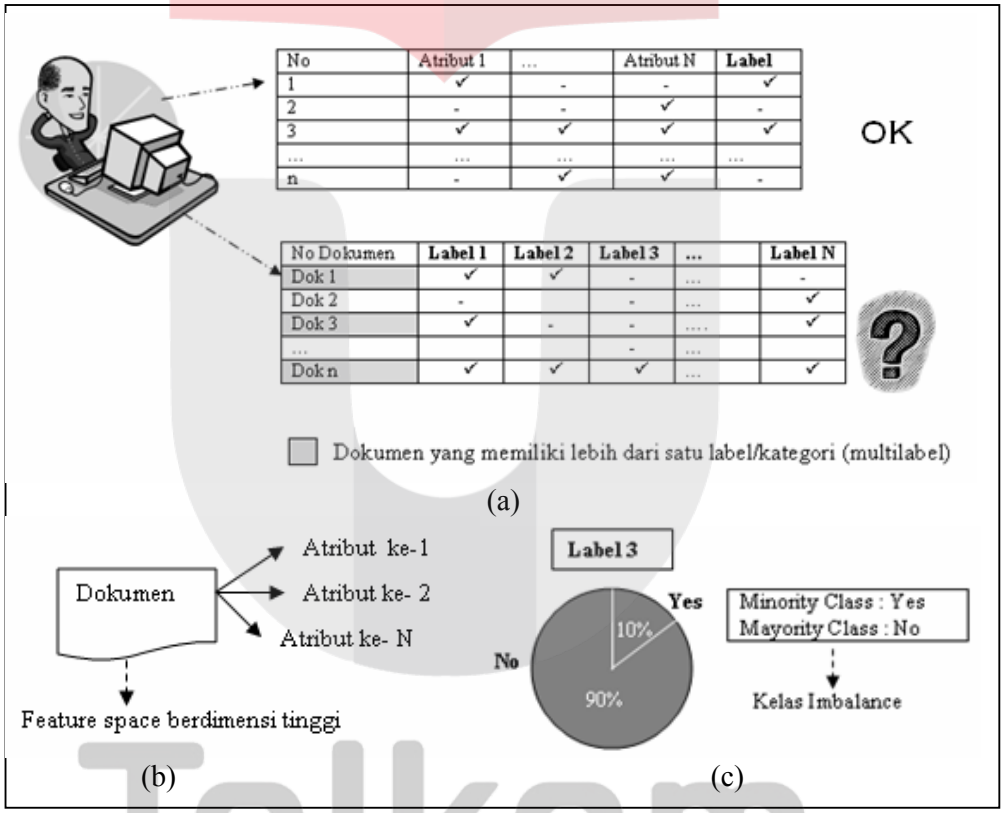
This final project do the comparison analysis feature selection methods, such as Odds Ratio (OR), GSS Coefficient, Information Gain (IG), improved OR (iOR), and improved SIG (iSIG). These feature selection methods are implemented in filter feature selection, whereas for wrapper feature selection implement Odds Ratio (OR). Implementation use multinomial naive bayes classification technique. The method use naive bayes alghorithm which for calculate upon amount of words that appear in document. Beside using multinomial naive bayes, Implementation in filter feature selection also use the process of document classification which available in software Weka 3.5. By using the comparison analysis feature selection methods, it find what method that the most reliable to handle the imbalance data by testing the accuracy level data after being classified by test set. Data that is used comes from Reuters 21578 that imbalace characteristic.

Keywords : text mining, classification, imbalance, feature selection, multinomial

1. Pendahuluan

1.1 Latar belakang

Dengan terus bertambahnya jumlah dan keanekaragaman dokumen, penggolongan secara manual tentu saja akan menjadi suatu masalah baru untuk user. Hal tersebut akan memakan banyak waktu dan menimbulkan kejenuhan. Dokumen yang tersebar dan tidak terkoordinasi dengan baik akan menyulitkan user dalam mendapatkan informasi yang diinginkan. Dengan dokumen yang telah terklasifikasi, pengguna informasi (*user*) dapat dengan mudah menemukan dokumen yang dibutuhkan karena dokumen tersebut telah dikelompokkan berdasarkan kategori yang mencerminkan isi dari suatu dokumen.



Gambar 1-1: Ruang Lingkup Permasalahan

Sekumpulan data biasanya diklasifikasikan dalam single label dengan jumlah atribut yang terbatas. Pada tugas akhir ini, dokumen yang akan digunakan dapat dikategorikan menjadi beberapa kategori atau label dengan atribut berupa kata-kata yang berasal dari dokumen teks, seperti pada gambar 1-1 bagian (a).

Salah satu kesulitan utama dalam permasalahan klasifikasi dokumen yaitu dokumen memiliki *feature space* yang berdimensi tinggi, seperti yang digambarkan pada gambar 1-1 bagian (b). Hal ini disebabkan karena masing-masing kata yang berbeda pada sekumpulan dokumen direpresentasikan sebagai satu dimensi pada *feature space*. Meskipun dokumen tersebut telah dilakukan

proses eliminasi kata-kata seperti *stop words*, akan tetapi fitur-fitur yang akan diterapkan pada teknik klasifikasi masih terlalu banyak [9].

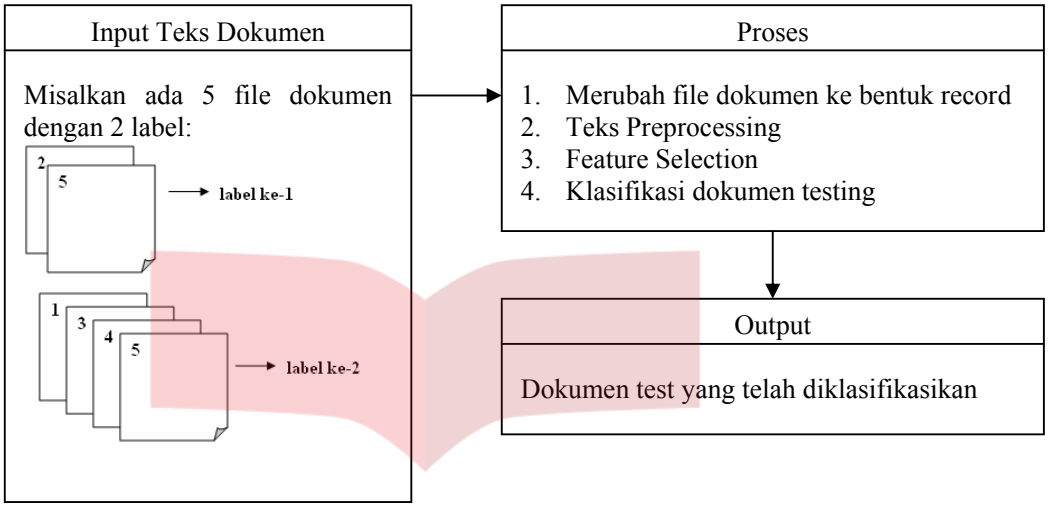
Sekumpulan dokumen selain memiliki *feature space* yang berdimensi tinggi dapat juga memiliki sifat data yang *imbalance,* seperti yang digambarkan pada gambar 1-1 bagian (c). Suatu data dikatakan *imbalance* jika data *training* terdistribusi tidak seimbang antara kelas positif dan kelas negatif. Contohnya, pada suatu label perbandingan persentase antara kelas negatif dan positif adalah 98% : 2%. Data *imbalance* akan berpengaruh terhadap klasifikasi data yang akan dibentuk. Dengan kondisi data yang tidak seimbang maka kecenderungan kelas data yang akan dihasilkan akan lebih condong ke bagian data yang memiliki komposisi data lebih besar (*majority class*). Untuk meningkatkan efisiensi dan keakuratan dalam klasifikasi dokumen diperlukan teknik *feature selection.* Teknik ini mengurangi jumlah fitur yang ada pada *feature space* dan juga merupakan salah satu pemecahan dalam menangani data *imbalance*.

Berawal dari masalah tersebut, maka dalam tugas akhir ini akan dilakukan analisis perbandingan metode *feature selection* untuk menangani data *imbalance* pada suatu klasifikasi dokumen. Diantaranya adalah metode *Odds Ratio* (OR) dan *GSS Coefficient* [9] yang termasuk kedalam *feature selection* menggunakan *one-sided metric* dimana *one-sided metric* hanya memilih fitur positif yang berpengaruh pada kelas. Kemudian *Information Gain* (IG) yang termasuk kedalam *two-sided metrics* dimana *two-sided metric* mengkombinasikan secara implisit fitur positif dan fitur negatif. Selain itu metode *improved OR* (iOR) dan *improved SIG* (iSIG) yang termasuk kedalam kombinasi antara fitur positif dan fitur negatif secara eksplisit .

Pendekatan *feature selection* yang akan dilakukan terdiri atas *filtering feature selection* [8] dan *wrapper feature selection* [6]. Pada penerapan *filtering feature selection,* selain menggunakan metode *multinomial naive bayes* juga akan dilakukan proses pengklasifikasian dokumen menggunakan metode yang ada dalam *tools* Weka 3.5. Metode *multinomial naive bayes* merupakan algoritma yang *naive* karena mengasumsikan indepedensi di antara kemunculan kata-kata dalam dokumen, tanpa memperhitungkan urutan kata dan informasi konteks dalam kalimat atau dokumen secara umum. Selain itu metode tersebut memperhitungkan jumlah kemunculan kata dalam dokumen [7].

Studi kasus yang akan digunakan merupakan data dari Reuters 21578 [3], yang berisi dokumen-dokumen teks dengan kondisi dokumen bersifat multilabel.

1.2 Perumusan masalah



Gambar 1-2: Bagan Sistem Klasifikasi Dokumen

Dengan mengacu pada latar belakang masalah diatas, maka permasalahan yang akan dibahas dan diteliti adalah :

1. Menangani data *imbalance* dengan penerapan metode *feature selection* dalam memilih *feature* yang relevan untuk meningkatkan keakuratan dalam klasifikasi dokumen. Hal ini disebabkan karena data *imbalance* dapat mengakibatkan hasil klasifikasi menjadi kurang akurat.
2. Mengklasifikasikan dokumen dengan menggunakan metode *multinomial naive bayes* untuk penerapan *filter* dan *wrapper feature selection*, serta menggunakan metode klasifikasi lain dengan *software* yang telah ada (Weka 3.5) untuk penerapan secara *filter*.
3. Menentukan metode *feature selection* yang mana yang baik untuk menangani data *imbalance*, dengan menganalisis hasil perhitungan *evaluation measures* sesuai dengan *feature selection* dan teknik klasifikasi yang digunakan.

Dalam Tugas Akhir ini batasan masalahnya sebagai berikut:

1. Input sistem berupa dokumen teks yang dapat berasal dari data bertipe *file text* atau *plain text*.
2. Tidak melakukan proses analisis secara sintatik ataupun semantik dan proses *stemming* sedangkan daftar *stopword* hanya dalam bentuk Bahasa Inggris.
3. Data yang akan dianalisis merupakan dokumen berupa teks dengan kondisi data *imbalance* dan penanganan data *imbalance* dilakukan secara *local feature selection*.
4. Klasifikasi yang akan dilakukan pada dokumen *multi-label* dengan penerapan *filter feature selection* menggunakan empat teknik klasifikasi yaitu *Support Vector machine* (SVM), KNN, *naive bayes* dan *multinomial naive bayes*. Pada *wrapper feature selection* [6], metode *feature selection* yang digunakan yaitu Odds Ratio dengan menerapkan teknik klasifikasi Multinomial Naive Bayes
5. Evaluation measures dihitung berdasarkan nilai precision, recall, F1 measure, macro-averaging, dan hamming loss.
6. Studi kasus yang akan digunakan merupakan data dari Reuters 21578.

1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari tugas akhir ini antara lain :

1. Menerapkan metode *feature selection* untuk memilih atribut secara tepat sehingga dapat memperbaiki atau meningkatkan tingkat akurasi data dalam menangani data *imbalance.*
2. Menganalisis perbandingan metode-metode *feature selection* untuk mengetahui metode mana yang paling handal dalam menangani data *imbalance* dengan menguji tingkat akurasi data melalui nilai *precision, recall,* F1 *measure*, *macro-averaging,* dan *hamming loss*, setelah dilakukan klasifikasi dengan data *test set* yang diberikan.

1.4 Metodologi penyelesaian masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut:

1. Studi literatur
   Pada tahap ini dilakukan pencarian, pengumpulan informasi dan pendalaman materi yang berupa literatur dari buku-buku referensi, artikel, ataupun *website* yang berhubungan dengan :
   a. *Feature selection* untuk data *imbalance*
   b. *Filtering feature selection*
   c. *Wrapper feature selection*
   d. Klasifikasi dokumen *multi-label*
   e. Evaluation measures *multi-label*.
2. Pencarian dan pengumpulan data.
   Data yang akan digunakan berasal dari Reuters 21578.
3. Melakukan pembangunan perangkat lunak sebagai pendukung untuk melakukan pengujian dan analisis metode *feature selection* serta permasalahan yang ada.
4. Melakukan pengujian dan analisis terhadap data set yang akan diklasifikasikan dengan bantuan *software* Weka 3.5 maupun perangkat lunak yang dibangun, untuk mencatat tingkat akurasi yang didapat sesuai dengan metode *feature selection* dan teknik klasifikasi yang digunakan.
5. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

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

1. *Improved Odds Ratio* dan *Improved SIG* merupakan metode *feature selection* yang tepat digunakan untuk data *imbalance*, dibandingkan metode *Information Gain*, *Odds Ratio*, dan *GSS Coefficient*, hal ini disebabkan karena:
   a. jumlah atribut terpilih yang dihasilkan sangat sedikit, akan tetapi efisien dan efektif dalam meningkatkan keakuratan hasil klasifikasi.
   b. selain memilih atribut positif juga memilih atribut negatif yang berpengaruh pada data *imbalance*.
   c. metode IOR merupakan perkembangan metode *one-sided metric* (OR) sedangkan metode ISIG merupakan perkembangan metode *two-sided metric* (IG) dengan memperbaiki kekurangan yang ada pada masing-masing metode tersebut.
2. Metode *filter OR* lebih baik dari pada *wrapper OR,* diakibatkan karena pada *wrapper OR* lebih mempertimbangkan untuk meningkatkan nilai *recall* saja, yang berarti hanya memperkecil kesalahan dalam memprediksi dokumen yang seharusnya bernilai positif menjadi negatif.
3. Pada data *imbalance* yang memiliki label *balance*, penggunaan metode *improved OR* dan *improved SIG* tetap menghasilkan nilai keakuratan yang lebih baik daripada metode lainnya karena metode tersebut memilih fitur atau atribut negatif dan positif secara tepat.
4. Keberadaan negatif fitur dalam suatu dokumen merupakan indikator yang baik untuk menentukan bahwa dokumen tersebut bukan termasuk dalam suatu kategori. Oleh karena itu performansi klasifikasi dokumen dapat ditingkatkan dengan penolakan terhadap dokumen-dokumen yang tidak relevan.

5.2 Saran

Berikut ini saran-saran yang perlu dipertimbangkan untuk aplikasi lebih lanjut :

1. Menangani data imbalance dengan melihat dari dua sisi sekaligus yaitu menyeimbangkan jumlah data (*sampling*) dan mengurangi *feature set* (*feature selection*) serta mencoba menerapkan metode wrapper yang lain.
2. Aplikasi yang dibuat hanya terbatas untuk dokumen bahasa inggris, akan tetapi bisa mengelompokkan dokumen yang bersifat multilingual dengan cara menambahkan stopword selain bahasa Inggris dan akan lebih baik pula jika menerapkan proses stemming pada setiap bahasa tertentu.

Daftar Pustaka

[1] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[2] Tan, Pang-Ning, et all. *Introduction to Data Mining*. Pearson Education, Inc., Boston, 2006.

[3] Reuters 21578.

[4] Even, Year and Zohar. *Introduction to Text Mining.* Automated Learning group, National Center for Supercomputing Applications, University of Illinois, 2002.

[5] Sebastiani, Fabrizio. *A Tutorial on Automated Text Categorization*. Pisa, Italy : Istituto di Elaborazione dell' Informazione, 1999.

[6] Korpella, Mikko. *Introduction to Variable Selection*. September 26, 2006.

[7] Mladenic Dunja and Grobelnik Marko, *Feature Selection for Unbalanced Class Distribution and Naïve Bayes*, 1998.

[8] Z. Zheng, X. Wu, and R. Srihari. *Feature Selection for Text Categorization on Imbalanced Data*. SIKDD Explorations, 2004.

[9] Z. Zheng and R. Srihari. *Optimally Combining Positive and Negative Features for Text Categorization*. In Proceedings of the ICML'03 Workshop on Learning From Imbalanced Data Sets, 2003

[10] Y. Yang and J. Pedersen. *A Comparative Study on Feature Selection in Text Categorization*. The Fourteenth International Conference on Machine Learning, 1997.

[11] G. Tsoumakas and I. Katakis. *Multi-label Classification: An Overview*. International Journal of Data Warehousing and Mining, page accepted for publication, 2007.

[12] Yiming Yang and Xin Liu. A re-examination of text categorization methods. InMarti A. Hearst, Fredric Gey, and Richard Tong, editors, Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, pages 42–49, Berkeley, US, 1999. ACM Press, New York, US.

[13] Alexander Bergo. Text categorization and prototypes. 2001. Retrieved September 4, 2003 from the www.illc.uva.nl/Publications/ResearchReports/ MoL-2001-08.text.pdf.