METODE BOOSTING UNTUK KATEGORISASI BERITA BERBAHASA INDONESIA YANG MULTI-LABEL

Intan Nurma Yulita[1], Moch Arif Bijaksanaech[2], Yuliant Sibaroni[3]

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

Abstrak

Saat ini penggunaan internet telah memicu pertumbuhan dan pertukaran informasi menjadi jauh lebih pesat dibandingkan sebelumnya. Begitu pula dengan volume berita elektronik berbahasa Indonesia. Banyaknya jumlah berita tersebut dapat menyebabkan user mengalami kesulitan dalam mencari berita yang mereka inginkan. Text Categorization merupakan salah satu solusi yang dapat dilakukan, yaitu dengan cara mengelompokan berita kedalam kategori tertentu. Salah satu permasalahan dalam bidang Text Categorization adalah karakteristik data yang mempunyai lebih dari satu label (multi-label).

Salah satu metode Text Categorization untuk kasus multi-label adalah BoosTexter. BoosTexter adalah metode Boosting yang didesain khusus untuk kategorisasi teks. Boosting merupakan salah satu Ensemble Method yang menghasilkan classifier dengan akurasi tinggi melalui kombinasi weak hypotheses.

Untuk mengevaluasi performansi BoosTexter yang diimplementasikan, digunakan Hamming Loss, One Error, dan Coverage. Hasil yang didapat menunjukkan bahwa BoosTexter dapat memprediksi semua label aktual dari tiap instance serta menempatkan label actual pada rangking teratas dengan baik. Namun kelemahannya adalah dalam melakukan perangkingan semua label instances. Selain itu, kenaikan iterasi pada BoosTexter tidak mampu memperbaiki error iterasi tapi dapat memperbaiki nilai rata-rata error secara keseluruhan.

Kata Kunci : Text Categorization, Multi-Label, Boosting, BoosTexter

Abstract

Today, internet's using has made the growth and exchanging of informations become higher than before. And as same as with the volume of Indonesian electronic news. Large number of information can causes the users get into trouble in finding information that they want. Text Categorization, which is one of the solution for this problem, which is the task of assigning news to pre-specified categories of news. One of problem in Text Categorization is the characteristic data which have more than one label (multi-label).

One of Text Categorization method for multi-label case is BoosTexter. BoosTexter is method which developed from original Boosting and designed for text categorization. Boosting is one of Ensemble Method for creating a highly precise classifier by combining weak hypotheses.

For evaluating the performance of implemented BoosTexter, we used Hamming Loss, One Error, and Coverage. The result show that BoosTexter can predict all of instances labels and the top-ranked label was in the set of possible labels. But the weakness of BoosTexter is less rank to all of instances labels. Beside that, BoosTexter can not improve iteration error but it can improve overall error.

Keywords : Text Categorization, Multi-Label, Boosting, BoosTexter.

1. Pendahuluan

1.1 Latar belakang

Dengan era teknologi sekarang, internet menjadi sumber informasi yang paling banyak digunakan. Internet dengan *HTTP*-nya dapat dikatakan sebagai keajaiban dunia dalam bidang teknologi informasi. Namun berlimpahnya informasi, justru membuat para pengguna internet mengalami kesulitan untuk mendapatkan halaman web yang mereka inginkan.

Salah satu solusi untuk permasalahan ini adalah *Search Engine*. *Search Engine* merupakan salah satu aplikasi yang paling banyak digunakan saat ini untuk melakukan pencarian terhadap suatu dokumen. *Search Engine* bekerja dengan mencari halaman-halaman web yang dianggap paling relevan dengan permintaan (*query*) pengguna. Selain *Search Engine*, aplikasi lainnya adalah *aggregator*. *Aggregator* merupakan sebuah aplikasi portal yang secara otomatis mengelompokkan suatu informasi berdasarkan kategori-kategorinya. Salah satu agregator yang populer adalah *news aggregator* yang mengelompokkan berita berdasarkan kategori-kategorinya.

News aggregator hanyalah salah satu contoh aplikasi *aggregator*. *Aggregator* di bidang lain masih sangat banyak, misalnya untuk artikel-artikel ilmiah bidang ilmu computer, keislaman, informasi seputar perguruan tinggi dengan event-eventnya di Indonesia dan lain-lain. Salah satu *News aggregator* yang dikenal masyarakat adalah Google News (http://news.google.com). *Google News* merupakan *aggregator* pencarian berita dengan sumber data dari berbagai sumber berita namun *Google News* tidak dapat memproses sumber berita berbahasa Indonesia. Untuk memenuhi kebutuhan akan berita dalam berita berbahasa Indonesia dan kemudahan mendapatkannya maka diperlukan suatu aplikasi yang sejenis dengan *Google News* yang dapat mengelompokkan berita yang berasal dari berita berbahasa Indonesia.

Pengelompokkan berita dapat dilakukan dengan berbagai macam cara, salah satunya melalui kategorisasi. Kategorisasi dapat dibedakan menjadi dua jenis yaitu kategorisasi *single-label* dan kategorisasi multi-label. Kategorisasi berita digolongkan kategorisasi multi-label karena suatu berita bisa memiliki lebih dari satu kategori. Contohnya adalah berita ”Roy Marten tertangkap polisi saat pesta Narkoba”. Berita tersebut dapat dikategorisasikan ke dalam berita kriminal dan *entertainment*. Penanganan kategorisasi multi-label ini dapat dilakukan melalui metode Boosting. Metode Boosting merupakan salah satu metode yang cukup handal[16]. Boosting menggunakan serangkaian *classifier* di dalam membuat modelnya dan secara bertahap merubah distribusi *training* data dengan fokus pada data yang sukar untuk diklasifikasikan sehingga penggabungan rule pada setiap iterasinya akan menghasilkan satu *hypothesis* dengan tingkat akurasi yang lebih tinggi. Metode Boosting memiliki banyak varian, salah satunya adalah BoosTexter. BoosTexter merupakan metode Boosting yang secara khusus menangani kategorisasi teks.

1.2 Perumusan masalah

Dengan mengacu latar belakang di atas, maka permasalahan yang dibahas dan diteliti adalah :

1. Bagaimana menerapkan BoosTexter untuk kategorisasi berita berbahasa Indonesia yang multi-label .
2. Bagaimana melakukan pengujian dan analisis dari implementasi BoosTexter.

Sedangkan batasan masalah dalam tugas akhir ini adalah :
1. Berita yang digunakan adalah berita berbahasa Indonesia.
2. Pengambilan data diambil dari beberapa portal berita berbahasa Indonesia.
3. Tidak melakukan kategorisasi secara online.
4. Data input untuk proses *preprocessing* berupa file .txt sedangkan untuk proses *training* dan *testing* berupa fie .arff
5. Hanya menangani kategori data multi-label dengan tiga label.
6. Pemisahan data *training* dan data *testing* dari *dataset* dilakukan manual.
7. *Text* preprocessing diimplementasikan dalam tugas akhir ini tapi tidak menjadi fokus permasalahan dalam tugas akhir ini.
8. Data *training* dan data *testing* berupa bobot dari masing-masing *term* yang diperoleh melalui *text* preprocessing.
9. Hanya mengimplementasikan BoosTexter dengan AdaBoost.MH prediksi dan kehadiran bernilai *real* sebagai *weak hypotheses*

1.3 Tujuan

Berdasarkan pada masalah yang telah didefinisikan di atas, maka tujuan tugas akhir ini adalah :
1. Menerapkan metode BoosTexter untuk kategorisasi berita berbahasa Indonesia yang multi-label.
2. Melakukan analisis performansi metode BoosTexter berdasarkan Hamming Loss, One Error, Coverage.

Hipotesis awal dari tugas akhir ini :
1. BoosTexter memiliki performansi yang baik dalam melakukan kategorisasi berita berbahasa Indonesia yang multi-label.
2. *Error* BoosTexter semakin kecil pada setiap kenaikan iterasi.

1.4 Metodologi penyelesaian masalah

Metode penyelesaian masalah yang digunakan sebagai berikut :
1. Studi literatur
   Mencari referensi dan sumber-sumber lain yang berhubungan dengan Data Mining khususnya Web Mining, multi-label, dan metode BoosTexter.
2. Pengumpulan data
   Mencari data dari *website* berita Indonesia, dan *Data Understanding*..
3. Analisis dan perancangan perangkat lunak
   Menganalisis permasalahan yang akan diselesaikan dan menganalisis tahapan-tahapan yang digunakan untuk menyelesaikan permasalahan dengan metode *Object Oriented.*
4. Implementasi sistem
   Melakukan implementasi sistem dengan membangun perangkat lunak sesuai dengan perancangan yang telah di lakukan.

5. Pengujian Sistem dan Analisis Hasil
   Pengujian dilakukan pada metode BoosTexter terhadap parameter-parameter Hamming Loss, One Error, dan Coverage dan analisis dilakukan terhadap hasil yang diperoleh dari pengujian tersebut.
6. Pengambilan kesimpulan dan pembuatan laporan tugas akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. BoosTexter*, dengan penggunaan AdaBoost.MH dengan prediksi dan kehadiran bernilai *real* sebagai *weak hypotheses*, memiliki performansi yang baik dalam penempatan label aktual menjadi label dengan rangking tertinggi dan prediksi BoosTexter* terhadap semua label yang mungkin dimiliki oleh suatu *instance*. Namun kelemahan BoosTexter* adalah dalam penempatan rangking dari tiap label-label dari suatu *instances.* Secara keseluruhan, dapat disimpulkan performansi BoosTexter* sangat baik dalam penanganan kasus kategorisasi teks yang multi-label.
2. Kenaikan iterasi pada BoosTexter* belum tentu dapat memperbaiki nilai Hamming Loss, One Error, dan Coverage. Namun kenaikan iterasi memiliki *trend* penurunan *error* dan dapat menurunkan nilai *errror* rata-rata iterasi dari tiga parameter pengukuran tersebut.
3. BoosTexter* tidak berpengaruh pada kenaikan jumlah *term/feature* yang dihasilkan oleh *feature selection* dengan metode Term Contribution.

5.2 Saran

Saran terhadap pengembangan yang akan dilakukan terhadap tugas akhir ini adalah :

1. BoosTexter yang telah dilakukan dapat dikembangkan pada dokumen dengan jumlah label lebih dari tiga.
2. BoosTexter yang telah dilakukan dapat dikembangkan pada dokumen dengan jumlah berita yang lebih banyak.
3. BoosTexter diimplementasikan pada *weak hypotheses* dan *weak learner* yang berbeda.

Daftar Pustaka

[1] Asian, Jelita. *Effective Techniques for Indonesian Text Retrieval*. 2007. School of Computer Science and Information Technology, Science, Engineering, and Technology Portfolio, RMIT University : Australia.

[2] Dewi, Rani Charisma. 2005. *Pengelompokan Berita Berbahasa Indonesia Menggunakan Clustering*. Jurusan Teknik Informatika Sekolah Tinggi Teknologi Telkom : Bandung.

[3] Fahrudin, Tora. 2007. *Analisis dan Implementasi Metoda Databoost-IM (Studi Kasus Churn Prediction Mobile Telecomunication)*. Jurusan Teknik Informatika Sekolah Tinggi Teknologi Telkom : Bandung.

[4] Feldman, Ronen And James Sanger. 2007. *The Text Mining Handbook : Advanced Approaches Analyzing Unstructured Data*. Cambridge University Press.

[5] Freund, Yoav and Robert E. Schapire. *A Short Introduction to Boosting*. Journal of Japanese Society for Artificial Intelligence,14(5):771-780, September, 1999.

[6] Gonalves, Teresa And Paulo Quaresma. 2002. *A Preliminary Approach to The Multilabel Classification Problem of Portuguese Juridical Documents*. Departamento De Inform_Atica, Universidade De _Evora : Portugal.

[7] Han, Jiawei And Micheline Kamber. 2006. *Data Mining : Concepts and Techniques.* Intelligent Database Systems Research Lab, School Of Computing Science, Simon Fraser University.

[8] Hartoyo, Agus. 2008. *Indonesian Grapheme-To-Phoneme (G2p) Menggunakan Model Ig-Tree + Strategi Tebakan Terbaik*. Departemen Teknik Informatika Institut Teknologi Telkom : Bandung.

[9] http://cs.ucsd.edu/%7eaarvey/jboost/presentations/Boostinglightintro.Pdf

[10] http://en.wikipedia.org/wiki/Alternating_Decision_Tree

[11] http://jboost.com

[12] Polikar, Robi. *Ensemble Based System in Decision Making*. Third Quarter 2006. Electrical And Computer Engineering, Rowan University: Glassboro.

[13] Puspitarini, Wita. 2007. *Analisis Perbandingan Metode K-Nearest Neighbor (k-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Data Multi-Label*. Departemen Teknik Informatika Sekolah Tinggi Teknologi Telkom : Bandung.

[14] Rahmani, Luthfia. 2007. *Metode Feature Selection dalam Menangani Data Imbalance Pada Klasifikasi Dokumen Multi-Label*. Departemen Teknik Informatika Sekolah Tinggi Teknologi Telkom : Bandung.

[15] Riswanto, Ricky. 2007. *Metode Sampling dalam Menyelesaikan Data Text Imbalance untuk Klasifikasi Multi-Label*. Departemen Teknik Informatika Sekolah Tinggi Teknologi Telkom : Bandung.

[16] ˇSochman, Jan And Ji˘R´I Matas. 2007. *Adaboost*. Centre For Machine Perception Czech Technical University : Prague.

[17] Taira, Hirotoshi. 2003. *Text Categorization Using Machine Learning*. Department Of Information Processing Graduate School Of Information Science Nara Institute Of Science And Technology : Japan.

[18] Tsoumakas, G. And I. Katakis. 2007. *Multi-Label Classification: An Overview*. International Journal Of Data Warehousing And Mining, 3(3):1-13, 2007.

[19] Tsoumakas, G. And Ioannis Vlahavas. 2007. *Random K-Labelsets: An Ensemble*

Method for Multilabel Classification. Department Of Informatics, Aristotle University Of Thessaloniki : Thessaloniki, Greece.

[20] Vipin Kumar, Tan, Pang Nim. 2005. *Introduction to Data Mining*. Pearson Addison Wesley.

[21] Y. Schapire, R.E. Singer. 2000. *BoosTexter: A Boosting-Based System for Text Categorization. Machine Learning*. 2000.