

SISTEM REKOMENDASI BERITA BERDASARKAN HASIL KLASIFIKASI TOPIK FAVORIT PENGGUNA DI TWITTER MENGGUNAKAN ALGORITMA BACKPROPAGATION

Gelar Gunadi Putra¹, Eko Darwiyanto², Erda Guslinar Perdana³

¹Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Pengguna aktif Twitter menciptakan data begitu banyak. Data tersebut memiliki potensi informasi yang bisa didapatkan, salah satunya adalah informasi topik obrolan favorit setiap pengguna Twitter. Penelitian ini mengklasifikasi data untuk mendapatkan informasi tersebut, dan memanfaatkannya untuk merekomendasikan berita dari berbagai sumber. Hasil penelitian ini berguna bagi media massa yang ingin merekomendasikan beritanya kepada pengguna Twitter. Sistem klasifikasi yang dibangun di penelitian ini menggunakan jaringan saraf tiruan yang dilatih menggunakan algoritma backpropagation. Lalu, penelitian ini membuat aplikasi per Rekomendasi berita bagi pengguna Twitter menggunakan framework CodeIgniter, dan bahasa pemrograman PHP. Penggunaan backpropagation menghasilkan akurasi uji terbaik yang didapatkan adalah 82,4324% disaat akurasi latihan mencapai 94,6341%. Akurasi tersebut didapatkan menggunakan backpropagation dengan nilai learning rate sebesar 0,5 dan momentum sebesar 0,25. Sedangkan menurut 32 pengguna Twitter, 9% diantaranya menilai rekomendasi beritanya sangat akurat, 66% akurat, 16% biasa saja, 6% kurang akurat, dan 3% sangat tidak akurat.

Kata Kunci : Menambang data, backpropagation, Twitter

Abstract

Twitter active users create so many data. Those data have potential informations which can be obtained, one of them is the information about Twitter user favorite chat topics. This research classify data to get that information, and use it to recommend news from many sources. The result is useful for mass media who want to recommend their news to Twitter users. The classification system for this research is build using artificial neural network which is trained using backpropagation algorithm. Then, this research creates news recommendation application for Twitter user using CodeIgniter framework, and PHP programming language. Backpropagation utilization is resulting the best test accuracy at 82,4324% when the train accuracy achieve 94,6341%. This accuracy is obtained using backpropagation with learning rate value at 0.5 and momentum value at 0.25. But, according to 32 Twitter users, 9% think the news which recommended by the application is very accurate, 66% accurate, 16% indifferent, 6% not accurate., and 3% is not very accurate.

Keywords : Data Mining, backpropagation, Twitter.

1. Pendahuluan

1.1 Latar Belakang

Dunia teknologi informasi berkembang pesat. Data tersebar dengan sangat cepat dan dengan jumlah yang begitu besar melalui Internet. Pada tahun 2013, jumlah data yang melewati Internet akan mencapai 667 *exabytes* [8]. Salah satu penyumbang data terbesar adalah jejaring sosial, salah satunya adalah Twitter. Pengguna Twitter menggunakannya untuk membicarakan berbagai hal. Data yang melimpah ini bisa dimanfaatkan untuk menggali informasi tentang topik obrolan favorit pengguna. Hasil *data mining* tersebut digunakan untuk menyaring berita yang sesuai dengan topik favorit pengguna. Aplikasi ini bermanfaat bagi portal berita yang ingin memberikan berita kepada pengguna Twitter. Sehingga pengguna tidak harus mengonsumsi semua berita baru.

Pada tugas akhir ini, Twitter dipilih sebagai jejaring sosial yang dimanfaatkan untuk menyaring berita. Alasannya adalah karena peran Twitter sebagai salah satu media utama tempat persebaran berita di era jejaring sosial. Para media massa dan jurnalis menyebarkan tautan beritanya melalui akun resmi Twitter mereka. Menurut [9], 27% pengguna Twitter mendapatkan tautan artikel berita dari akun media massa dan jurnalis. Jumlah ini, melebihi data yang sama pada Facebook, yaitu sebesar 13%.

Penggunaan algoritma *Backpropagation* dilakukan karena menurut [5], algoritma ini memiliki kemampuan generalisasi yang baik. Di sisi lain, kata kunci pada dokumen Tweet tidak mutlak kaitannya dengan suatu rubrik berita. Hal ini memunculkan kasus XOR yang bisa diatasi dengan algoritma yang non-linear, dan *backpropagation* termasuk algoritma tersebut.

1.2 Perumusan masalah

Berdasarkan latar belakang di atas, maka rumusan masalah pada penelitian tugas akhir ini adalah:

1. Bagaimana cara mengambil data Twitter pengunjung situs berita ?
2. Bagaimana cara mengolah dan mengklasifikasi data Twitter menggunakan algoritma *backpropagation* ?
3. Bagaimana memanfaatkan data Twitter masing – masing pengunjung situs berita agar berita yang ditawarkan relevan sesuai topik obrolan favoritnya?
4. Bagaimana keakuratan algoritma *backpropagation* dalam memberikan berita yang relevan sesuai topik obrolan favorit pengunjung situs berita ?

1.3 Batasan Masalah

Penelitian tugas akhir ini memiliki batasan masalah sebagai berikut:

1. Data yang diteliti oleh aplikasi berasal dari jejaring sosial Twitter.
2. Data yang diteliti berupa teks berbahasa Indonesia.
3. Sistem tidak menganalisa teks yang cara penulisannya tidak umum.

4. Situs media massa yang digunakan sebagai sumber berita memiliki tautan RSS terpisah untuk setiap rubrik berita.

1.4 Tujuan

Penelitian tugas akhir ini bertujuan sebagai berikut:

1. Menerapkan algoritma *backpropagation* untuk mengklasifikasi data teks Twitter agar mengetahui topik obrolan favorit pengguna.
2. Menganalisa keakuratan algoritma *backpropagation* dalam mengklasifikasi topik obrolan favorit pengguna Twitter. Analisa keakuratan ini sebagai dasar penilaian keberhasilan portal berita dalam menyampaikan berita relevan.
3. Mencari nilai *learning rate* dan momentum terbaik dalam menghasilkan akurasi uji tertinggi.
4. Mencari arsitektur JST yang menghasilkan akurasi uji tertinggi.

1.5 Metodologi penyelesaian masalah

Penelitian tugas akhir ini diselesaikan dengan metodologi penyelesaian masalah sebagai berikut :

1. *Studi literature* atau mempelajari teori dasar mengenai *data mining*, *preprocessing* data teks, klasifikasi data menggunakan algoritma *backpropagation*, RSS aggregator, API (*Application Programming Interface*) pencarian konten Twitter dan topik lain yang terkait dengan judul penelitian dari berbagai sumber referensi, seperti: buku, jurnal, dan artikel.
2. Pengumpulan data latih dan data uji. Selain itu, mengumpulkan tautan RSS berbagai situs media massa.
3. Membuat perangkat lunak berupa aplikasi pengumpul berita berbasis web. Sistem perangkat lunak tersebut terdiri dari tiga proses besar, yaitu: *preprocessing*, klasifikasi menggunakan *backpropagation*, dan pengambilan konten berita yang relevan dengan hasil klasifikasi. Pada proses *preprocessing*, data teks Twitter diolah melalui tiga proses, yaitu: tokenisasi, penyaringan, dan *stemming*. Proses tokenisasi adalah proses menghapus tanda baca, memecah teks menjadi kumpulan kata, dan mengganti karakter non-teks menjadi spasi tunggal. Proses penyaringan adalah menghilangkan kata yang tingkat kepentingannya rendah, dan proses *stemming* menghilangkan imbuhan pada semua kata hingga menjadi kata dasar. Data hasil *preprocessing* diklasifikasi menggunakan algoritma *backpropagation* menjadi 9 kelas rubrik berita, yaitu: bola, dunia, nasional, otomotif, teknologi, ekonomi, metro, olahraga, dan travel. Hasil klasifikasi tersebut digunakan untuk memunculkan berita relevan dari berbagai sumber media massa.
4. Analisa keakuratan algoritma *backpropagation* dalam mengklasifikasikan data Twitter. Selain itu, menganalisa hasil survey tentang relevansi berita yang dimunculkan sistem buatan penelitian tugas akhir ini kepada beberapa pengguna Twitter sebagai responden.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Penelitian ini menghasilkan beberapa kesimpulan sebagai berikut:

1. Akurasi uji terbaik yang didapatkan adalah 82,4324% disaat akurasi latihan mencapai 94,6341%. Akurasi tersebut didapatkan menggunakan *backpropagation* dengan nilai *learning rate* sebesar 0,5 dan momentum sebesar 0,25.
2. Arsitektur JST yang menghasilkan akurasi uji terbaik terdiri dari 251 neuron di lapisan masukan (250 neuron masukan dan 1 neuron bias), 14 neuron di lapisan tersembunyi (13 neuron tersembunyi dan 1 neuron bias), dan 9 neuron di lapisan keluaran.
3. Dari 32 responden yang menggunakan Twitter, sebanyak 22% dari seluruh responden menyatakan sangat tertarik dengan konsep aplikasi ini, 66% menyatakan tertarik, dan 12% menyatakan biasa saja. Tidak ada responden yang menyatakan tidak tertarik, dan sangat tidak tertarik.
4. Dari seluruh responden, 9% diantaranya menilai rekomendasi beritanya sangat akurat, 66% akurat, 16% biasa saja, 6% kurang akurat, dan 3% sangat tidak akurat.

5.2 Saran

Ada beberapa saran yang penulis berikan untuk kelanjutan penelitian ini, antara lain sebagai berikut:

1. Kurangi jumlah neuron masukan agar memperdalam proses latihan JST.
2. Pisahkan proses klasifikasi setiap kelas berita, dan lakukan klasifikasi data menjadi sub kelas berita. Contohnya: proses klasifikasi kelas “ekonomi” dipisah dengan kelas lainnya. Lalu klasifikasi data menjadi subkelas: “makro”, “mikro”, “keuangan”, “wirausaha”, dll. Selain itu, pisahkan 250 kata kunci (koleksi kata dalam “*trending_topic_id*”) setiap kelas. Dengan trik ini, 250 kata kunci tersebut hanya mewakili setiap kelas, bukan 9 kelas. Trik ini dapat memperkaya koleksi kata yang bisa diklasifikasi oleh aplikasi, tanpa memperberat proses latihan JST.
3. Jika pengguna tidak aktif di Twitter dalam seminggu terakhir, gunakan data selain tweet untuk mengklasifikasi topik favorit pengguna Twitter, misalnya menggunakan data akun yang diikutinya.
4. Gunakan algoritma *Support Vector Machine* (SVM) dengan *kernel trick*. Karena algoritma ini mengklasifikasi data secara nonlinear, dan dapat mengatasi kasus XOR (lihat gambar 2-4).

Daftar pustaka

- [1] Alexander, D., Data Mining, 2013, <http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/>, diakses tanggal 8 April 2013.
- [2] Elder, J., Miner, G., Nisbet, B., 2012, "Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications", Oxford: Elsevier Inc., h. xxxi.
- [3] Keke, D., Chikita, R., Prayogo, A.D., 2012, "Sistem Temu Balik Informasi: Algoritma nazief dan Andriani", Yogyakarta: Universitas Gajah Mada.
- [4] Liu, M., Yang, J., 2012, "An Improvement of TFIDF Weighting in Text Categorization", International Conference on Computer Technology and Science 2012.
- [5] Ramasundaram, S., Victor, S.P., 2010, "Text Categorization by Backpropagation Network", International Journal of Computer Applications (0975 – 8887), Volume 8-No.6, Oktober 2010.
- [6] Soucy, P., Mineau, G.W., 2005, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model", 19th International Joint Conference on Artificial Intelligence, hh. 1130-1135.
- [7] Suyanto, ST, MSc., 2011, "Artificial Intelligence : Searching, Reasoning, Planning, dan Learning", Bandung: Informatika Bandung, h. 164.
- [8] The Economist, 2010, "Data, Data Everywhere", <http://www.economist.com/node/15557443>, diakses tanggal 8 April 2013.
- [9] The Journalism, 2012, "Study: Half of News Articles Shared on Twitter are BBC News Stories", <http://www.journalism.co.uk/news/bbc-news-accounts-for-half-twitter-referrals/s2/a549543/>, diakses tanggal 8 April 2013.
- [10] Willamette, 2014, "Momentum and Learning Rate Adaptation", <http://www.willamette.edu/~gorr/classes/cs449/momrate.html>, diakses tanggal 16 Januari 2014.

Telkom
University