

KOMPARASI ALGORITMA KLASIFIKASI *TEXT MINING* UNTUK ANALISIS SENTIMEN PADA REVIEW RESTORAN

Dinda Ayu Muthia

Manajemen Informatika
AMIK BSI Bekasi
<http://www.bsi.ac.id>
dinda.dam@bsi.ac.id

Abstract—*Online review sites are growing in popularity as more and more people seek advice from fellow users about their services and products. The results of recent years research have also evolved in the field of sentiment analysis in order to find the right solution in the system that can automatically analyze the above and extract the most relevant information for the user. In a previous study of sentiment analysis on restaurant reviews, the algorithm of Naive Bayes is superior to the Support Vector Machine. In this research used two algorithms, namely Naive Bayes and Support Vector Machine. The goal is to determine the best algorithm that can be used for data review of Indonesian text. From the results of data processing, Naive Bayes algorithm is superior to the Support Vector Machine with 87% accreditation rate. While the Support Vector Machine algorithm only produces 56% accounting. The author makes the application of sentiment analysis using Java programming language as research support.*

Keywords: *Sentiment Analysis, Naive Bayes, Support Vector Machine, Review, Text Mining.*

Intisari—Situs review online terus bertambah populer karena semakin banyak orang mencari saran dari sesama pengguna mengenai layanan dan produk. Sejumlah penelitian beberapa tahun terakhir juga sudah berkembang dalam bidang analisis sentimen guna menemukan solusi yang tepat dalam membuat sistem yang dapat secara otomatis menganalisis review di internet dan mengekstrak informasi yang paling relevan bagi pengguna. Dalam penelitian sebelumnya mengenai analisis sentimen pada review restoran, akurasi algoritma *Naive Bayes* lebih unggul dari *Support Vector Machine*. Pada penelitian ini digunakan dua algoritma, yakni *Naive Bayes* dan *Support Vector Machine*. Tujuannya adalah untuk menentukan algoritma terbaik yang bisa digunakan untuk data review teks bahasa Indonesia. Dari hasil pengolahan data, algoritma *Naive Bayes* lebih unggul dari *Support Vector Machine* dengan tingkat akurasi sebesar 87%. Sedangkan algoritma *Support Vector Machine* hanya menghasilkan

akurasi 56%. Penulis membuat aplikasi analisis sentiment menggunakan bahasa pemrograman Java sebagai penunjang penelitian.

Kata Kunci: *Analisis Sentimen, Naive Bayes, Support Vector Machine, Review, Text Mining.*

PENDAHULUAN

Situs review online terus bertambah populer karena semakin banyak orang mencari saran dari sesama pengguna mengenai layanan dan produk (Brody, 2010). Menemukan informasi yang relevan dan tepat waktu dari berbagai review ini sangat penting (Azam & Yao, 2012). Hal ini menyebabkan peningkatan penelitian di bidang analisis sentimen dan *opinion mining*, dengan tujuan menyediakan sistem yang dapat secara otomatis menganalisis review pengguna dan mengekstrak informasi yang paling relevan bagi pengguna (Brody, 2010).

Seiring berkembangnya situs penyedia review online, penelitian pun sudah banyak dilakukan dalam bidang analisis sentimen. Penelitian sebelumnya diantaranya, analisis sentimen pada review restoran, di mana review tersebut ditulis dalam bahasa Canton dan diklasifikasikan menggunakan algoritma *Naive Bayes* dan *Support Vector Machine* (Zhang, Ye, Zhang, & Li, 2011). Analisis sentimen pada review tempat tujuan perjalanan menggunakan algoritma *Naive Bayes*, *Support Vector Machine*, dan *Character Based N-gram Model* (Ye, Zhang, & Law, 2009). Analisis sentimen pada review film dan review multi domain lainnya, seperti buku, DVD, barang elektronik, dan lain-lain yang diambil dari situs Amazon.com menggunakan algoritma *Lexicon Labeling*, *Heuristic Labeling*, *Self-Labeling*, *Naive Bayes*, *Support Vector Machine*, dan *Maximum Entropy* (He & Zhou, 2011). Analisis sentimen pada review film menggunakan algoritma *Naive Bayes*, *Support Vector Machine*, dan *SO-PMI-IR* (Waila, Singh, & Singh, 2012). Analisis sentimen pada review kamera, kamp, dokter, oba, laptop, pengacara, film, musik, radio, dan TV menggunakan algoritma *Naive Bayes*,

Maximum Entropy, Decision Tree, K Nearest Neighbor, dan Support Vector Machine (Wang, Sun, Ma, Xu, & Gu, 2014).

Algoritma *Naive Bayes* sangat sederhana dan efisien (Chen, Huang, Tian, & Qu, 2009). Di samping itu, algoritma *Naive Bayes* merupakan teknik *machine learning* yang populer untuk klasifikasi teks, dan memiliki performa yang baik pada banyak domain (Ye et al., 2009). Akan tetapi, *Naive Bayes* ternyata mempunyai kekurangan, yaitu sangat sensitif dalam pemilihan fitur (Chen et al., 2009). Jumlah fitur yang terlalu banyak dalam proses klasifikasi, tidak hanya meningkatkan waktu penghitungan tetapi juga menurunkan akurasi (Uysal & Gunal, 2012).

Support Vector Machine telah digunakan secara efisien dalam banyak studi klasifikasi teks karena manfaat utamanya seperti robust pada area dengan perspektif tinggi, setiap fungsi sesuai, kuat bila ada sekumpulan sampel secara sporadis, dan sebagian besar masalah klasifikasi teks bebas linear. Apalagi *Support Vector Machine* sudah mendapatkan hasil yang baik dalam *opinion mining* dan metode ini telah melampaui metode *machine learning* lainnya. Namun *Support Vector Machine* mempunyai kekurangan, yaitu hasil akurasi dipengaruhi oleh pemilihan parameter yang sesuai (Basari, Hussin, Ananta, & Zeniarja, 2013).

Dalam penelitian analisis sentimen pada review restoran (Zhang et al., 2011), akurasi algoritma *Naive Bayes* lebih unggul dari *Support Vector Machine*. Oleh karena itu, penelitian ini bertujuan untuk menentukan algoritma terbaik dalam analisis sentimen pada review restoran dengan teks bahasa Indonesia menggunakan algoritma *Naive Bayes* dan *Support Vector Machine*.

BAHAN DAN METODE

Metode penelitian yang penulis lakukan adalah metode penelitian eksperimen, dengan tahapan sebagai berikut:

- a. Pengumpulan Data
Penulis menggunakan data yang diambil dari situs review restoran www.zomato.com yang terdiri dari 100 review positif dan 100 review negative dengan teks bahasa Indonesia.
- b. Pengolahan Awal Data
Dataset ini dalam tahap *preprocessing* harus melalui 2 proses, yaitu:
 - 1) Tokenization
Yaitu mengumpulkan semua kata yang muncul dan menghilangkan tanda baca maupun simbol apapun yang bukan huruf.
 - 2) Generate N-grams

Yaitu menggabungkan kata sifat yang seringkali muncul untuk menunjukkan sentimen, seperti kata “sangat” dan kata “bagus”. Kata “bagus” memang sudah menunjukkan sentimen bentuk opini positif. Kata “sangat” tidak akan berarti jika berdiri sendiri. Namun jika dua kata tersebut digabung menjadi “sangat bagus”, maka akan sangat menguatkan opini positif tersebut. Penulis menggunakan penggabungan tiga kata, yang disebut 2-grams (Bigrams).

- c. Metode yang Digunakan
Metode yang penulis bandingkan adalah algoritma *Naive Bayes* dan *Support Vector Machine*. Di mana kedua metode tersebut sudah sangat populer digunakan dalam penelitian di bidang analisis sentimen, klasifikasi teks, ataupun *opinion mining*.
- d. Eksperimen dan Pengujian Metode
Untuk eksperimen data penelitian, penulis menggunakan Rapid Miner Studio untuk mengolah data.
- e. Evaluasi dan Validasi Hasil
Validasi dilakukan menggunakan 10 *fold cross validation*. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC untuk mengukur nilai AUC.

HASIL DAN PEMBAHASAN

1. Hasil

Data training yang digunakan dalam pengklasifikasian teks ini terdiri dari 100 review restoran positif dan 100 review restoran negatif dengan teks bahasa Indonesia. Data tersebut masih berupa sekumpulan teks yang terpisah dalam bentuk dokumen. Sebelum diklasifikasikan, data tersebut harus melalui beberapa tahapan proses agar bisa diklasifikasikan dalam proses selanjutnya. Tabel 1 menunjukkan hasil proses pengolahan data awal *tokenization*. Tabel 2 menunjukkan proses *generate 2-grams*.

Tabel 1. Hasil Proses *Tokenization*

Review	Tokenization
Pilihan yg cukup oke untuk ngobrol2 dan nyemil2 di daerah depok. Karna selain menawarkan serabi ditempat ini juga terdapat beberapa tenant seperti batagor ihsan dan bubur barito Gw sering ihsan dan bubur barito. Gw sering bgt kesini waktu s1 di kampus ui depok. Selain karna	Pilihan yg cukup oke untuk ngobrol dan nyemil di daerah depok Karna selain menawarkan serabi ditempat ini juga terdapat beberapa tenant seperti batagor ihsan dan bubur barito Gw sering bgt kesini waktu s di kampus ui depok Selain karna dekat makanan disini juga murah dan cukup variatif Yang pernah gw coba adalah

deket, makanan disini juga murah dan cukup variatif.

Yang pernah gw coba adalah berbagai macam serabi manis, bakso malang, tempe mendoan, dan batagor. Serabinya enak semua. Bakso malang juga enak. Tp kalo untuk batagornya biasa aja dan agak ga worth it utk harga pinggir jalan. Kalo pelayanannya cukup sigap dan cepat.

Overall sih oke lah untuk dicoba di daerah depok. Tp selain depok ini juga ada di daerah rawamangun.

berbagai macam serabi manis bakso malang tempe mendoan dan batagor Serabinya enak semua Bakso malang juga enak Tp kalo untuk batagornya biasa aja dan agak ga worth it utk harga pinggir jalan Kalo pelayanannya cukup sigap dan cepat Overall sih oke lah untuk dicoba di daerah depok Tp selain depok ini juga ada di daerah rawamangun

waktu waktu_s s s_di di di_kampus kampus kampus_ui ui ui_depok depok depok_Selain Selain Selain_karna karna karna_deket deket deket_makanan makanan makanan_disini disini disini_juga juga juga_murah murah murah_dan dan dan_cukup cukup cukup_variatif variatif variatif_Yang Yang Yang_pernah pernah pernah_gw gw_coba coba coba_adalah adalah adalah_berbagai berbagai berbagai_macam macam macam_serabi serabi serabi_manis manis manis_bakso bakso bakso_malang malang malang_tempe tempe tempe_mendoan mendoan mendoan_dan dan dan_batagor batagor batagor_Serabinya Serabinya Serabinya_enak enak enak_semua semua semua_Bakso Bakso Bakso_malang malang malang_juga juga juga_enak enak enak_Tp Tp Tp_kalo kalo kalo_untuk untuk untuk_batagornya batagornya batagornya_biasa biasa biasa_aja aja aja_dan dan dan_agak agak agak_ga ga ga_worth worth worth_it it it_untuk utk utk_harga harga harga_pinggir pinggir pinggir_jalan jalan jalan_Kalo Kalo Kalo_pelayanannya pelayanannya pelayanannya_cukup cukup cukup_sigap sigap sigap_dan dan dan_cepat cepat cepat_Overall Overall Overall_sih sih sih_oke oke oke_lah lah lah_untuk untuk untuk_dicoba dicoba dicoba_di di di_daerah daerah daerah_depok

Sumber: (Muthia, 2018)

Tabel 2. Hasil Proses *Generate 2-grams*
Tokenization *Generate 2-grams*

Pilihan yg cukup oke untuk ngobrol dan nyemil di daerah depok Karna selain menawarkan serabi ditempat ini juga terdapat beberapa tenant seperti batagor ihsan dan bubur barito Gw sering bgt kesini waktu s di kampus ui depok Selain karna deket makanan disini juga murah dan cukup variatif Yang pernah gw coba adalah berbagai macam serabi manis bakso malang tempe mendoan dan batagor Serabinya enak semua Bakso malang juga enak Tp kalo untuk batagornya biasa aja dan agak ga worth it utk harga pinggir jalan Kalo pelayanannya cukup sigap dan cepat Overall sih oke lah untuk dicoba di daerah depok Tp selain depok ini juga ada di daerah rawamangun	Pilihan Pilihan_yg yg yg_cukup cukup cukup_oke oke oke_untuk untuk untuk_ngobrol ngobrol ngobrol_dan dan dan_nyemil nyemil nyemil_di di di_daerah daerah daerah_depok depok depok_Karna Karna Karna_selain selain selain_menawarkan menawarkan menawarkan_serabi serabi serabi_ditempat ditempat ditempat_ini ini ini_juga juga juga_terdapat terdapat terdapat_beberapa beberapa beberapa_tenant tenant tenant_seperti seperti seperti_batagor batagor batagor_ihsan ihsan ihsan_dan dan dan_bubur bubur bubur_barito barito barito_Gw Gw Gw_serang sering sering_bgt bgt bgt_kesini kesini kesini_waktu
---	---

depok	depok_Tp	Tp
Tp_selain	selain	selain
selain_depok	depok	depok
depok_ini	ini	ini_juga
juga_juga_ada	ada	ada_di
di_di_daerah	daerah	daerah
daerah_rawamangun		rawamangun
rawamangun		

Sumber: (Muthia, 2018)

Data-data di atas selanjutnya akan masuk ke proses klasifikasi yang berguna untuk menentukan sebuah kalimat sebagai anggota *class* positif atau *class* negatif berdasarkan nilai perhitungan probabilitas dari rumus *Bayes* yang lebih besar. Jika hasil probabilitas kalimat tersebut untuk *class* positif lebih besar dari pada *class* negatif, maka kalimat tersebut termasuk ke dalam *class* positif. Jika probabilitas untuk *class* positif lebih kecil dari pada *class* negatif, maka kalimat tersebut termasuk ke dalam *class* negatif.

Pada penelitian ini, pengujian model dilakukan dengan menggunakan teknik *10 cross validation*. Proses ini membagi data secara acak ke dalam 10 bagian. Proses pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya. Kemudian proses akurasi dihitung dengan melihat seberapa banyak data yang sudah terklasifikasi dengan benar.

Hasil pengujian model akan dibahas melalui *confusion matrix* untuk menunjukkan seberapa baik model yang terbentuk. Tabel 3 menunjukkan *confusion matrix* algoritma *Naïve Bayes*. Tabel 4 menunjukkan *confusion matrix* algoritma *Support Vector Machine*.

Tabel 3. *Confusion Matrix* Algoritma *Naïve Bayes*

Akurasi Naive Bayes: 87.00% +/- 7.48% (mikro: 87.00%)

	<i>True Positive</i>	<i>True Negative</i>	<i>Class precision</i>
<i>Pred. Positive</i>	90	16	84.91%
<i>Pred. Negative</i>	10	84	89.36%
<i>Class recall</i>	90.00%	84.00%	

Sumber: (Muthia, 2018)

Tabel 4. *Confusion Matrix* Algoritma *Support Vector Machine*

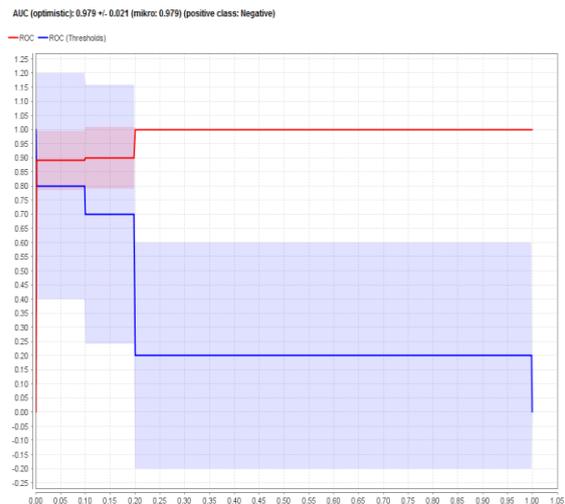
Akurasi Naive Bayes: 56.00% +/- 5.83% (mikro: 56.00%)

	<i>True Positive</i>	<i>True Negative</i>	<i>Class precision</i>

<i>Pred. Positive</i>	99	87	53.23%
<i>Pred. Negative</i>	1	13	92.86%
<i>Class recall</i>	99.00%	13.00%	

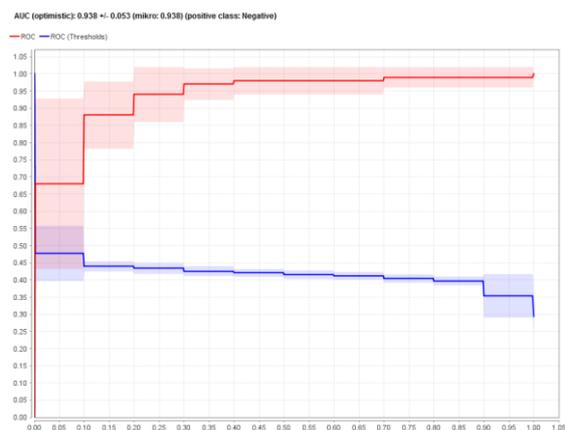
Sumber: (Muthia, 2018)

Berikut adalah tampilan kurva ROC dari hasil uji data. Gambar 1 adalah kurva ROC dari algoritma *Naïve Bayes*. Gambar 2 adalah kurva ROC dari algoritma *Support Vector Machine*.



Sumber: (Muthia, 2018)

Gambar 1. Kurva ROC dari algoritma *Naïve Bayes*



Sumber: (Muthia, 2018)

Gambar 2. Kurva ROC dari algoritma *Support Vector Machine*

2. Pembahasan

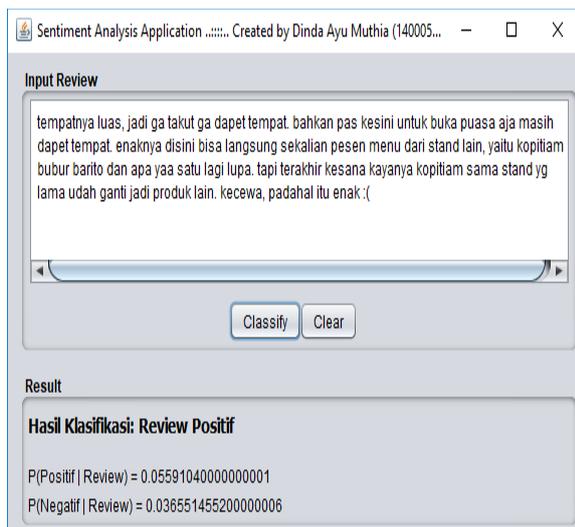
Algoritma *Naïve Bayes* dan *Support Vector Machine* memang populer dalam mengklasifikasikan teks, khususnya untuk bidang penelitian analisis sentiment. Keduanya memiliki performa yang baik dan hasil akurasi yang tinggi. Namun, tiap penelitian menggunakan data yang berbeda-beda, sehingga hasil akurasi pun beragam.

Dalam penelitian ini, digunakan data review teks bahasa Indonesia, di mana umumnya data yang digunakan teks dalam bahasa Inggris. Hal tersebut kemungkinan menjadi salah satu penyebab hasil akurasi algoritma *Support Vector Machine* tidak optimal. Selain itu, penggunaan *generate 2-grams* membuat data yang diujikan menjadi lebih banyak dari sebelumnya. Bisa dilihat, bahwa algoritma *Naïve Bayes* unggul dengan menghasilkan akurasi sebesar 87%. Sedangkan algoritma *Support Vector Machine* hanya menghasilkan akurasi sebesar 56%.

Akurasi di tingkat 0.80-0.90 dianggap sebagai "*good classification*" atau klasifikasi yang bagus, sedangkan akurasi di tingkat 0.50-0.60 dianggap "*failure*" atau kegagalan. Bisa diartikan, dalam penelitian ini algoritma *Naïve Bayes* masuk ke dalam pengklasifikasi yang bagus, dengan hasil akurasi di atas 0.80 atau 80%.

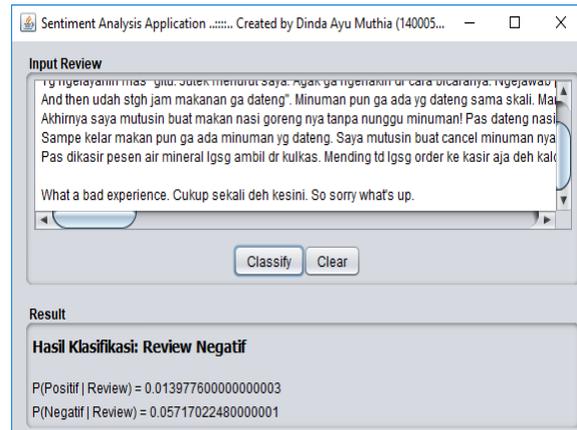
a. Implementasi

Untuk menunjang penelitian ini, penulis merancang aplikasi sederhana berbasis desktop. Berikut adalah tampilan rancangan aplikasi dan implementasinya. Penulis membuat aplikasi analisis sentimen menggunakan bahasa pemrograman Java. Gambar 3 menunjukkan sebuah review positif restoran diinput ke dalam aplikasi tersebut langsung menampilkan hasilnya berdasarkan perhitungan algoritma *Naïve Bayes*. Sedangkan Gambar 4 menunjukkan hasil klasifikasi review negatif.



Sumber: (Muthia, 2018)

Gambar 3. Aplikasi Analisis Sentimen menunjukkan hasil klasifikasi review positif



Sumber: (Muthia, 2018)

Gambar 4. Aplikasi Analisis Sentimen menunjukkan hasil klasifikasi review negatif

KESIMPULAN

Dari pengolahan data yang sudah dilakukan, terbukti algoritma *Naïve Bayes* lebih unggul dari algoritma *Support Vector Machine* dalam mengklasifikasi review restoran dengan teks bahasa Indonesia. Akurasi algoritma *Naïve Bayes* mencapai 87%, sedangkan algoritma *Support Vector Machine* hanya menghasilkan akurasi sebesar 56%. Perbedaan akurasinya cukup jauh, sebesar 31%. Untuk mendukung penelitian ini, penulis mengembangkan aplikasi review restoran untuk mengklasifikasikan review positif dan negatif menggunakan bahasa pemrograman Java.

REFERENSI

- Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. *Expert Systems with Applications*, 39(5), 4760–4768. <https://doi.org/10.1016/j.eswa.2011.09.160>
- Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, 53, 453–462. <https://doi.org/10.1016/j.proeng.2013.02.059>
- Brody, S. (2010). An Unsupervised Aspect-Sentiment Model for Online Reviews, (June), 804–812.
- Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with *Naïve Bayes*. *Expert Systems with*

- Applications*, 36(3), 5432–5435.
<https://doi.org/10.1016/j.eswa.2008.06.054>
- He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4), 606–616.
<https://doi.org/10.1016/j.ipm.2010.11.003>
- Muthia, D. A. (2018). *Laporan Akhir Penelitian Dosen Yayasan*. Jakarta.
- Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, 36, 226–235.
<https://doi.org/10.1016/j.knosys.2012.06.005>
- Waila, P., Singh, V. K., & Singh, M. K. (2012). Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews. *2012 IEEE International Conference on Computational Intelligence and Computing Research*, (c), 1–6.
<https://doi.org/10.1109/ICCIC.2012.6510235>
- Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, 57, 77–93.
<https://doi.org/10.1016/j.dss.2013.08.002>
- Ye, Q., Zhang, Z., & Law, R. (2009). Expert Systems with Applications Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems With Applications*, 36(3), 6527–6535.
<https://doi.org/10.1016/j.eswa.2008.07.035>
- Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, 38(6), 7674–7682.
<https://doi.org/10.1016/j.eswa.2010.12.147>