Prosiding Komputeran UTM
Inovasi di dalam Teknologi dan Aplikasi Komputeran
Volume: I | Year: 2016 | ISBN: 978-967-0194-82-0

Pengelasan Bahasa Kesat Menggunakan Pemberat Istilah Sebagai Pemilihan Ciri Bagi Kandungan Laman Web

RIDZWAN BIN MOHAMED @ HUSSIN [1] ***, ROSELINA SALLEH@SALLEHUDDIN*** [*2]

Jabatan Sains Komputer, Fakulti Komputeran, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia

[1] *ridzwanmh@gmail.com,* [2] *roselina@utm.my*

Abstrak

Kini kebanyakan urusan seharian kita melibatkan penggunaan laman web di internet. Bagaimanapun, terdapat beberapa laman web yang tidak sesuai dilayari kerana mengandungi bahasa kesat yang melampau. Pendedahan seumpama ini terhadap golongan remaja dan kanak-kanak boleh menyebabkan atau menyumbang kepada berlakunya gejala-gejala negatif seperti kejadian rogol dan pembuangan bayi. Justeru, bagi mencegah kanak-kanak daripada bebas melayari laman web yang tidak baik tersebut, pengesanan isi kandungan laman web menggunakan model pengelasan Support Vector Machine (SVM) boleh dilaksanakan. Untuk meningkatkan prestasi pengelas SVM bagi pengesanan bahasa kesat dalam isi kandungan laman web, dua skim pemberat istilah digunakan sebagai pemilih ciri iaitu Kekerapan Istilah (TF) dan Kekerapan Istilah Songsang Kekerapan Dokumen (TFIDF). Prestasi ketepatan SVM menggunakan kedua-dua skim pemberat ini diukur dan dibandingkan menggunakan data yang diperolehi dari laman web yang sama. Keputusan eksperimen menujukkan kedua-dua skim pemberat menghasilkan ketepatan yang sama iaitu 70%. Keputusan ini menunjukkan bahawa TF dan TFIDF sesuai digunakan sebagai pemilih ciri bagi meningkatkan prestasi pengelasan SVM untuk pengesanan kandungan bahasa kesat dalam laman web.

Kata Kunci: Bahasa Kesat, Skim Pemberat Istilah, TF, TFIDF

1.0 Pendahuluan

Kini kebanyakan urusan seharian kita dijalankan dengan menggunakan laman web. Hampir semua golongan bebas melayari laman web termasuk juga golongan remaja dan kanak-kanak. Walau bagaimanapun, terdapat sesetengah laman web mengandungi penggunaan bahasa kesat yang melampau pada isi kandungannya. Hal ini boleh memberikan kesan negatif seperti kejadian rogol dan pembuangan bayi. Bagi mengatasi masalah penggunaan bahasa kesat yang melampau pada isi kandungan laman web, pengelasan bahasa kesat memnggunakan skim pemberat istilah sebagai pemilihan ciri bagi isi kandungan laman web harus dilaksanakan. Dengan menggunakan skim pemberat

istilah sebagai pemilihan ciri, istilah-istilah yang dianggap kesat pada laman web boleh dikelaskan. Dalam kajian ini, skim pemberat istilah Kekerapan Istilah (TF) dan Kekerapan Istilah Songsang Kekerapan Dokumen (TFIDF) digunakan. Daripaad itu, objektif kajian ini adalah (i)mengenal pasti dan mengelaskan istilah atau bahasa kesat dengan melibatkan kepakaran manusia, (ii) menganalisis istilah atau bahasa kesat itu tadi dengan menggunakan skim pemberat istilah Kekerapan Istilah (TF) dan Kekerapan Istilah Songsang Kekerapan Dokumen (TFIDF) dengan menggunakan *Support Vector Machine* (SVM), dan (iii) membuat perbandingan antara kedua-dua skim pemberat istilah.

2.0 Metodologi Pembangunan

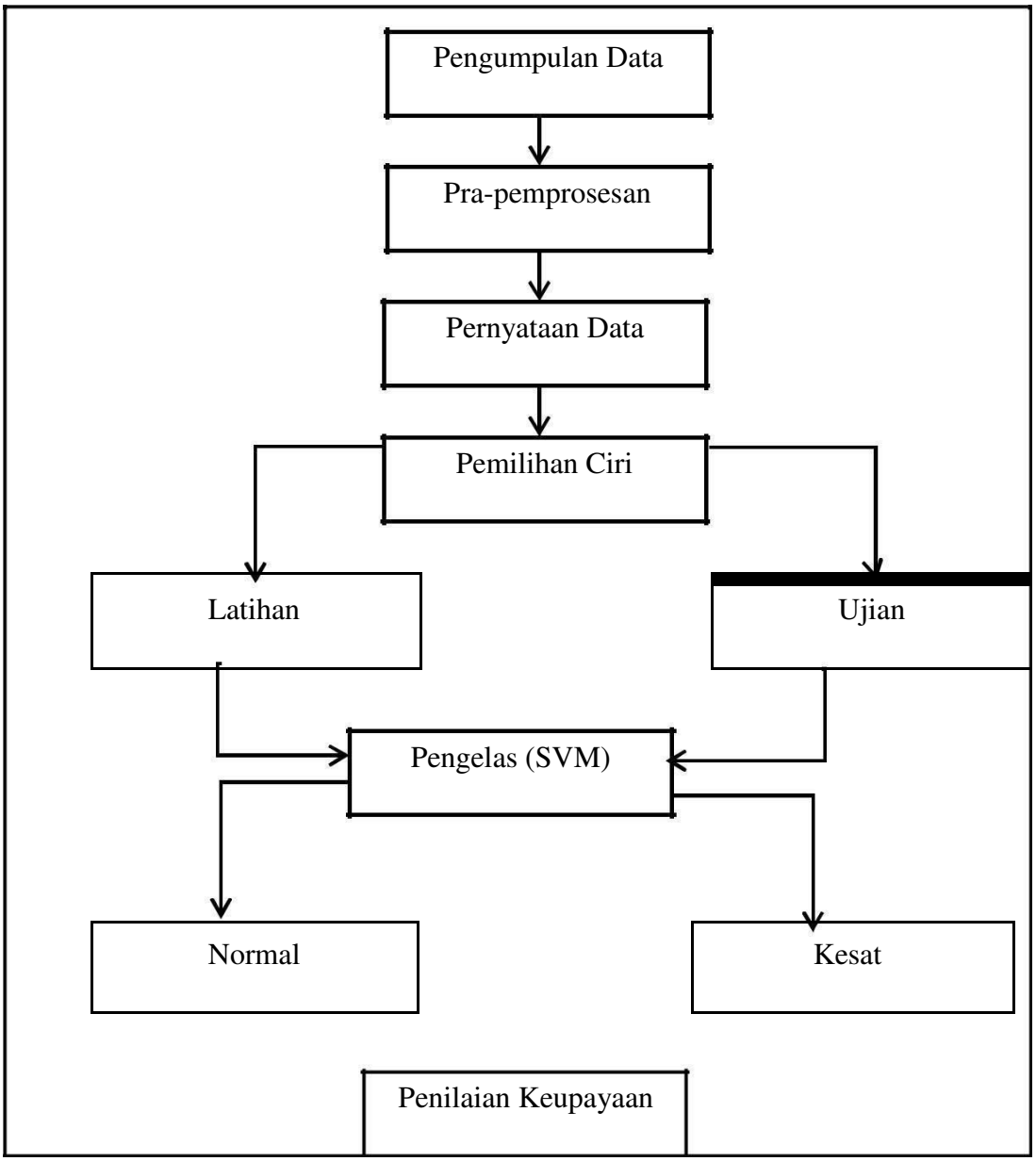
Secara umum, aliran kajian ini dapat dilihat dalam Rajah 1. Ia bermula dengan pengumpulan data di mana data-data yang diperoleh daripada laman web seperti laman web yang mengandungi penggunaan bahasa kesat yang melampau dan laman web yang berunsurkan pendidikan atau kesihatan.

Pra-pemprosesan pula adalah fasa di mana data yang dalam bentuk *Hypettext Markup Language* (HTML) ditukarkan menjadi bentuk teks. Penghuraian HTML berlaku bagi membuang semua sintaks yang terdapat dalam HTML. Dalam fasa ini juga, proses stemming dan stopping juga berlaku. Stemming adalah bertujuan untuk menjadikan perkataan itu sebagai kata akar manakala stopping pula adalah untuk membuang perkataan-perkataan yang sering digunakan seperti **"***is***",** **"***are***",** **"***and***"** dan sebagainya.

Pernyataan data adalah fasa yang di mana data yang telah melalui fasa sebelumnya ditunjukkan. Data tersebut kini dalam bentuk teks dan barulah fasa seterusnya boleh dilaksanakan. Bagi melaksanankannya, satu pakej yang dikenali sebagai *Voyant Server* digunakan bagi mengenal pasti antara perkataan atau istilah yang akan digunakan bagi fasa seterusnya.

Dalam fasa pemilihan ciri, skim pemberat istilah Kekerapan Istilah (TF) dan Kekerapan Istilah Songsang Kekerapan Dokumen (TFIDF) digunakan. Sebanyak 20 perkataan yang akan dipilih bagi melaksanakan fasa ini. Kekerapan perakataan ini akan dikira berdasarkan skim pemberat yang dipilih. Daripada 100 data yang dikumpulkan, data-data tersebut akan dibahagikan kepada dua iaitu 60 bagi data latihan dan 40 bagi data ujian. Kedua-dua data ini akan menjalani proses yang sama iaitu TF dan TFIDF. Perbezaannya adalah jumlah data sahaja.

Selepas melakukan pemilihan ciri iaitu TF dan TFIDF, data tersebut akan dikelaskan dengan menggunakan *Support Vector Machine* (SVM). Bagi melaksanakan SVM, perisian LibSVM akan digunakan. Daripada SVM, ini data tersebut dapat dikelaskan sama ada normal atau kesat. Fasa penilaian keupayaan pula adalah fasa bagi membandingkan keteapatan antara kedua-dua pemberat istilah yang digunakan iaitu TF dan TFIDF.



Rajah 1 Aliran Kajian

3.0 Keputusan

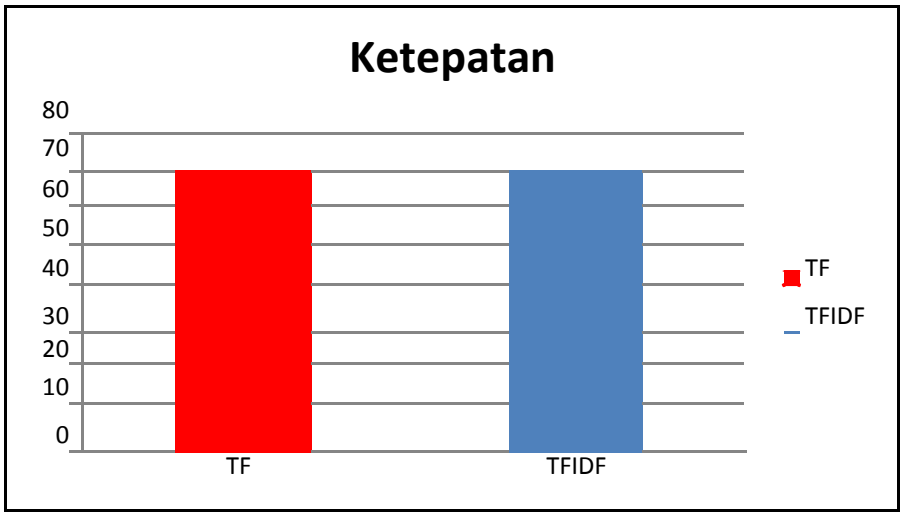
Label **"**1**"** adalah bermaksud kesat manakala label **"**0**"** bermaksud tidak kesat. Hasil daripada perisian LibSVM itu akan dibandingkan dengan data latihan. Keputusan bagi skim pembearat istilah TF adalah seperti dalam Jadual 1 manakala Jadual 2 menunjukkan keputusan bagi TFIDF. Rajah 2 pula menunjukkan perbandingan ketepatan antara kedua-dua skim pemberat yang digunakan.

Jadual 1 Keputusan Ketepatan bagi TF

Latihan	Keputusan	Ketepatan
0	0	T
1	0	F
1	0	F
1	0	F
0	0	T
0	0	T
1	0	F
0	0	T
0	0	T
0	0	T
0	0	T
1	0	F
0	0	T
0	0	T
0	0	T
0	0	T
0	0	T
1	0	F
0	0	T
0	0	T

Jadual 2 Keputusan Ketepatan bagi TFIDF

Latihan	Keputusan	Ketepatan
0	0	T
1	0	F
1	0	F
1	0	F
0	0	T
0	0	T
1	0	F
0	0	T
0	0	T
0	0	T
0	0	T
1	0	F
0	0	T
0	0	T
0	0	T
0	0	T
0	0	T
1	0	F
0	0	T
0	0	T



Rajah 2 Perbandingan Ketepatan

4.0 Perbincangan

Hasil daripada keptusan yang dibuat SVM, kedua-dua skim pemberat itu menunjukkan ketepatan sebanyak 70%. Hal ini berdasarkan daripada formula berikut.

$$\text{Ketepatan} = \frac{\quad}{\quad} \times 100\%$$

$$= 70\%$$

Berdasarkan keputusan yang diperoleh daripada TF dan TFIDF, sebanyak 14 data daripada jumlah keselurahan iaitu 20 memberi keputusan yang betul. Justeru, ketepatannya adalah sebanyak 70% .

5.0 Kesimpulan

Berdasarkan kajian yang dijalankan, hasil ketepatan yang diperoleh daripada kedua-dua skim pemberat istilah TF dan TFIDF menunjukkan ketepatan sebanyak 70%. Antara cadangan yang boleh ditambah bagi menambahkan lagi ketepatan keputusan ini adalah dengan menambah bilangan data.

Rujukan

Aghdam, M. H., Aghaee N.G., Basiri M.E, (2009) Text feature selection using ant colony optimization. Expert Systems with applications.

Hu, W., Wu, O., Fu, Z. & Maybank, S. (2007). Recognition of pornographic web pages by classifying texts and image. IEEE Transaction on Pattern Analysis and Machine Intelligence. pp-1019-1034

Joachims, T. (1998). Text categorization with support vector machines: Learning with many Relevant Features in Proceesings of ECML-98, 10th European Conference on Machine Learning

Lee, Z. S. (2008) Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classification. ISDA '08 proceeding of the 2008 Eighth International Conference on Intelligent Systems Design and Applications – Volume 02. Washington, DC, USA: IEEE, 145-150.

Lee, Z.S. (2010). Enhanced Featured Selection Method for Illicit Web Content Filtering. Doctor Philosophy, Universiti Teknologi Malaysia, Skudai.

Razavi, A., Ink Ipen, D., Uritsky, S., and Matwin, S., (2010). Offensive language detetction using multi-level classification. Proceedings of 23rd Canadian conference on Advances in Artificial Intelligence. Berlin, Heidelberg, Springer Verlag, 16-27.

Rohan S., Kalyani N., Shivani S., Shantanu N., Gopal U., (2015) A System to Detect Inappropriate Messages in Online Social Networks. 18th IRF International Conference. Pune, India

Selamat, A. and Omatu, S. (2004). Web page feature selection and classification using neural networks. Information Sciences. 158, 69-88.

Siti, F. (2012). Comparative Study On Term Weighting Schemes As Feature Selection Method For Malay Illicit Web Content Filtering. Master, Universiti Teknologi Malaysia, Skudai.

Wohnee, L., Samuel S.L., Seungjong C., & Dongun A., (2007). Harmful Contents Classification Using the Harmful Word Filtering and SVM, Chonbuk National University, South Korea

Yadav, S. H. and Parne, B. L. (2014). A Survey on Different Text Categorization Techniques for Text Filtration. International Journal of Computer Science and Technologies, Vol. 5(6), 8233-8235. Nagpur, India.

Xu, Z. and Zhu, S., (2010) Filtering Ofensive Language in Online Communities Using Grammatical Relations. Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference July 13-14, 2010, Washington, USA.