Malikussaleh Journal of Mechanical Science and Technology

ISSN : 2337-6945 Vol.5 No.2 (2021) 34-38

KLASIFIKASI INFORMASI KESEHATAN PADA DATA MEDIA SOSIAL MENGGUNAKAN SUPPORT VECTOR MACHINE DAN K-FOLD CROSS VALIDATION

Pauzi Ibrahim Nainggolan*, Desta Sandya Prasvita[2], Dhani Syahputra Bukit[3]
[1]Universitas Sumatera Utara, Medan, Indonesia
[2]Universitas Pembangunan Nasional Veteran Jakarta, Indonesia.
[3]Universitas Sumatera Utara, Medan, Indonesia
Corresponding Author: nainggolan@usu.ac.id

Abstract – *Media sosial saat ini memberikan informasi yang mampu mempengaruhi masyarakat. Sehingga, kini media sosial memiliki peranan signifikan sebagai sumber rujukan yang baru oleh maysarakat. Informasi kesehatan merupakan informasi yang sering dicari oleh pengguna media sosial. Penderita penyakit mencari informasi melalui media sosial terlebih dahulu sebelum bertemu dengan tenaga kesehatan. Tetapi kebanyakan informasi tidak dapat dipastikan sebagai informasi yang sesuai. Kesalahan terkait informasi kesehatan bisa membahayakan penderita. Ini bermakna, informasi yang terdapat pada media sosial perlu mendapatkan pengesahan pakar atau tenaga kesehatan. Penelitian ini betujuan untuk mengetahui penggunaan media sosial oleh tenaga kesehatan sebagai media konsultasi dan memberikan informasi yang tidak bertentangan dengan etika profesionalisme. Penelitian ini menggunakan teknik kalasifikasi Support Vector Machine (SVM). Validasi klasifikasi data yang diperoleh dilaksanakan menggunakan K-Fold Cross Validation. Hasil penelitian menunjukkan bahwa penggunaan SVM dalam klasifikasi kesesuaian informasi kesehatan dengan akurasi 70% pada data yang digunakan..*

Keywords: *Support Vector Machine, Informasi Kesehatan, Media Sosial*

1 Pendahuluan

Media sosial merupakan salah satu media informasi mengenai kesehatan. Perkembangan informasi kesehatan dalam media sosial saat ini semakin berkembang pesat. Melalui media sosial, dokter dapat berkomunikasi secara efektif dengan pesakit. Media sosial juga dapat digunakan oleh pasien untuk berbagi pengalaman bersama pasien yang lain. Hal ini mengakibatkan informasi kesehatan tidak lagi sesuai karena pemberi informasi tidak lagi kompeten atau bukan tenaga Kesehatan yang mengenali pasiennya malah terdapat kemungkinan informasi tidak sesuai untuk pasien tersebut.

Media sosial sebagai media penyebaran informasi juga mempunyai masalah terkait kebenaran informasi dan juga tahapan penggunaan informasi tersebut[1]. Penelitian ini dilakukan untuk menyelesaikan permasalahan dalam pengesahan kebenaran informasi melalui media sosial. Informasi dalam media sosial bisa disebarkan oleh siapa saja untuk tujuan tertentu. Pengesahan informasi adalah sangat penting terutama dalam bidang kesehatan. Informasi dan pengetahuan yang disebar tanpa pengesahan bisa membawa keadaan yang lebih buruk kepada penggunanya.

Informasi Kesehatan dalam media sosial yang interaktif, mudah alih, menarik, bersesuaian dengan kontekstual, dan dapat digunakan oleh masyarakat umum bisa meningkatkan kualitas penjagaan kesehatan dan promosi kesehatan [2]. Dalam media sosial, berbagai pihak bisa mendapatkan informasi yang sesuai dengan keperluan kesehatan mereka. Informasi kesehatan mudah dikemas kini namun sulit disesuaikan dengan masalah berkaiatan kesehatan yang sering berubah-ubah. Media sosial boleh menggalakkan penyertaan yang lebih besar di antara penyedia informasi kesehatan yang boleh berhubung secara langsung dengan pengguna. Keperluan menjaga hubungan di antara pakar perubatan dengan pesakit dalam media sosial adalah penting. Namun, masih terdapat pakar perubatan yang sulit untuk memanfaatkan penggunaan media sosial. Oleh itu, kajian

Manuscript received October 2021, revised October 2021

secara mendalam perlu dilakukan untuk mengetahui kebenaran informasi yang tersebar dalam media sosial.

2 Kajian Pustaka

Kajian literatur yang dijalankan dibahagi kepada tiga tahapan berasaskan keperluan penelitian. Pertama, adalah pengumpulan data dan praproses data. Tahapan kedua klasifikasi data dengan menggunakan metode SVM. Tahapan ketiga ialah evaluasi kinerja klasifikasi untuk memperoleh hasil akurasi yang maksimal.

2.1 Praproses Data

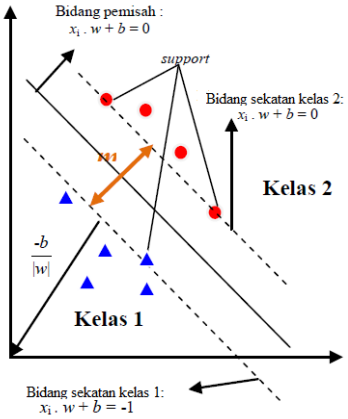
Praproses data dilakukan untuk meningkatkan nilai evaluasi klasifikasi teks. Dalam tahapan praproses dilakukan penghilangan stopword dan stemming. Penghilangan stopword ialah untuk menghilangkan kata kata umum seperti kata "The", "did", "do" dan sebagainya yang tidak memiliki nilai informasi.

Stemming adalah cara mengubah kata dalam sebuah dokumen teks ke kata dasarnya. Sebagai contoh kata connected, connecting, connection dan connections memiliki kata dasar yang sama iaitu connect.

2.2 Support Vector Machine(SVM)

SVM adalah metode klasifikasi yang mencari hyperplane dari maksimum margin pemisah antara dua kelas data. SVM menggunakan ruang hipotesis dari suatu fungsi linear dalam suatu ruang dimensi berfitur tinggi [3]. Pendekatan struktural risk minimization (SRM), digunakan untuk membangun sebuah pemisah hyperplane optimum dengan ketepatan pengelasan yang tinggi.

Sebagai contoh, jika data dinotasikan sebagai $x_i \in \Re^n$, untuk label kelas dari data x_i dinotasikan $y \in \{+1,-1\}$ dengan i = 1,2,…,l dengan l adalah banyak data. Pemisahan data secara linear pada kaedah SVM dapat dilihat pada Gambar 1.



Gambar 1 Contoh data dengan hyperplane

Margin adalah jarak antara hyperplane dan pola data terdekat dari kelas masing-masing. Pola data yang paling dekat disebut vektor sokongan. Nilai margin antara dua kelas adalah $m=\frac{2}{||w||}$, dengan *w* adalah nilai vektor yang tegak lurus terhadap hyperplane atau bidang pemisah. Margin dapat dimaksimumkan menggunakan fungsi pengoptimuman *Lagrangian* berikut:

$$\min_{w,b} L(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{1} \alpha_i \, (y_i((x.w+b) - 1 \quad (1)$$

Dengan meminimumkan *L* terhadap *w* dan *b,* diperoleh:

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{1} \alpha_i y_i w_i = 0$$

$$\frac{\partial L}{\partial w} = \sum_{i=1}^{1} \alpha_i y_i = 0$$

Persamaan (1) dapat dimodifikasi untuk memaksimumkan *L* yang mengandung α_i sebagai persamaan (2).

$$\max_{\alpha} L = \sum_{i=1}^{1} \alpha_i - \frac{1}{2} \sum_{i=1}^{1} \sum_{j=1}^{1} \alpha_i \alpha_j y_i y_j x_i x_j \qquad (2)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i > 0$$

Nilai α yang dihasilkan digunakan untuk mencari *w*. Data yang memiliki nilai $\alpha_i \geq 0$ merupakan vektor sokongan, manakala selebihnya mempunyai nilai $\alpha_i = 0$. Setelah nilai α_i ditemukan, kelas dari data pengujian *x* dapat ditentukan berdasarkan nilai fungsi keputusan:

$$f(x_d) = \sum_{i=1}^{NS} \alpha_i y_i x_i . x_d + b$$

Dengan:

x_i = vektor sokongan;

NS = jumlah support vector;

x_d = data yang akan dikelaskan

2.3 K-Fold Cross Validation

Klasifikasi silang K-Fold adalah metode yang digunakan untuk membagi data menjadi data latih dan data uji. Klasifikasi silang K-Fold membagi data contoh secara acak ke dalam K subset yang saling bebas. Satu subset digunakan sebagai data uji dan K-1 subset sebagai data latih. Proses klasifikasi silang akan diulang hingga K kali. Data awal dibagi menjadi K subset yang saling bebas secara acak yaitu, S1, S2,…, Sk, dengan ukuran setiap subset sama. Pelatihan dan pengujian dilakukan sebanyak K kali. Pada proses ke-i, subset Si diperlakukan sebagai data uji dan subset lainnya diperlakukan sebagai data latih. Pada proses pertama, S2,…, Sk menjadi data latih dan S1 menjadi data uji, pada proses kedua S1, S3,…,

Sk menjadi data latih dan S2 menjadi data uji, dan seterusnya [4].

3 Metodologi Penelitian

3.1 Data Penelitian

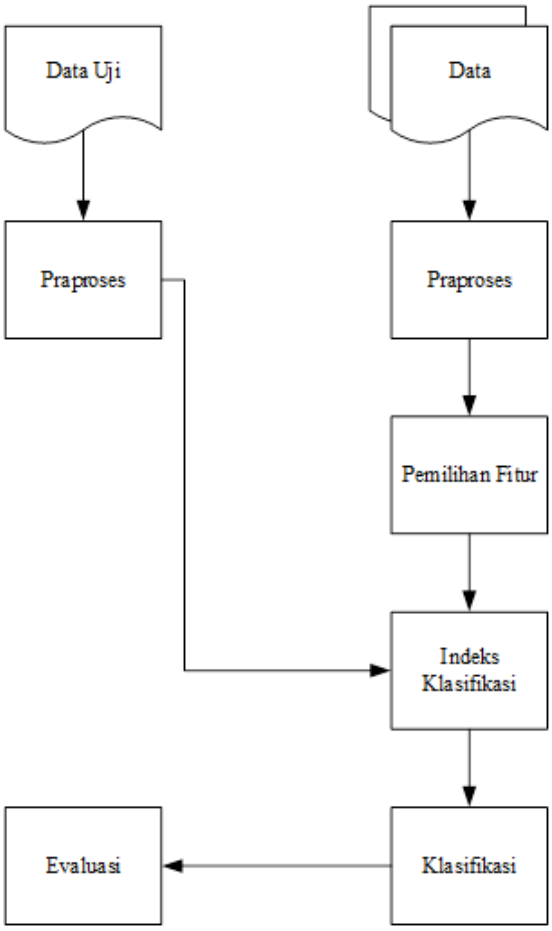
Data dikumpulkan berdasarkan laporan kesehatan yang diambil daripada University of California, School of Information and Computer Sciences, Knowledge Discovery in Databases Archive. Jumlah data secara keseluruhan adalah berjumlah 1000 dokumen. Dokumen yang dikumpulkan dalam format .txt. Kemudian data tersebut diklasifikasikan dalam kelas data untuk pelajar bidang kesehatan.

Daripada jumlah data yang diperoleh 50 di antaranya dipilih untuk dijadikan data ujian kepada sistem pengelasan yang dijalankan. Data yang diujikan telah diketahui kelas sebenarnya. Pengujian yang dijalankan bertujuan untuk mengetahui kecekapan sistem pengelasan yang dijalankan. Ujian selanjutnya dijalankan ke atas laporan kesehatan/ perbincangan kesehatan yang dilakukan oleh pakar kesehatan dalam sosial media Facebook

3.2 Deskripsi Sistem

Pelaksanaan penelitian diawali dengan pengumpulan data. Data yang didapatkan dibedakan menjadi dua yaitu, data latih dan data uji. Setiap data dilakukan praproses terlebih dahulu untuk mendapatkan kata-kata yang sesuai maknanya dengan penelitian yang dilakukan. Selanjutnya data latih yang telah melewati tahapan praproses masuk kepada tahapan pemilihan fitur.

Tahapan pemilihan fitur untuk mengetahui batasan fitur bagi setiap kelas. Sistem klasifikasi dijalankan di luar talian untuk menghasilkan fitur setiap kelas. Fitur setiap kelas dimasukkan kepada indeks klasifikasi untuk menghasilkan hasil klasifikasi yang akan di uji. Secara keseluruhan alur proses di ditunjukkan pada Gambar 2.



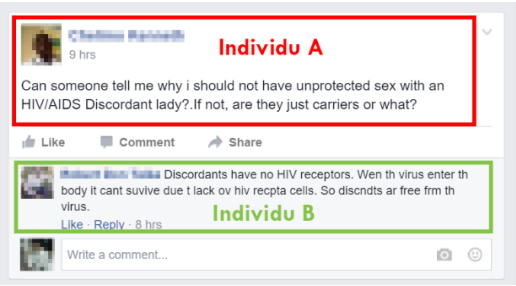
Gambar 2. Deskripsi Sistem

4 Hasil dan Pembahasan

Penelitian menghasilkan klasifikasi data dan hasil pengujian sistem. Pengujian dilakukan dari setiap tahapan mulai dari tahapan praproses data, tahapan klasifikasi yang menggunakan metode SVM dan tahapan pengujain hasil klasifikasi dengan K-Fold cross validation.

4.1 Data dan Praproses Data Pengujian

Data pengujian diambil dari media social yang mengandung informasi kesehatan. Data diekstrak dari postingan atau komentar postingan yang mengandung informasi atau pertanyaan mengenai Kesehatan. Pada Gambar 3 menunjukkan Individu A yang melakukan posting mendapat informasi balasan dari individu B yang memberikan pendapat.



Gambar 3 Contoh data pengujian

Setiap posting pengguna diuji dengan model klasifikasi untuk mengetahui posting tersebut berkaitan dengan bidang kesehatan ataupun tidak. Data yang terdiri daripada 1000 dokumen dipisahkan oleh pakar kesehatan, yang mengesahkan posting tersebut berkaitan dengan kesehatan atau tidak. Hasil kajian penentuan yang dijalankan menunjukkan 715 posting adalah berkaitan dan mengandungi informasi kesehatan manakala posting yang tidak berkaitan dengan informasi kesehatan 285 dokumen.

4.1.1 Penghapusan stop Word

Stop Word atau kata henti adalah kata dalam data yang tidak memiliki arti. Kata henti tidak dapat mencirikan sesuatu dokumen pada kelasnya. Contoh kata henti THE, I, SO, THEN dan lainnya. Pengenalan kata henti pada aplikasi pemprofilan dihapuskan terlebih dahulu sebelum dokumen diproses.

Tindakan penghapusan kata henti dalam dokumen yang akan dikelaskan dapat meningkatkan ketepatan pengelasan satu dokumen [5]. Praproses dalam penelitian ini menggunakan 780 kata. Perkataan yang sudah diidentifikasi sebagai kata henti atau stop word dihapuskan dari dokumen. Dokumen yang akan diuji juga mengalami praproses penghapusan kata henti

4.1.2 Stemming

Setelah praproses penghapusan kata henti selanjutnya data diproses Kembali menggunakan algoritma Porter Stemming untuk mendapatkan kata yang sesuai [6]. Algoritma porter stemming memiliki 60 aturan dan 6 tahpan tanpa perulangan. Tahapan dalam algoritma porter stemming adalah seperti berikut:

1. Menghilangkan kata jamak dan kata yang memiliki akhiran -ed atau –ing

Contoh: Interesting → interest; agreed→ agree;

2. Mengubah huruf pada kata yang berakhiran y kepada i yang apabila terdapat kata vokal lain dalam perkataan tersebut.

Contoh: Interestingly → Interestingli; Happy → Happi; Grey→ grei;

3. Memetakan akhiran ganda kepada kata tunggal, Seperti: -ization, -ational, dan lain-lain

Contoh: Operational→operate; vietnamization→ vietnamize;

4. Mengembalikan akhiran kepada kata dasar, seperti akhiran: Ful, -ness dan lain-lain.

Contoh: goodness→good; Playful→play;

5. Menanggalkan akhiran, sepert: -ant, -ence, dan lain-lain

Contoh: inference→ infer; irritant→irrit;

6. Membuang yang memiliki akhiran –e

Contoh: Controllable→controll→control; Deflate→deflat; Parable→parable;

Tujuan mengembalikan kata kepada kata dasar adalah untuk menghilangkan kekaburan perhitungan kata tersebut. Kemunculan kata yang diberi akhiran dengan kata dasar menjadi satu perwakilan dalam setiap penilaian klasifikasi dokumen. Sebuah dokumen yang diwakili oleh sebuah vektor memiliki nilai yang sama untuk kata yang berakhiran dan yang tidak memiliki akhiran. Kemunculan kata dalam satu vektor dihitung untuk menjelaskan kelas dokumen tersebut.

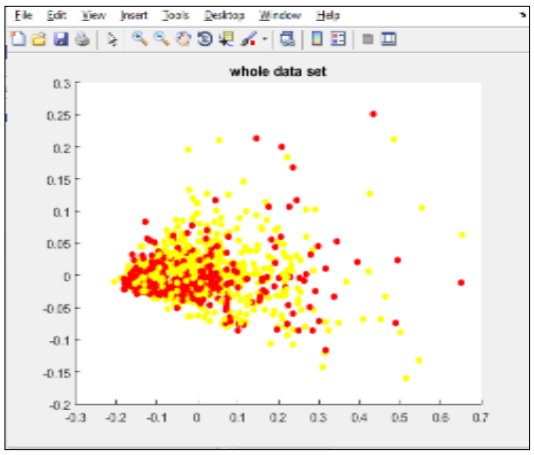
4.2 Pengujian Klasifikasi

Klasifikasi pada Penelitian yang memfokuskan di bidang kesehatan ini terbagi kepada dua tahapan klasifikasi. Tahapan pertama ialah membedakan kelas data kesehatan dengan yang bukan bidang kesehatan. Tahapan kedua fokus kepada bidang penelitian kepakaran dalam kesehatan. Kedua tahpan klasifikasi dijalankan terhadap data latih untuk mendapatkan indeks data, dan data uji bagi mendapatkan kelas data dari pengguna.

Tidak semua bidang kesehatan dimasukkan dalam penelitian ini. Kelengkapan data yang lebih banyak tentu menjadi hal utama untuk mendapatkan indeks data dari setiap kelas kepakaran dalam bidang kesehatan. Dalam penelitian ini, setiap kelas kepakaran yang tidak memiliki indeks data, dikategorikan dalam kelas memiliki kepakaran secara umum.

Penelitian ini menggunakan teks dalam Bahasa Inggris. Data yang menggunakan bahasa lain tidak disertakan. Pengelasan SVM dijalankan dengan fungsi kernel radial basis dengan nilai gamma (γ) sebanyak 0.7. Penelitian ini menggunakan 1000 dokumen yang diklasifikasikan kepada kelas bidang kesehatan dan kelas bukan bidang kesehatan. Pembagian kedua kelas merujuk kepada makna dari dokumen yang diklasifikasikan. Setiap kata yang menjelaskan makna dokumen tidak seluruhnya memiliki fitur atau ciri yang memberikan makna khusus dalam dokumen tersebut.

Tahapan pertama klasifikasi menghasilkan 715 dokumen sebagai kelas kesehatan. Hasil proses klasifikasi tersebut, terdapat 698 dari 715 dokumen memiliki ciri dalam bidang kesehatan. Bagi kelas yang bukan bidang kesehatan, setelah dilakukan praproses terdapat hanya 276 dari 285 dokumen yang memiliki ciri bukan bidang kesehatan. Penyebaran ciri dokumen ditunjukkan pada Gambar 4. Titik kuning pada Gambar 4 meunjukkan penyebaran data kelas kesehatan dan titik merah menunjukkan data bukan kelas Kesehatan.



Gambar 2 Sebaran Ciri Dokumen

Tahapan kedua klasifikasi pada data latih terdapat di 46 kelas bidang kesehatan. Data didapati memiliki ciri dari jumlah data bagi untuk setiap bidang kepakaran kesehatan. Rincian data latih pada bidang kepakaran ditunjukkan melalui Tabel 1

Tabel 1 Sebaran Ciri Dokumen dalam bidang kesehatan

No	Bidang Pakar Kesehatan	Data	No	Bidang Pakar Kesehatan	Data
1	Accident and emergency medicine	**4**	24	Neurosurgery	**5**
2	Allergology	**58**	25	Obstetrics and gynecology	**14**
3	anaesthetics	**2**	26	Oncology	**1**
4	biological hematology	**1**	27	Opthalmology	**21**
5	Cardiology	**15**	28	Orthopaedics	**22**
6	Child psychiatry	**4**	29	Otorhinolaryngology	**10**
7	Clinical biology	**11**	30	Paediatrics	**10**
8	Clinical chemistry	**42**	31	Pathology	**20**
9	Clinical neurophysiology	**8**	32	Pharmacology	**101**
10	Dental	**2**	33	Physical medicine and rehabilitation	**24**
11	Dermatology	**22**	34	Plastic surgery	**1**
12	Dermato-venerology	**1**	35	Podiatrics surgery	**2**
13	Endocrinology	**15**	36	Psychiatry	**6**
14	farmakology	**5**	37	Public health / Preventive Medicine	**47**
15	Gastroenterology	**9**	38	Radiology	**20**
16	General hematology	**5**	39	rehabilitation	**1**
17	General surgery	**7**	40	Respiratory medicine	**11**
18	Immunology	**6**	41	Rheumatology	**4**
19	Infectious diseases	**30**	42	Stomatology	**6**
20	Internal medicine	**17**	43	Tropical medicine	**2**
21	Laboratory medicine	**11**	44	Urology	**37**
22	Microbiology	**34**	45	Venereology	**15**
23	Neurology	**26**			

4.3 Evaluasi Sistem

Evaluasi system dilakukan melalui hasil pengujian data yang dilakukan. Pengujian menggunakan 30 data dokumen yang bukan bidang kesehatan dan 70 data bidang kesehatan. Hasil Pengujian kelasifikasi menunjukkan ketepatan kelasifikasi keseluruhan dokumen yang diuji sebesar 70%. Keseluruhan data dokumen yang diuji berjumlah 100, seluruh dokumen memiliki ciri tertentu dan telah melalui praproses.

5 Kesimpulan

Berdasarakan perancangan penelitian yang telah selesaikan dilaksanakan. Proses klasifikasi menghasilkan nilai evaluasi yang cukuo baik. Pengujian yang dilakukan mendapati bahwa penelitian mengalami underfitting, karena data informasi Kesehatan yang digunakan tidak dapat mengetahui kesamaan data.

Selanjutnya pengembangan penelitian dapat memperkaya data informasi Kesehatan. Peningkatan data dapat merepresentasikan evaluasi yang baik. Penggunaan multi kelas SVM bisa digunakan untuk mencoba peningkatan nilai akurasi pengujian.

Daftar Pustaka

[1] Scanfeld, D., Scanfeld, V., & Larson, E. L. 2010. Dissemination of health information through social networks: twitter and antibiotics. American Journal of Infection Control, 38(3).

[2] Kreps, G. L., & Neuhauser, L. 2010. New directions in eHealth communication: opportunities and challenges. Patient Education and Counseling, 78(3), 329–36.

[3] Cortes C, Vapnik V. 1995. Support-vector networks. Machine Learning 20: 273-297.

[4] Fu, L.M. 1994. Neural Network In Computer Intelligence. Singapura: McGraw Hill.

[5] Moh, T.-S., & Bhagvat, S. (2012). Clustering of technology tweets and the impact of stop words on clusters. Proceedings of the 50th Annual Southeast Regional Conference, 226–231.

[6] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

[7] Ilham, M., Islami, N., Abdurrahman, F., & Suryadi, S. (2021). E-aedes framework based on Geographical Information System: Stakeholder Perceptions. Journal of Multidisciplinary Academic, 4(6), 453-456.