JURNAL PINTER
VOL 3. NO.2 DESEMBER 2019
e-ISSN: 2597- 4475
http://doi.org/10.21009/pinter.3.2.5

Kinerja Algoritma Support Vector Machine dalam Menentukan Kebenaran Informasi Banjir di Twitter

Muhamad Prasetio Dwi Cahyo[1], Widodo[2], Bambang P. Adhi[3]

[1,2,3] Pendidikan Teknik Informatika dan Komputer Fakultas Teknik
Universitas Negeri Jakarta

Email: mprasetiodc@gmail.com , [2] widodo@unj.ac.id, [3]bambangpadhi@unj.ac.id

ABSTRAK

Menurut survei yang dilakukan oleh Asosiasi Penyelenggara Jaringan Internet Indonesia (APJII) mengatakan bahwa sepanjang 2016 sebanyak 132,7 juta orang telah terhubung ke internet. Dalam penggunaan internet seseorang dapat berkomunikasi melalui jejaring sosial. Jejaring sosial adalah sarana untuk bersosialisasi satu sama lain secara online didunia maya (internet). Twitter adalah salah satu dari macam-macam jejaring sosial. Menurut catatan Badan Penanggulangan Bencana Daerah (BPBD) Jakarta terdapat 700 kasus banjir selama periode Januari-Agustus 2016. Pada twitter terdapat user yang memberikan informasi tentang banjir dengan mengirimkan sebuah *tweet*. Namun penggunaan kata "banjir" tidak semua dimaksudkan untuk memberi informasi mengenai banjir. Ada yang menggunakannya hanya sebagai kata kiasan. Penelitian ini menggunakan algoritma *Support Vector Machine* untuk mengklasifikasi *tweet* yang benar memberikan informasi mengenai banjir atau tidak. Algoritma *Support Vector Machine* adalah suatu algoritma yang memiliki tingkat akurasi yang tinggi. Dalam penelitian ini untuk mengevaluasi tingkat akurasi dari algoritma *Support Vector Machine* menggunakan *Confusion Matrix*. Hasilnya adalah tingkat akurasi dari *Support Vector Machine* dalam menentukan tweet mengenai informasi banjir sebesar 0,96.

Kata Kunci : Twitter, Banjir, Informasi, *Support Vector Machine, Confusion Matrix.*

1. PENDAHULUAN

Teknologi di era globalisasi saat ini sudah berkembang, terutama dalam penggunaan internet. Penggunaan internet di berbagai elemen masyarakat sudah berpengaruh besar terhadap kehidupan sehari-hari. Menurut survei yang dilakukan oleh Asosiasi Penyelenggara Jaringan Internet Indonesia (APJII) yang dimuat oleh www.tekno.kompas.com mengungkap bahwa lebih dari setengah penduduk Indonesia kini telah terhubung ke internet. Survei yang dilakukan sepanjang 2016 itu menemukan bahwa 132,7 juta orang Indonesia telah terhubung ke internet. Adapun total penduduk Indonesia sendiri sebanyak 256,2 juta orang. Dengan menggunakan internet saat ini dapat memudahkan seseorang dalam berkomunikasi satu sama lain, salah satunya melalui jejaring sosial.

Jejaring sosial adalah sarana untuk bersosialisasi satu sama lain secara online didunia maya (internet). Twitter adalah salah satu dari macam-macam jejaring sosial. Twitter adalah layanan jejaring sosial yang memungkinkan penggunanya untuk mengirim pesan, yang dikenal dengan sebutan kicauan (*tweet*). Dengan jejaring sosial Twitter penggunanya dapat dengan mudah berkomunikasi atau memberi informasi baru terkait keadaan lingkungan sekitarnya. Informasi yang sering kita peroleh dari jejaring sosial twitter mengenai fenomena alam seperti banjir.

Seperti yang diketahui fenomena banjir sering terjadi di daerah DKI Jakarta. Menurut catatan Badan Penanggulangan Bencana Daerah (BPBD) Jakarta terdapat 700 kasus banjir selama periode Januari-Agustus 2016. Banyak pengguna twitter memberikan

Avalaiable at:
http://journal.unj.ac.id/unj/index.php/pinter/article/view/16384

informasi mengenai banjir di daerah lingkungannya dengan cara membuat *tweet*.

Namun tidak semua *tweet* yang berisi kata "banjir" adalah informasi mengenai lokasi yang sedang terkena banjir. Sebagian ada pengguna jejaring sosial twitter lainnya yang menggunakan kata "banjir" tetapi tidak memberikan informasi mengenai banjir namun penggunaanya hanya sebagai kiasan.

Atas dasar tersebut, maka peneliti merasa perlu untuk mengklasifikasi *tweet* yang berisi informasi tentang banjir yang akurat, agar memudahkan masyarakat untuk dapat informasi yang mereka inginkan. Dalam mengklasifikasi *text* terdapat beberapa metode algortima, seperti *Support Vector Machine, Naive Bayes, Decision Tree,* etc. Untuk mengklasifikasi *tweet* yang berisikan informasi mengenai banjir, peneliti menggunakan algortima *Support Vector Machine*. *Support Vector Machine* memiliki kelebihan yaitu mampu mengidentifikasikan hyperplane terpisah yang memaksimalkan margin antara dua kelas yang berbeda (Chou et al., 2014). Pertimbangan peneliti menggunakan algoritma *Support Vector Machine* karena algortima tersebut memiliki tingkat akurasi yang tinggi seperti dalam penelitian yang dilakukan oleh Achamd Nurhadi (2015) yang menggunakan algortima tersebut dalam mengklasifikasi konten berita digital bahasa Indonesia, menghasilkan nilai akurasi sebesar 95.42%. Jadi penelitian ini difokuskan untuk implementasi algoritma *Support Vector Machine* dalam pengklasifikasi *tweet* mengenai banjir.

2. DASAR TEORI

2.1. Kajian Teoritik

2.1.1. Banjir

Banjir adalah fenomena alam yang terjadi di kawasan yang banyak dialiri oleh aliran sungai. Sedangkan secara sederhana, banjir didefinisikan sebagai hadirnya air suatu kawasan luas sehingga menutupi permukaan bumi kawasan tersebut. Berdasarkan SK SNI M-18-1989-F (1989) dalam Suparta 2004, bahwa banjir adalah aliran air yang relatif tinggi, dan tidak tertampung oleh alur sungai atau saluran.

2.1.2. Text Mining

Menurut Han & Kamber (2006) yang diacu dalam Kestrilia Rega Prilianti & Hendra Wijaya (2014: 1) *text mining* adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen.

2.1.3. Support Vector Machine

Menurut (Nello Christianini dan John S. Taylor, 2000) yang diacu dalam Dwi Astuti (2007:2) *Support Vector Machine* (SVM) adalah sistem pembelajaran yang pengklasifikasiannya menggunakan ruang hipotesis berupa fungsi-fungsi linear dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik.

Langkah awal suatu algoritma SVM adalah pendefinisian persamaan suatu hyperplane pemisah yang dituliskan dengan:

$$W . X + b = 0 \quad (1)$$

W adalah suatu bobot vektor, yaitu W = {W1, W2,…,Wn}; n adalah jumlah atribut dan b merupakan suatu skalar yang disebut dengan bias. Jika berdasarkan pada atribut A1, A2 dengan permisalan *tupel* pelatihan X = (x1, x2), x1 dan x2 merupakan nilai dari atribut A1 dan A2, dan jika b dianggap sebagai suatu bobot tambahan w0, maka persamaan suatu *hyperplane* pemisah dapat ditulis ulang sebagai berikut:

$$w_0 + w_1 w_1 + w_2 w_2 = 0 \quad (2)$$

Setelah persamaan dapat didefinisikan, nilai x1 dan x2 dapat dimasukkan ke dalam persamaan untuk mencari bobot w1, w2, dan w0 atau b.

SVM menemukan *hyperplane* pemisah maksimum, yaitu *hyperplane* yang mempunyai jarak maksimum antara tupel pelatihan terdekat. *Support vector* ditunjukkan dengan batasan tebal pada titik tupel. Dengan demikian, setiap titik yang terletak di atas *hyperplane* pemisah memenuhi rumus:

$$w_0 + w_1 w_1 + w_2 w_2 > 0 \quad (3)$$

Sedangkan, titik yang terletak di bawah *hyperplane* pemisah memenuhi rumus:

$$w_0 + w_1 w_1 + w_2 w_2 < 0 \quad (4)$$

Melihat dua kondisi di atas, maka didapatkan dua persamaan *hyperplane* yaitu:

$$H_1 : w_0 + w_1 w_1 + w_2 w_2 \geq 1 \text{ untuk } y_i = +1 \quad (5)$$

$$H_2 : w_0 + w_1 w_1 + w_2 w_2 \leq -1$$

untuk $y_i = -1$ (6)

Perumusan model SVM menggunakan trik matematika yaitu formula Lagrangian. Berdasarkan Lagrangian *formulation*, *Maksimum Margin Hyperplane* (MMH) dapat ditulis ulang sebagai suatu batas keputusan (*decision boundary*) yaitu:

$$d(X^T) = \sum_{i=1}^{l} y_i a_i X_i X^T + b_0 \quad (7)$$

yi adalah label kelas dari support vector Xi. XT merupakan suatu tupel test. αi dan b0 adalah parameter numerik yang ditentukan secara otomatis oleh optimalisasi algoritma SVM dan l adalah jumlah *vector support*.

2.1.4. Informasi

Menurut (Tata Sutabri 2005:23) diacu dalam Slamet Pebrianto (2010:2) Informasi adalah data yang telah diklasifikasikan atau diolah atau diinterprestasi untuk digunakan dalam proses pengambilan keputusan.

2.1.5. Twitter

Menurut Domikus Juju & MataMaya Studio (2009: 2-3) di acu dalam Dika Putri Utama, dkk (2015: 5) definisi dari twitter adalah sebuah web dan layanan mikroblog yang bisa digunakan untuk melakukan pembaharuan (update) berupa sebuah teks panjang maksimum sebanyak 140 karakter, pembaharuan (update) di twitter dikenal sebagai tweets.

2.1.6. Confusion Matrix

Metode ini menggunakan tabel matriks seperti pada Tabel 1 jika data set hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007). Confusion Matrix adalah tools yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah.

Tabel 2.1. Model *Confusion Matrix* (Bramer, 2007)

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	TP *true positives*	FN *false negatives*
-	FP *false positives*	TN *true negatives*

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah record negatif yang diklasifikasikan sebagai *negative*. Evaluasi dan validasi hasil dihitung menggunakan rumus akurasi, *precision*, *recall* dan *f-measure* berikut ini:

- Akurasi

Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang diklasifikasi secara benar dengan total sample data testing yang diuji.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision*

Menghitung nilai *precision* dengan cara membagi jumlah data benar yang bernilai positif (*True Positive*) dibagi dengan jumlah data benar yang bernilai positif (*True Positive*) dan data salah yang bernilai positif (*False Negative*).

$$Precision = \frac{TP}{TP + FP}$$

- *Recall*

Sedangkan *recall* dihitung dengan cara membagi data benar yang bernilai positif (*True Positive*) dengan hasil penjumlahan dari data benar yang bernilai positif (*True Positive*) dan data salah yang bernilai negatif (*False Negative*).

$$Recall = \frac{TP}{TP + FN}$$

- *F-Measure*

Nilai *F-Measure* didapat dari perhitungan pembagian hasil dari perkalian precision dan recall dengan hasil penjumlahan *precision* dan *recall*, kemudian dikalikan dua.

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

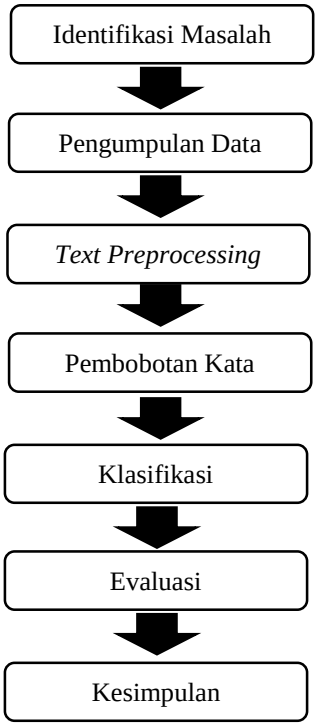
2.2. Prosedur

Berdasarkan identifikasi masalah yang telah dijelaskan pada bab 1, langkah pertama adalah mengumpulkan data *tweet* dari twitter yang akan digunakan sebagai data train dan data test.

Setelah data terkumpul, proses selanjutnya adalah *text preprocessing* pada setiap *tweet* meliputi proses *case fold*, *tokenizing*, *filtering*. Selanjutnya *tweet* dibuat menjadi sebuah *Vector Space Model* dengan pendekatan *bag of word* dan setiap kata pada *bag of word* akan di beri bobot dengan metode TF-IDF.

Setelah melewati tahap pembobotan kata, kemudian dilakukan teknik klasifikasi dengan menggunakan *Support Vector Machine*. Hasil dari klasifikasi tersebut lalu

dievaluasi terlebih dahulu. Tahap terakhir penelitian penulis akan menarik kesimpulan berdasarkan evaluasi hasil dari klasifikasi pada tahap sebelumnya.



Gambar 2.1. Prosedur Penelitian

3. Metodologi Penelitian

3.1. Tempat dan Waktu Penelitian

Penelitian ini dilakukan di Program Studi Pendidikan Teknik Informatika dan Komputer Fakultas Teknik Universitas Negeri Jakarta. Penelitian ini dilaksanakan sejak bulan Oktober 2016 hingga Januari 2017.

3.2. Alat dan Bahan Penelitian

Hardware

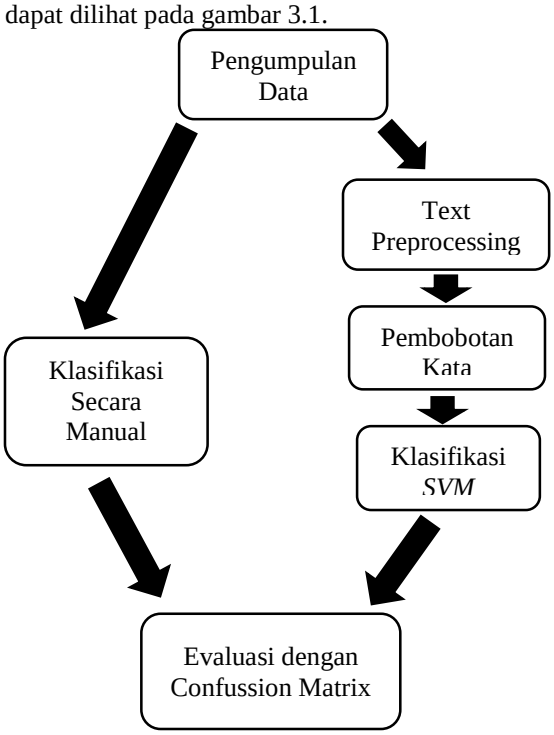
1. Processor AMD E2-1800, 1.7GHz
2. Memory RAM 4 GB DDR 3
3. Harddisk 500 GB

Software

1. Sistem Operasi Windows 7 Ultimate 64-bit Operating System
2. Microsoft Excel 2016 64-bit
3. Notepad
4. Anaconda 3 64-bit
5. PyCharm Community Edition 2016.2.3
6. Menggunakan modul dan fungsi yang ada pada www.scikit-learn.org.

3.3. Diagram alir penelitian

Langkah-langkah pada penelitian ini dapat dilihat pada gambar 3.1.



Gambar 3.1 Diagram Alir Penelitian

3.4. Teknik dan Prosedur Pengumpulan data

Data yang dibutuhkan adalah *tweet-tweet* yang diambil dari website Twitter berbahasa Indonesia. Data yang diambil berupa *tweet* sebanyak 400 *tweet*. Proses pengumpulan data dilakukan secara manual.

3.4.1. Klasifikasi Secara Manual

Pada proses 1 dari data yang telah terkumpul akan dilakukan pengklasifikasian untuk menentukan *tweet* mana yang akurat dan tidak akurat. Maksud dari akurat disini adalah *tweet* yang mengandung kata "banjir" yang benar memberikan informasi mengenai banjir, bukan *tweet* yang mengandung kata "banjir" tetapi tidak ada hubungannya dengan peristiwa banjir (penggunaan kata "banjir" sebagai kata khiasan).

4. Hasil Penelitian

4.1. Deskripsi Hasil Penelitian

Penelitian ini terdiri dari pengumpulan data, melalui tahap *case folding*, tahap *tokenizing*, tahap *filtering*, pembobotan kata lalu mengklasifikasi data dengan algoritma *Support Vector Machine*. Data sebanyak 400 *tweet* akan diuji di dalam program implementasi algoritma *Support Vector Machine* untuk menentukan apakah *tweet*

tersebut termasuk "*false*" atau "*true*". Yang dimaksud dengan "*false*" adalah apabila *tweet* tersebut menggunakan kata "banjir" tetapi tidak berkaitan dengan bencana alam berupa banjir. Sedangkan yang dimaksud dengan "*true*" adalah apabila *tweet* tersebut menggunakan kata "banjir" yang berisikan informasi mengenai suatu daerah yang sedang terjadi bencana alam berupa banjir.

4.2. Analisis Data Penelitian

Setelah menghitung nilai IDF dan diberi label, selanjutnya masuk ke tahap klasifikasi data. Kumpulan data *tweet* tersebut akan di klasifikasi menggunakan algoritma *Support Vector Machine.* Pada pengujian algoritma *Support Vector Machine* di penelitian ini menggunakan metode *K-fold Cross Validation.* Sehingga pengklasifikasian akan dilakukan sebanyak 10 kali (dengan presentase *data test* sebanyak 10% dari data keseluruhan). Dengan menggunakan data test yang berbeda-beda diharapkan hasil dari klasifikasi akan semakin valid.

Instrumen yang digunakan oleh peneliti untuk mengaplikasikan algoritma *Support Vector Machine* dalam bahasa pemrograman python adalah PyCharm Community Edition 2016.2.3. Dalam pemrosesan klasifikasi, data yang digunakan adalah hasil dari VSM yang mengitung nilai IDF lalu akan diklasifikasi menggunakan algoritma *Support Vector Machine.*

Gambar 4.5. Hasil Klasifikasi Algoritma *Support Vector Machine*

Pada gambar 4.5 adalah hasil akhir dari klasifikasi menggunakan algoritma *Support Vector Machine*. Berikut adalah detail hasil klasifikasi dengan menggunakan *Confusion Matrix:*

Tabel 4.3 Tabel *Confusion Matrix*

		Diklasifikasikan sebagai	
		False	*True*
Klasifikasi yang benar	*False*	198	2
	True	12	188

Pada kelas *false* algoritma *Support Vector Machine* berhasil mengklasifikasikan sebanyak 210 *tweet*. Namun yang sesuai dengan kelas *false* hanya sebanyak 198 *tweet* dan 12 *tweet* lainnya algoritma salah dalam mengklasifikasinya. Dan 12 *tweet* seharusnya kelas true dianggap kelas *false* oleh algoritma *Support Vector Machine*. Pada kelas true algoritma *Support Vector Machine* mengklasifikasi sebanyak 190 *tweet*. Dengan kelas yang sesuai sebanyak 188 *tweet* dan 2 lainnya salah dalam pengklasifikasiannya. Berikut adalah hasil *Precision, Recall* dan Akurasi tabel *Confusion Matrix*:

Tabel 4.4 Tabel Hasil *Precision, Recall*

Kelas	*Precision*	*Recall*
False	0.94	0.99
True	0.99	0.94
avg	0.97	0.96

Berdasarkan tabel 4.4 pada kelas false memiliki nilai *precision* sebesar 0.94 dan nilai *recall* 0.99. dan pada kelas true memiliki nilai *precision* sebesar 0.99 dan nilai *recall* 0.94.

Tabel 4.5 Tabel Hasil Akurasi

Cross Validation	Akurasi
1	0.925
2	0.975
3	0.875
4	1
5	1
6	1
7	0.975
8	0.975
9	0.975
10	0.95
Akurasi Akhir	0.96

Berdasarkan hasil akurasi dari seluruh *Cross Validation* mendapatkan nilai rata-rata sebesar 0.96.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil penelitian ini dengan judul "Kinerja Algoritma *Support Vector Machine* dalam Menentukan Informasi Banjir di Twitter" dapat disimpulkan bahwa hasil klasifikasi dari kinerja algoritma *Support Vector Machine* dalam menentukan tweet mengenai banjir dianggap bagus karena memiliki nilai *precision* sebesar 0.97, *recall* sebesar 0.96 dan akurasi 0.96.

5.2. Saran

Peneliti memiliki saran untuk penelitian lainnya yang akan dilakukan terkait dengan penggunaan algoritma *Support Vector Machine* dalam mengklasifikasi teks, yaitu :

1. Menggunakan metode *K-Fold Cross Validation* sehingga hasil klasifikasi menjadi semakin valid.
2. Memperbanyak jumlah data, sehingga algoritma dapat mempelajari berbagai macam karakteristik jenis data. Dan semakin akurat juga hasilnya.
3. Menambahkan satu atau beberapa algoritma sebagai pembanding hasil akurasi, sehingga menemukan algoritma yang terbaik.

6. Daftar Pustaka

[1] Dika Putri Utama, A., & Permana, P. Penggunaan Twitter sebagai Media Pembelajaran untuk Meningkatkan Kemampuan Menulis Kalimat Sederhana dalam Pembelajaran Bahasa Jerman.

[2] Dina, M. (2015). Skripsi Penerapan Data Mining untuk Rekomendasi Beasiswa Pada SMA Muhammadiyah Gubug Menggunakan Algoritma C4. 5., Jurusan Teknik Informatika, Fakultas Ilmu Komputer, UNIDUS, Semarang.

[3] Prilianti, K. R., & Wijaya, H. (2014). Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering. Jurnal Cybermatika, 2(1).

[4] Widiastuti, D. (2007). Analisa Perbandingan Algoritma SVM, Naive Bayes, dan Decision Tree dalam Mengklasifikasikan Serangan (Attacks) pada Sistem Pendeteksi Intrusi. Jur. Sist. Inf. Univ. Gunadarma, 1-8.

[5] Zakapedia, 2015. Pengertian Banjir, Penyebab, Dampak, Cara Menanggulangi. http://www.artikelsiana.com/2015/08/pengertian-banjir-penyebab-dampak-cara.html (diakses pada tanggal 4/1/17)