Perbandingan Metode Random Forest dan KNN pada Analisis Sentimen Twitter

Dwi Ahmad Dzulhijjah[1], Hafidz Sanjaya[2], Aji Said Wahyudi Hidayat[3], Almi Yulistia Alwanda[4], Ema Utami*[5]
[1,2,3,4,5,6]Magister of Informatics Engineering, Universitas Amikom Yogyakarta
E-mail: [1]dwi.ahmad@gmail.com, [2]hafidz@gmail.com, [3]ajisaid.wahyudi1@gmail.com,
[4]almi.yulistia@students.amikom.ac.id, *[5]**ema.u@amikom.ac.id**

Abstrak

Twitter menjadi salah satu platform media sosial yang sering digunakan untuk menyampaikan berbagai keresahan terhadap berbagai permasalahan yang ada termasuk dengan program-program yang dibuat oleh pemerintah. Tweets adalah salah satu layanan yang disediakan kepada penggunanya dimana tweets ini berisi ungkapan pendapat pengguna yang dapat juga dibaca oleh pengguna lain. Pada penelitian ini dilakukan perbandingan antara dua algoritma klasifikasi yaitu Support Vector Machine dan K-nearest Neighbor dari segi akurasi. Perbandingan ini tidak lain bertujuan untuk mengetahui algoritma klasifikasi mana yang dapat menghasilkan akurasi terbaik dalam mengklasifikasi analisis sentiment data twitter. Setelah dilakukan pengujian dan evaluasi didapatkan hasil akurasi dari algoritma SVM sebesar 83% dan KNN sebesar 49%.

Kata Kunci*—Analisis Sentimen; KNN; Support Vector Machine*

1. PENDAHULUAN

Media sosial merupakan sebuah platform digital yang dapat memfasilitasi pengguna dalam berinteraksi, salah satunya adalah twitter. Twitter menjadi salah satu platform media sosial yang sering digunakan untuk menyampaikan berbagai keresahan terhadap berbagai permasalahan yang ada termasuk dengan program-program yang dibuat oleh pemerintah. Tweets adalah salah satu layanan yang disediakan kepada penggunanya dimana tweets ini berisi ungkapan pendapat pengguna yang dapat juga dibaca oleh pengguna lain.

Analisis sentiment saat ini menjadi bidang penelitian yang cukup populer, dimana setiap isu yang berkembang dijadikan sasaran untuk dapat diteliti. Analisis sentimen ini sendiri merupakan cara yang digunakan untuk dapat menentukan apakah nada emosional pesan yang disampaikan pengguna twitter dalam menanggapi suatu permasalahan tersebut positif, negative atau bahkan netral.

Semakin banyak tweets yang berkembang terhadap suatu permasalahan, maka akan semakin banyak pula informasi yang akan dihasilkan. Hal ini karena tweets sendiri mengandung sentimen yang dapat dijadikan tolak ukur pandangan masyarakat yang dapat djadikan sebagai bahan evaluasi kedepannya..

2. METODE PENELITIAN

Pada penelitian ini dilakukan perbandingan antara dua algoritma klasifikasi yaitu Support Vector Machine (SVM) dan K-nearest Neighbor (KNN) dari segi akurasi. Perbandingan ini tidak lain bertujuan untuk mengetahui algoritma klasifikasi mana yang dapat menghasilkan akurasi terbaik dalam mengklasifikasi analisis sentimen data twitter.

Dimulai dengan pengumpulan data dari media sosial Twitter menjadi sebuah dataset, kemudian dibuat metode evaluasinya dengan menggunakan confusion matrix, dimana outputnya berupa sejumlah data yang bernilai true positif, true negatif, false positif, dan false negatif, setelah itu dihitung nilai akurasi, presisi, recall, dan F1-Score dari dataset tersebut, sehingga diperoleh nilai tersebut.

2.1. Media sosial

Media sosial menurut Kaplan dan Haenin didefinisikan sebagai "sebuah kelompok aplikasi berbasis internet yang berdasar pada ideology dan teknologi Web 2.0 yang memungkinkan penciptaan dan pertukaran konten berupa teks, gambar, video, dan sebagainya". Secara sederhananya media sosial adalah sebuah platform digital yang dapat memungkinkan penggunanya dalam mengekspresikan diri entah itu berupa teks, gambar, maupun video dan juga untuk berinterkasi dengan sesama pengguna [1].

2.2. Text mining

Text mining didefinisikan sebagai menambang data berupa teks yang bersumber dari dokumen (Langgeni, Baizal & W, 2010). Mining yang berarti penambangan dimana secara sederhananya text mining ini adapat diartikan sebagai salah satu teknik yang digunakan untuk menambang atau menggali sebuah informasi yang berasal dari dokumen teks [1].

2.3. Klasifikasi teks

Klasifikasi teks secara sederhana dapat didefinisikan sebagai proses pengelompokkan data ke dalam kelas yang telah ditentukan sebelumnya. Klasifikasi teks terbagi ke dalam 2 jenis yaitu supervised dan unsupervised. Supervised sendiri adalah proses klasifikasi teks dengan menggunakan metode learning pada data teks yang sudah memiliki kelas atau label, sedangkan untuk unsupervised merupakan metode klasifikasi teks yang tidak memiliki kelas atau label [2].

2.4. KNN

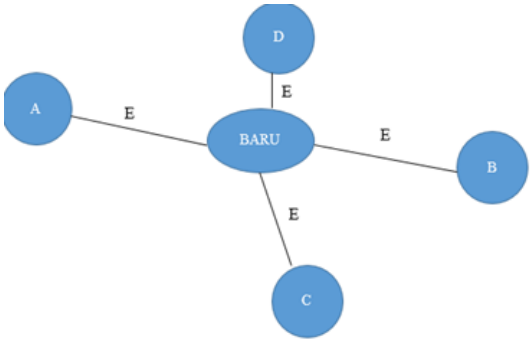
K-Nearest Neighbor atau yang biasa disingkat KNN adalah salah satu algoritma klasifikasi yang digunakan untuk dapat mencari kelompok k objek dalam data training yang paling dekat atau paling mirip dengan objek pada data baru atau data testing. KNN sendiri termasuk kelompok instance-based learning dan juga termasuk ke salah satu teknik lazy learning [2].

Sebagai contoh kasus, ketika diinginkan untuk memecahkan masalah satu mencari solusi untuk pasien baru dengan menggunakan solusi yang sudah digunakan pada pasien lama. Untuk dapat mencari solusi dari pasien baru dilakukan kedekatan dengan kasus pasien lama, ketika kasus pasien lama memiliki kedekatan dengan kasus pasien baru, maka solusi kasus lama dapat digunakan untuk mengatasi kasus pasien baru.

Terdapat pasien baru dan 4 pasien lama, sebutlah A, B, C dan D itu pasien lama dan E adalah pasien baru yang akan dicarikan kedekatan kasus nya dengan 4 pasien lama tersebut.

Digambarkan E1 adalah jarak antara pasien baru dengan pasien A, E2 jarak antara pasien baru dengan pasien B, E3 jarak antara pasien baru dan pasien C, dan E4 adalah jarak antara pasien baru dengan pasien D. dapat dilihat pada ilustrasi gambar 1.1. :

P-ISSN: 2089-676X
E-ISSN: 2549-0796



Gambar 1..Ilustrasi algoritma KNN

Dapat dilihat pada gambar 1.1, bahwa yang memiliki jarak terdekat antara pasien baru dan pasien lama adalah pasien D. sehingga nantinya solusi yang dapa digunakan untuk mengatasi kasus pasien baru adalah solusi pada kasus pasien D.

2.5. SVM

Support Vector machine merupakan salah satu metode yang digunakan dalam klasifikasi dan regresi, dimana untuk metode SVM ini senditi mencoba untuk menemukan hyperplane (ruang dengan dimensi N) yang memisahkan data pada setiap kelas sedemikian rupa sehingga margin antara kelas tersebut maksimal [3].

2.6. Preprocessing

Tahap pertama yang harus dilakukan sebelum data dapat digunakan adalah preprocessing dimana pada tahapan ini data akan dipersiapkan hingga pada akhirnya bisa untuk digunakan. Seperti nama Preprocessing atau sebelum data itu di proses ada beberapa tahapan yang perlu dilewati. Preprocessing ini sendiri bertujuan untuk dapat membersihkan data yang tidak relevan, memeriksa apakah terdapat missing value dan mempersiapkan data yang tidak terstruktur menjadi data terstruktur [1]. Pada penelitian ini beberapa tahapan Preprocessing yang dilakukan sebagai berikut :

2.6.1. Case folding

Pada tahapan ini dilakukan konversi teks menjadi bentuk yang standar dimana secara sederhananya pada tahapan ini dilakukan untuk menyeragamkan penggunaan huruf kapital dalam teks. Pada tahap ini biasanya dipilih lowercase untuk membuat hurud kapital menjadi lowercase [2] [4].

2.6.2. Stopword removal

Stopword merupakan salah satu tahapan dari preprocessing dimana pada tahapan ini dilakukan penghapusan kata yang tidak memiliki makna dalam kalimat. Misalnya seperti “jika”, “di”, “yaitu”, “dari”, dan sebagainya. Pada tahapan ini kata-kata pada dataset stopword akan dihapus [4].

2.6.3. Stemming

Stemming merupakan tahapan dimana semua imbuhan kata akan dihilangkan dan akan dirubah menjadi kata dasar. Misalnya seperti kata “kebersamaan” akan diubah menjadi kata “sama” karena kata “kebersamaan” itu didasari oleh kata “sama” [2].

2.6.4. Normalization

Pada tahapan ini kalimat pada kolom tweet dataset akan di normalisasi. Dimana kata yang tidak baku nantinya akan diubah menjadi kata baku. Contohnya seperti "aamiin adek abis 3x" setelah dilakukan normalisasi maka akan menjadi "amin adik habis tiga kali" [5].

2.7. Evaluation

Pada penelitian ini metode evaluasi yang digunakan adalah confusion matrix dimana confusion matrix ini sendiri memberikan gambaran atau informasi mengenai hasil klasifikasi yang sudah dilakukan oleh sistem. Gambaran yang diberikan oleh confusion matrix ini tergantung dari jumlah kelas data yang diklasifikasi, dimana keluarannya atau labelnya dapat berupa dua kelas atau lebih. Evaluasi sendiri bertujuan untuk mengetahui kinerja atau performa dari model yang diusulkan [2] [4].

Hasil proses klasifikasi pada confusion matrix itu direpresentasikan dengan 4 istilah, yaitu True Positif, False Positif, True Negatif dan False Negatif. Keluaran confusion matrix tergantung dari jumlah kelas yang dimiliki dataset, pada tabel 1 merupakan contoh confusion martrix dengan keluaran 2 kelas. Confusion matrix dapat dipahami melalui tabel 1. :

Tabel 1. Daftar evaluasi klasifikasi

Kelas	Terklasifikasi positif	Terklasifikasi negative
Positif	TP (*True* Positif)	FN (*False* Negatif)
negatif	FP (*False* Positif)	TN (*True* Negatif)

Nilai True Positif (TP) dan True Negatif (TN) adalah hasil klasifikasi yang benar sedangkan nilai False Positif (FP) merupakan nilai yang dimana dia di klasifikasikan positif akan tetapi sebenarnya dia bernilai negatif dan untuk False Negatif (FN) merupakan nilai yang dimana dia diklasifikasikan negatif akan tetapi sebenarnya dia bernilai positif.

Dengan menggunakan confusion matrix, ada beberapa nilai yang dihasilkan yaitu seperti nilai akurasi, presisi, recall dan F1 score. Akurasi merupakan gambaran sebarapa akurat system dalam melakukan klasifikasi pada data dengan benar. Kemudian untuk presisi adalah rasio dari jumlah prediksi positif yang benar terhadap keseluruhan jumlah prediksi positif. Recall atau sensitivity sendiri adalah gambaran dari keberhasilan model dalam menemukan kembali sebuah informasi sedangkan F-1 Score adalah perbandingan rata-rata precision dan recall yang dibobotkan [3].

Nilai akurasi, presisi, recall dan F-1 Score didapatkan dengan persamaan pada gambar 2., 3., dan 4. :

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\,\%$$

Gambar 2. Rumus akurasi

$$Presisi = \frac{TP}{FP+TP} * 100\,\%$$

Gambar 3. Rumus presisi

$$Recall = \frac{TP}{FN + TP} * 100\%$$

Gambar 4. Rumus Recall.

Rumus F-1 *Score* = (2 * *Recall* * *Precision*) / (*Recall* + *Precision*)

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan sejumlah 3.935 record dan dari proses pelabelan, didapatkan hasil bahwa tweet yang termasuk ke dalam kelas positif sebanyak 42,2% dan tweet yang termasuk ke dalam kelas negatif sebanyak 57,7%. Dari 3.935 record yang digunakan, dilakukan data split untuk membagi data ke dalam dua model yaitu data training dan data testing. Dimana data training yang digunakan sejumlah 30% dan data testing 70%.

Algoritma SVM dan KNN yang sudah dipilih diimplementasikan menggunakan Bahasa pemrograman Python dan machine learning library "sklearn". Kemudian hasil dari pengujian akan dievaluasi dengan menggunakan confusion matrix.

Didapatkan confusion matrix dari pengujian sistem dengan menggunakan algoritma SVM dan KNN seperti yang tertera pada tabel 1.2 dan tabel 1.3 :

Tabel 2. Confusion matrix dengan algoritma SVM

n = 3935	Aktual : Positif (1)	Aktual : Negatif (0)
Prediksi : Positif (1)	TP : 2047	FP : 439
Prediksi : Negatif (0)	FN : 221	TN : 1228
	2268	1667

Tabel 3. Confusion matrix dengan algoritma KNN

n = 3935	Aktual : Positif (1)	Aktual : Negatif (0)
Prediksi : Positif (1)	TP : 331	FP : 53
Prediksi : Negatif (0)	FN : 1937	TN : 1614
	2268	1667

Didapatkan perbandingan hasil perhitungan akurasi, presisi, recall dan F-1 Score berdasarkan confusion matrix yang dihasilkan pada tabel 1.2 dan tabel 1.3 di atas.

Tabel 4. Hasil perhitungan akurasi, presisi, recall, dan F1-score

Evaluasi	*SVM*	KNN
Akurasi	83%	49%
Presisi positif Presisi negatif	0.74 0.90	0,97 0.15
Recall positif *Recall negatif*	0,85 0.82	0,45 0.86
f1-score positif *f1 score negatif*	0,79 0.86	0.62 0.25

Dapat dilihat pada tabel 1.4 yaitu hasil evaluasi menunjukkan bahwa metode SVM menghasilkan nilai akurasi sebesar 83% sedangkan metode KNN menghasilkan akurasi sebesar 49%. Hal ini menunjukkan bahwa metode SVM dapat menghasilkan akurasi terbaik pada kasus ini.

4. KESIMPULAN

Dari hasil klasifikasi perbandingan metode SVM dan KNN pada analisis sentimen Twitter, dapat disimpulkan bahwa metode SVM memiliki perhitungan evaluasi yang lebih baik dibanding dengan KNN, meskipun KNN memiliki nilai presisi positif dan recall negatif lebih baik dibanding SVM, namun KNN kurang memiliki akurasi yang baik. Diharapkan perhitungan evaluasi analisis sentimen ini tidak hanya menggunakan confusion matrix, namun juga menggunakan algoritma lainnya, seperti Naive Bayes, dan C4.5.

DAFTAR PUSTAKA

[1] S. A. Putri, P. D. Kusuma, and C. Setianingsih, "CLUSTERING TOPIK PADA DATA SENTIMEN BPJS KESEHATAN MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION," e-Proceeding of Engineering, vol. 8, no. 5, pp. 6097–6105, 2021.

[2] O. S. Y. Prakasa and K. M. Lhaksamana, "KLASIFIKASI TEKS DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR PADA KASUS KINERJA PEMERINTAH DI TWITTER," e-Proceeding of Engineering, vol. 5, no. 3, pp. 8237–8248, 2018.

[3] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 5, no. 2, pp. 640–651, Apr. 2021, doi: 10.30865/mib.v5i2.2937.

[4] M. I. Fikri, T. S. Sabrila, and Y. Azhar, "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter," SMATIKA Jurnal, vol. 10, no. 2, pp. 71–76, 2020.

[5] J. C. W. Pantouw, "PERBANDINGAN KLASIFIKASI ROCCHIO DAN MULTINOMIAL NAÏVE BAYES PADA ANALISIS SENTIMEN DATA TWITTER BAHASA INDONESIA," 2017.