COMPARISON OF SVM, KNN, AND NAIVE BAYES METHOD WITH N-GRAM IN TRAFFIC ACCIDENT CLASSIFICATION

Dhika Malita Puspita Arum*; Andri Triyono

ABSTRACT

Traffic accidents that occur in Indonesia are still relatively high, the information can be easily obtained through social media, one of which is Twitter. The amount of traffic accident information can be processed and classified according to certain categories. Traffic accident data classification is done using SVM, KNN and Naïve Bayes methods using n-gram feature extraction. The results of this study indicate the best accuracy is 87.63 using the KNN method.

***Keywords**; Traffic Accident, Classification, SVM, KNN, Naïve Bayes, N-Gram*

Correspondence:
Dhika Malita Puspita Arum
Universitas An Nuur, Email; dhika.malita11@gmail.com

PENDAHULUAN

Kecelakaan lalu lintas di Indonesia yang menyebabkan kematian, luka berat, luka ringan serta kerugian material masih relatif tinggi. Pada tahun 2018, 109.215 kasus kecelakaan tercatat dengan 29.472 korban meninggal, 13.315 korban luka berat, 130.571 korban luka ringan dan 213 866 juta kerugian material (bps.go.id). Dengan perkembangan teknologi yang ada saat ini, informasi mengenai kecelakaan lalu lintas dapat dengan mudah di dapatkan salah satunya adalah dari media social. Salah satu media social yang banyak penggunanya adalah twitter . Twitter termasuk 10 besar situs web yang paling sering di kunjungi (Godara and Kumar, 2020) Banyaknya informasi yang didapat melalui twitter khususnya informasi kecelakaan lalu lintas dapat di proses dan di klasifikasikan menurut kategori tertentu.

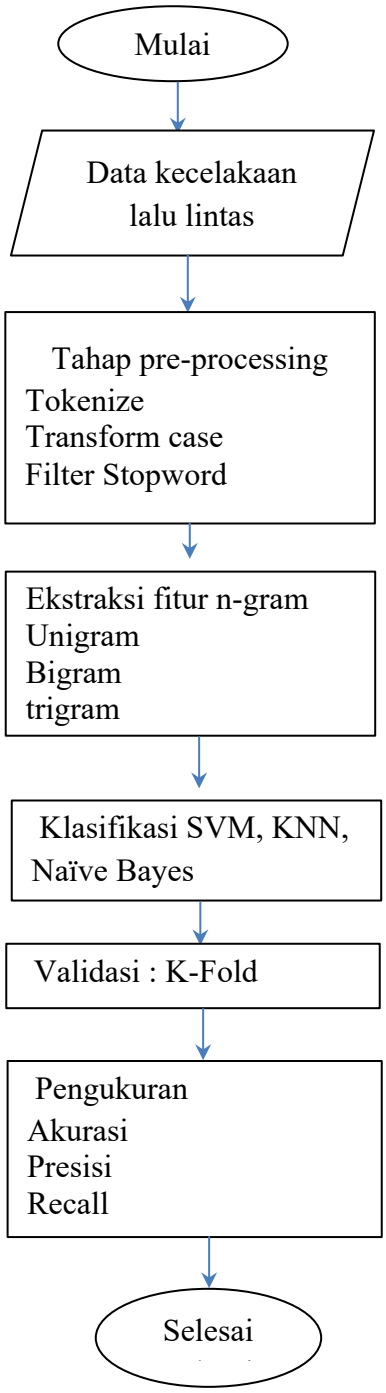
Pada penelitian sebelumnya dengan dataset yang sama dengan dataset yang diusulkan peneliti dan metode yang sama pula yaitu SVM, KNN dan Naive Bayes tanpa ekstraksi fitur n-gram Di dapatkan hasil akurasi tertinggi dengan menggunakan metode SVM yaitu 85 % (Ganeswangga *et al.*, 2020), penetian lain tentang analisis sentiment pada media social twitter menggunakan naive bayes classifier dengan ekstraksi fitur N-gram . pada penelitian ini data twitter didapat dengan memanfaatkan API search twitter yang berhubungan dengan tariff dasar

listrik, hasil pengujian dengan menguunakan metode naïve bayes mencapai 89, 67 % sebelum di terapkan n-gram dan tingkat akurasi meningkat 2,33 % menjadi 92 % setelah di terapkan n-gram (Nugroho, 2018).

METODE PENELITIAN

Data

Penelitian ini menggunkan data public yang dapat di akses oleh siapa saja. Dataset diambil dari alamat https://www.kaggle.com/dodyagung/accident?select=twitter_label_manual.csv. Dataset ini berkaitan dengan kecelakaan lalu lintas yang di ambil dari twitter yang berjumlah 1000 record



Gambar 1: Metode yang di usulkan

Tahap Pre-processing

Pre- Processing merupakan langkah pertama yang di lakukan untuk pemprosesan dokumen teks. Proses pre-processing pada penelitian ini diantaranya

yaitu Tokenize , Transform Case dan Filter Stopword. Tokenize adalah memecah kalimat menjadi bentuk yang lebih sederhana, yaitu kata atau *term.* Tahap preprocessing selanjutnya yaitu Transform *case* atau mengubah huruf ke dalam bentuk huruf kecil. Tahap selanjutnya yang tidak kalah penting yaitu *filter stopwords* adalah membuang informasi yang tidak penting, tidak relevan, dan tidak dibutuhkan dalam suatu dokumen teks(Trianto, Triyono and Arum, 2020).

N-Gram

N-gram(Chandra, Indrawan and Sukajaya, 2016) adalah potongan sejumlah n karakter dari sebuah string . Ngram merupakan sebuah metode yang diaplikasikan untuk pembangkitan kata atau karakter. Metode n-gram ini digunakan untuk mengambil potongan-potongan karakter huruf sejumlah n dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen.

Klasifikasi dengan SVM, KNN dan Naive Bayes

Support Vector Machine

salah satu metode *machine learning* untuk permasalahan klasifikasi yang paling popular yaitu SVM (Support Vector Machine(Hafidz and Liliana, 2021), yang mencoba mendapatkan *hyperplane* yang paling jauh memisahkan data poin terdekat dari setiap kelas (Hutto, C.J. and Gilbert, 2014).

KNN

K-Nearest Neigbhor (KNN) salah satu metode yang populer untuk menghasilkan klasifikasi teks yaitu K-Nearest Neighbor (Sreemathy, 2012),yaitu dengan melakukan proses pembelajaran dari data latih untuk menentukan kelompok k objek. Sehingga dalam menentukan hasil klasifikasi KNN melihat jarak terdekat dari objek dengan masing-masing kelompok. Jarak tersebut diperoleh dari hasil kedekatan antara data masukan dengan data yang berada dalam kelompok berdasarakan nilai sejumlah fitur yang ada. Tetapi KNN juga miliki kekurangan salah satunya yang besar dalam aspek komputasi perhitungan

Naive Bayes

Naive bayes classifier merupakan sebuah metode klasifikasi yang berakar pada teorema bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal

sebagai Teorema Bayes. Ciri utama dari naive bayes classifier ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi atau kejadian (Chandra, Indrawan and Sukajaya, 2016).

Validasi dan Pengukuran Performa

Validasi yang digunakan yaitu k-fold validation dengan k=10 untuk proses pelatihan dan pengujian ,yang berarti dataset di pecah menjadi 10 bagian , 9 bagian di gunakan untuk proses pelatihan dan 1 proses pengujian , proses ini dilakukan sebanyak 10 kali (Wahono, Herman and Ahmad, 2014) .

HASIL DAN PEMBAHASAN

Tabel 1; Menggunakan Metode SVM dengan n-gram

Model	Akurasi	Presisi	Recall	f-measure
SVM dengan unigram	**85,53**	**82,67**	98,13	**89,73902**
SVM dengan bigram	75,35	72,41	99,69	83,88789
SVM dengan trigram	71,86	69,52	**100**	82,01982

Table 1; menunjukkan hasil pengujian klasifikasi dengan menggunakan metode SVM dan n -gram . hasil terbaik di dapat pada pemakaian unigram pada metode SVM yaitu dengan nilai akurasi sebesar 85,53, nilai presisi sebesar 82,6, nilai recall pada pemakaian trigram yaitu 100

Tabel 2; Menggunakan Metode Naive Bayes dengan n-gram

Model	akurasi	Presisi	recall	f-measure
Naïve bayes dengan unigram	83,53	87,99	86,13	87,05007
Naïve bayes dengan bigram	**84,83**	89,93	**86,14**	**87,99421**
Naïve bayes dengan trigram	84,73	**92,54**	83,02	87,52188

Table 2; menunjukkan hasil pengujian klasifikasi dengan menggunakan metode Naive Bayes dan n -gram . hasil terbaik di dapat pada pemakaian bigram pada metode Naive Bayes yaitu dengan nilai akurasi sebesar 84,83, pemakaian trigram dengan nilai presisi sebesar 92,54 dan, nilai recall pada pemakaian bigram yaitu 86,14

Tabel 3; Menggunakan metode KNN dengan n-gram

Model	Akurasi	Presisi	recall	f-measure
KNN (k=4) dengan unigram	85,13	89,96	86,44	88,16488
KNN (k=4) dengan bigram	86,13	91,39	86,61	88,93582
KNN (k=4) dengan trigram	85,83	91,22	86,3	88,69182
KNN (k=5) dengan unigram	84,63	89,4	86,29	87,81747
KNN (k=5) dengan bigram	85,73	91,45	85,82	88,5456
KNN (k=5) dengan trigram	86,53	91,09	87,69	89,35767
KNN (k=6) dengan unigram	85,63	89,96	87,37	88,64609
KNN (k=6) dengan bigram	85,63	91,12	87,54	89,29413
KNN (k=6) dengan trigram	87,13	91	88,78	89,87629
KNN (k=7) dengan unigram	86,33	90,19	88,31	89,2401
KNN (k=7) dengan bigram	**87,63**	**91,76**	88,78	90,24541
KNN (k=7) dengan trigram	**87,63**	90,85	**89,87**	**90,35734**

Table 3 menunjukkan hasil pengujian klasifikasi dengan menggunakan metode KNN dan n -gram . hasil terbaik di dapat pada pemakaian bigram dan trigram pada metode KNN yaitu dengan nilai akurasi sebesar 87,63, pemakaian bigram dengan nilai presisi sebesar 91,76 dan nilai recall pada pemakaian trigram yaitu 89,87

SIMPULAN DAN SARAN

Dari ketiga metode yaitu Support Vectore Machine, K-Nearest Neigbhor (KNN) dan Naïve Bayes di dapatkan hasil akurasi tertinggi yaitu dengan menggunakan metode KNN dengan bigram yaitu 87,6 %

DAFTAR PUSTAKA

Chandra, D. N., Indrawan, G. and Sukajaya, I. N. (2016) 'Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram', *Jurnal Ilmiah Teknologi Informasi Asia*, 10(1), pp. 11–19.

Ganeswangga, A. *et al.* (2020) 'International Journal of Emerging Trends in Engineering Research Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter24842020.pdf Evaluation of E-Wallet Implementation : Proposed New Model', 8(4), pp. 1096–1102.

Godara, N. and Kumar, S. (2020) 'ISSN NO : 0042-9945 Twitter Sentiment Classification using Machine Learning Techniques', XI(10), pp. 10–20.

Hafidz, N. and Liliana, D. Y. (2021) 'Klasifikasi Sentimen pada Twitter Terhadap WHO', *JURNAL RESTI (*Rekayasa *Sistem dan Teknologi Informasi)*, 5(2), pp. 213–219.

Hutto, C.J. and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-based Model for', *Eighth International AAAI Conference on Weblogs and Social Media*, p. 18.

Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

Nugroho, A. (2018) 'Analisis Sentimen Pada Media Sosial Twitter Menggunakan Naive Bayes Classifier Dengan Ekstrasi Fitur N-Gram', *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, 2(2), p. 200. doi: 10.30645/j-sakti.v2i2.83.

Sreemathy, J. (2012) 'an Efficient Text Classification Using Knn and Naive Bayesian', 4(03), pp. 392–396.

Trianto, R. B., Triyono, A. and Arum, D. M. P. (2020) 'Klasifikasi Rating Otomatis pada Dokumen Teks Ulasan Produk Elektronik Menggunakan Metode N-gram dan Naïve Bayes', *Jurnal Informatika Universitas Pamulang*, 5(3), p. 295. doi: 10.32493/informatika.v5i3.6110.

Wahono, R. S., Herman, N. S. and Ahmad, S. (2014) 'A comparison framework of classification models for software defect prediction', *Advanced Science Letters*, 20(10–12), pp. 1945–1950. doi: 10.1166/asl.2014.5640