J-ENSISTEC (Journal of Engineering and Sustainable Technology)
Vol. 11|No. 01, Desember 2024
ISSN : 2477-359X (Online)
ISSN: 2407-6007 (Print)
pp. 10202-10206

STUDI KLASIFIKASI TOPIK BERITA DENGAN ALGORITMA MACHINE LEARNING

Guruh Wijaya[1], Dudi Irawan[2], Zainul Arifin[3], Hardian Oktavianto[4], Miftahur Rahman[5], Ginanjar Abdurrahman[6]
[1,2,3,4,5,6]*Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember*
Email: [1]guruh.wijaya@unmuhjember.ac.id, [2] dudi.irawan@unmuhjember.ac.id,
[3]zainul.arifin@unmuhjember.ac.id, [4]hardian@unmuhjember.ac.id, [5]miftahurrahman@unmuhjember.ac.id,
[6]abdurrahmanginanjar@unmuhjember.ac.id

ABSTRACT

As a result of the use and access of social media, it also has an impact on increasing the amount of data and information, especially text data. Text has become one of the most natural forms of data that is stored, so that the field of text mining is believed to be an advanced field of data mining. Facts that emerge from research studies that have been conducted show that 80% of company information is presented in text documents. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, and clustering. The text mining classification method is one technique that can be used to carry out classification. Text classification specifically works to group text documents based on categories, and within the scope of news datasets, categories are generally divided into politics, economics, military, sports and others. Statistical methods are one of the most frequently applied methods in text emotion classification. As a method in statistics, Naive Bayes is a classification algorithm that is easy to understand in text classification. Apart from that, Naïve Bayes has good classification effects and performance for processing large-scale data. The conclusion of this research is, Naïve Bayes gets an accuracy value of 77.78%. Random Forest gets an accuracy of 70.1%. KNN gets an accuracy of 24.88% and SVM gets an accuracy value of 80.60%. Meanwhile, the respective running times are Naïve Bayes 0.046 seconds, Random Forest 150 seconds, KNN 15 seconds, and SVM 0.43 seconds.

Keywords*: knn, naïve bayes, random forest, sentiment analysis, support vector machine*

Riwayat Artikel :

Tanggal diterima : 29-11-2024

Tanggal revisi : 02-12-2024

Tanggal terbit : 05-12-2024

DOI :

https://doi.org/10.31949/jensitec.v11i01.12037

1. PENDAHULUAN

Berita telah menjadi sebuah lembaga yang digunakan untuk menyebarkan informasi terkini yang dibutuhkan masyarakat. Melalui berbagai instansi, berita disampaikan dengan menggunakan media seperti Media Online, Televisi, Surat Kabar, Radio, dan berbagai media lainnya. Secara umum berita yang disampaikan di media terdiri dari beberapa kategori seperti politik, olahraga, ekonomi, kesehatan, dan lain sebagainya [1]. Dan dengan seiring berkembangnya teknologi informasi dan komunikasi masyarakat lebih sering memilih media online khususnya media sosial untuk memenuhi kebutuhan informasi setiap saat, karena dapat diakses kapanpun dan

dimanapun [2] [3] [4]. Akibat penggunaan dan akses media sosial tersebut maka berdampak pula pada peningkatan jumlah data dan informasi khususnya data teks [5].

Penambangan teks mengacu pada proses umum mengekstraksi pola atau pengetahuan yang terkandung dalam dokumen teks yang tidak terstruktur. Teks telah menjadi salah satu bentuk data paling alami yang disimpan, sehingga bidang text mining dipercaya sebagai bidang lanjutan dari data mining. Fakta yang muncul dari penelitian penelitian yang telah dilakukan menunjukkan bahwa 80% informasi perusahaan disajikan dalam dokumen teks. Penambangan teks adalah bidang multidisiplin, yang melibatkan pengambilan informasi, analisis teks, ekstraksi informasi, dan pengelompokan [6]. Metode klasifikasi text mining merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi. Penambangan teks adalah variasi penambangan data yang mencoba mengidentifikasi pola menarik dalam kumpulan data tekstual yang besar. Selain klasifikasi, text mining juga digunakan untuk menangani masalah yang berkaitan dengan clustering, ekstraksi informasi, dan pengambilan informasi [1].

Klasifikasi teks secara khusus bekerja mengelompokkan dokumen teks berdasarkan kategori, dan dalam lingkup dataset berita, kategori umumnya dibedakan menjadi politik, ekonomi, militer, olahraga dan lain-lain. Metode statistik adalah salah satu metode yang paling sering diterapkan dalam klasifikasi emosi teks. Sebagai salah satu metode di dalam statistik, Naive Bayes merupakan salah satu algoritma klasifikasi yang mudah dipahami dalam klasifikasi teks, selain itu Naïve Bayes memiliki efek klasifikasi dan kinerja yang baik untuk memproses data skala besar [7].

Penelitian – penelitian yang menjadi rujukan pada penelitian ini ada beberapa. Yang pertama adalah penelitian yang dilakukan Saputra pada tahun 2020 yang melakukan klasifikasi kepuasan pelanggan provider XL dengan melakukan perbandingan diantara menggunakan Naïve Bayes dengan SMOTE yang digabungkan Nave Bayes. Hasil penelitian menunjukkan bahwa gabungan SMOTE dengan Naïve Bayes mempunyai kinerja yang lebih baik, nilai akurasi 86.33%, nilai presisi 82.85%, dan nilai recall 92.38% [8]. Yang kedua adalah penelitian oleh Syahrini pada 2020 yang berjudul Sentiment Analysis Of Facebook Comments On Indonesian Presidential Candidates Using The Naive Bayes Method, peneliti mengkaji analisis sentimen komentar facebook 2 calon presiden asli indonesia pada tahun 2014 melalui data komentar facebook pada beberapa status yang diposting pada 2 akun resmi calon presiden indonesia, karena facebook merupakan media sosial yang banyak digunakan oleh masyarakat indonesia, terbukti media sosial Facebook menduduki peringkat ke 3 di Indonesia. Dalam proses pengklasifikasian teks peneliti menggunakan metode Naive Bayes menggunakan gabungan metode seleksi fitur yaitu information gain dan algoritma genetika mempunyai tingkat akurasi klasifikasi sentimen sebesar 83,67% yang lebih baik dari hasil sebelumnya sebesar 60,00% [9].

2. METODE PENELITIAN

Tahapan atau langkah – langkah penelitian yang akan dilakukan secara umum terdiri dari 4 buah tahapan, mulai dari studi literatur, kemudian pengumpulan dan pemrosesan dataset, dilanjutkan dengan implementasi klasifikasi, dan yang terakhir adalah penarikan kesimpulan atau analisis hasil. Tahapan atau langkah – langkah penelitian ini secara umum dapat dilihat pada gambar 1.

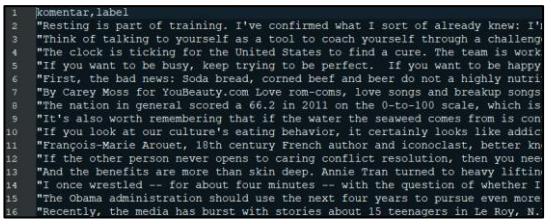
Gambar 1. Metodologi Penelitian

Studi literatur merupakan langkah yang dilakukan untuk mempelajari referensi berupa jurnal penelitian, paper, buku-buku referensi yang lain terkait dengan penelitian untuk melengkapi pengetahuan awal, guna memahami teori yang dapat digunakan untuk menunjang penelitian.

Tahap praproses data dilakukan untuk mempersiapkan data agar siap dipakai dalam keseluruhan proses klasifikasi teks. Pada penelitian ini digunakan dataset yaitu data teks berupa data sekunder yang diambil dari

ISSN : 2477-359X (Online)
ISSN: 2407-6007 (Print)

https://www.kaggle.com/datasets/rmisra/news-category-dataset. Dataset ini terdiri dari 210.000 baris teks. Data awal nantinya tidak akan dipakai secara keseluruhan, melainkan hanya akan diambil sebagian, baik secara jumlah baris maupun label kelas.



Gambar 2. Dataset

Data masukan berupa data teks akan diproses Tokenizing, kemudian dilanjutkan dengan proses Stemming, setelah itu dilakukan proses Delete Stop Words, proses selanjutnya adalah membentuk vektor representasi dari masing – masing kata, kemudian selanjutnya adalah ekstraksi fitur, dan yang terakhir adalah proses klasifikasi menggunakan algoritma tertentu, dan pada penelitian ini digunakan Naive Bayes, KNN, Random Forest, Support Vector Machine.

Analisis hasil yaitu melakukan pendeskrispsian terhadap hasil pengelompokan yang terbentuk. Pada tahap ini juga dilakukan penghitungan akurasi untuk mengetahui seberapa baik performa atau kinerja dari metode yang diterapkan.

2.1 Naïve Bayes

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat dalam teknik klasifikasi. Naive Bayes adalah klasifikasi dengan metode probabilitas dan statistik Teorema Bayes memprediksi peluang di masa depan berdasarkan pengalaman sebelumnya.

$$P(H|X) = \frac{P(X\backslash H)P(H)}{P(X)}$$

Dimana

X = Data yang dicari labelnya

H = Data pada label kelas tertentu

P (H | X) = Probabilitas H terhadap kondisi x

P (H) = Probabilitas H

P (X | H) = Probabilitas X terhadap H

P (X) = Probabilitas X

2.2 Support Vector Machine

Support Vector Machine (SVM) adalah algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi, tetapi lebih sering digunakan untuk tugas klasifikasi. SVM mencoba menemukan garis pemisah terbaik (hyperplane) yang memisahkan data dari dua kelas dengan jarak terbesar (margin) antara kedua kelas tersebut. Hyperplane yang dipilih adalah yang memiliki jarak terlebar ke titik-titik data terdekat dari kedua kelas. Titik-titik ini disebut support vectors.

Cara Kerja SVM:

Dalam data dua dimensi, SVM mencari garis lurus yang memisahkan data dari dua kelas.

Dalam dimensi yang lebih tinggi, SVM mencari sebuah bidang (plane) atau hyperplane untuk memisahkan kelas-kelas.

Ketika data tidak dapat dipisahkan secara linier (misalnya, titik – titiknya membentuk pola melingkar), SVM menggunakan fungsi matematika yang disebut kernel untuk memetakan data ke dimensi yang lebih tinggi agar dapat dipisahkan. Dengan kata lain, SVM mencoba membuat keputusan yang paling aman untuk memisahkan data, sehingga kesalahan melakukan klasifikasi dapat diminimalisir.

2.3 Random Forest

Random Forest adalah algoritma pembelajaran mesin yang sering digunakan untuk klasifikasi dan regresi. Sederhananya, ini adalah kumpulan banyak pohon keputusan yang digunakankan untuk membuat prediksi yang lebih akurat dan stabil. Random Forest bekerja dengan cara membangun banyak pohon keputusan dimana setiap pohon dilatih pada subset data yang berbeda, pohon diambil secara acak dengan pengembalian. Setiap pohon juga hanya menggunakan subset fitur atau variabel yang dipilih secara acak ketika membuat keputusan.

Secara teknis, implementasi random forest berbeda ketika digunakan pada klasifikasi dengan ketika digunakan pada regresi. Pada klasifikasi Random Forest menggunakan konsep voting, dimana setelah semua pohon memberikan

prediksi, algoritma mengambil hasil mayoritas atau suara terbanyak sebagai prediksi akhir klasifikasi. Sedangkan pada regresi, Random Forest menghitung rata-rata dari prediksi semua pohon.

2.4 KNN

K-Nearest Neighbors (KNN) adalah algoritma pembelajaran mesin yang sangat sederhana untuk klasifikasi dan regresi, yang didasarkan pada kesamaan antara data. Cara kerja KNN adalah algoritma KNN akan mencari K tetangga terdekat dengan titik data baru yang ingin diprediksi, dimana kemiripan ini umumnya dihitung menggunakan jarak, misalnya menggunakan formula jarak Euclidean.

KNN bekerja sebagai berikut:

1. Tentukan nilai K.
2. Hitung jarak titik baru ke semua titik dalam dataset.
3. Pilih K tetangga terdekat berdasarkan jarak terpendek.
4. Untuk klasifikasi, prediksi ditentukan berdasarkan mayoritas label dari K tetangga. Untuk regresi, prediksi ditentukan dengan mengambil rata-rata nilai tetangga.

Berdasarkan konsep dan cara kerja KNN maka pemilihan nilai K sangat penting, jika K terlalu kecil,misalnya 1, hasilnya bisa terlalu sensitif terhadap data tunggal, jika K terlalu besar, prediksi bisa menjadi terlalu umum.

3. HASIL DAN PEMBAHASAN

3.1 Praproses Data

Dataset awal terdiri dari 210.000 baris data dengan 42 kelas, sedangkan pada penelitian ini hanya akan diambil 20.000 baris data yang terdiri dari 4 kelas saja, yang bertujuan untuk menguji kinerja algoritma naïve bayes. Dataset ini terdiri dari 20.000 baris teks, teks masing – masing berisi kalimat berita yang diikuti dengan label kategori berita. Terdapat 4 kategori berita WELLNESS, POLITICS, ENTERTAINMENT, dan TRAVEL. Adapun jumlah baris masing – masing kategori adalah 5000 baris teks.

3.2 Klasifikasi

Proses klasifikasi teks dilakukan dengan menggunakan bantuan bahasa pemrograman python dengan memanfaatkan beberapa library yang telah tersedia. Beberapa library yang digunakan adalah pandas, numpy, dan sklearn.

```python
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

```python
x = df['komentar'].values
y = df['label'].values

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

tfidf_vectorizer = TfidfVectorizer()
tfidf_train_vectors = tfidf_vectorizer.fit_transform(x_train)
tfidf_test_vectors = tfidf_vectorizer.transform(x_test)

clf_random_forest = RandomForestClassifier()
clf_random_forest.fit(tfidf_train_vectors, y_train)
y_pred = clf_random_forest.predict(tfidf_test_vectors)

df_compare = pd.DataFrame(
    data={
        'komentar': x_test,
        'predicted_komentar': y_pred,
        'real_komentar': y_test
    },
    columns=['komentar', 'predicted_komentar', 'real_komentar'])
df_compare
```

Gambar 3. Tahap Klasifikasi

Proses klasifikasi dimulai dengan import dataset kemudian dilakukan pembobotan menggunakan TF-IDF, setelah itu baru dilakukan klasifikasi menggunakan naïve bayes. Metode seleksi fitur yang digunakan adalah TF-IDF. Adapun pembagian data latih dengan data uji adalah 80% data latih dan 20% data uji.

```python
clf_random_forest = RandomForestClassifier()
clf_random_forest.fit(tfidf_train_vectors, y_train)
y_random_forest_pred = clf_random_forest.predict(tfidf_test_vectors)
accuracy_score(y_test, y_random_forest_pred)

clf_knn = KNeighborsClassifier(n_neighbors=19)
clf_knn.fit(tfidf_train_vectors, y_train)
y_knn_pred = clf_knn.predict(tfidf_test_vectors)
accuracy_score(y_test, y_knn_pred)

clf_nb = MultinomialNB()
clf_nb.fit(tfidf_train_vectors, y_train)
y_nb_pred = clf_nb.predict(tfidf_test_vectors)
accuracy_score(y_test, y_nb_pred)

clf_svc = LinearSVC(dual=True)
clf_svc.fit(tfidf_train_vectors, y_train)
y_svc_pred = clf_svc.predict(tfidf_test_vectors)
accuracy_score(y_test, y_svc_pred)
```

Gambar 4. Tahapan Klasifikasi dan Uji

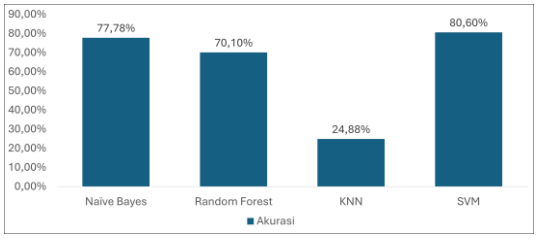
Pengujian juga dilakukan dengan menggunakan classifier lainnya, yaitu Random Forest, KNN, dan Support Vector Machine. Adapun pembagian data latih dan data uji tetap sama yaitu 20% digunakan sebagai data uji dan sisanya sebagai data latih. Running time dari masing – masing classifier juga dibandingkan, yaitu dengan memanggil library timeit.

ISSN : 2477-359X (Online)
ISSN: 2407-6007 (Print)

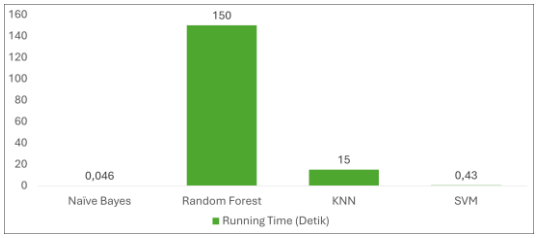
```
# In the beginning - import and default values
import timeit

scores = {}
times = {}

# ...
# Then, each algorithm will start with `start`
#   and end with assigning the `score` and `times`

start = timeit.default_timer()
```

Gambar 5. Library timeit

Hasil perbandingan diantara 4 algoritma classifier yang digunakan. Naïve Bayes mendapatkan nilai akurasi 77.78% Random Forest mendapatkan akurasi 70.1% KNN mendapatkan akurasi 24.88% dan SVM mendapatkan nilai akurasi 80.60%. Sedangkan running time masing – masing adalah Naïve Bayes 0.046 detik, Random Forest 150 detik, KNN 15 detik, dan SVM 0.43 detik.



Gambar 6. Hasil Perbandingan Akurasi



Gambar 7. Hasil Perbandingan Running Time

4. KESIMPULAN

Kesimpulan dari penelitian ini adalah, Naïve Bayes mendapatkan nilai akurasi 77.78% Random Forest mendapatkan akurasi 70.1% KNN mendapatkan akurasi 24.88% dan SVM mendapatkan nilai akurasi 80.60%. Sedangkan running time masing – masing adalah Naïve Bayes 0.046 detik, Random Forest 150 detik, KNN 15 detik, dan SVM 0.43 detik.

Adapun saran yang dapat dilakukan oleh peneliti selanjutnya adalah melakukan analisis terhadap nilai TPR/ROC selain analisis nilai akurasi, dan dari sisi metode seleksi fitur yang dipakai, bisa ditambahkan dengan metode lain kemudian dilakukan perbandingan.

5. REFERENSI

[1] Y. Ying, T. N. Mursitama, Shidarta and Lohansen, "Effectiveness of the News Text Classification Test Using the Naïve Bayes' Classification Text Mining Method," in *J. Phys.: Conf. Ser. 1764 012105*, 2021.

[2] E. D. Bintari, Gunawan and A. Indriani, "Classification of News on "Radar" Tarakan Online Using KNearest Neighbor Method with N-Gram Features," in *IOP Conf. Ser.: Mater. Sci. Eng. 676 012008*, 2019.

[3] J. P. Haumahu, S. D. H. Permana and Y. Yaddarabullah, "Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)," in *IOP Conf. Ser.: Mater. Sci. Eng. 1098 052081*, 2021.

[4] P. Yu, V. Y. Cui and J. Guan, "Text Classification by using Natural Language Processing," in *J. Phys.: Conf. Ser. 1802 042010*, 2021.

[5] I. M. Rabbimov and S. S. Kobilov, "Multi-Class Text Classifcation of Uzbek News Articles using Machine Learning," in *J. Phys.: Conf. Ser. 1546 012097*, 2020.

[6] X. Wu and J. Chen, "Text Classification on Large Scale Chinese News Corpus using Character-level Convolutional Neural Network," in *J. Phys.: Conf. Ser. 1693 012171*, 2020.

[7] J. Li and H. Zhu, "A Practical Application for Text-Based Sentiment Analysis Based on Bayes-LSTM Model," in *J. Phys.: Conf. Ser. 1631 012035*, 2020.

[8] D. D. Saputra, W. Gata, N. K. Wardhani, K. S. Parthama, H. Setiawan, S. Budilaksono, D. Yogatama, A. Hadiyatna, E. P. Purnamasari, B. Pratama and D. Novianti, "Optimization Sentiments of Analysis from Tweets in myXLCare using Naïve Bayes Algorithm and Synthetic Minority Over Sampling Technique Method," in *J. Phys.: Conf. Ser. 1471 012014*, 2020.

[9] Syahriani, A. A. Yana and T. Santoso, "Sentiment Analysis Of Facebook Comments On Indonesian Presidential Candidates Using The Naive Bayes Method," in *J. Phys.: Conf. Ser. 1641 012012*, 2020.