BAB I
PENDAHULUAN

1.1 Latar Belakang

Berita (*news)* merupakan sebuah sajian utama dari media massa di samping views (opini) (Asep, 2018). Berdasarkan data dari perusahaan konsultan Maverick Indonesia yang melibatkan 453 responden di kawasan Jakarta dan Bandung menghasilkan beberapa klasifikasi berita yaitu pembaca dengan rentang usia 18 hingga 24 tahun tertarik pada topik bahasan edukasi yaitu sebesar 35% sedangkan rentang usia 25 hingga 34 tahun topik bahasan tertinggi adalah kesehatan (31%), keduanya memiliki topik bahasan terendah yaitu gaya hidup 4% dan 7% (Kasih, 2020). Tahun 2006 perkembangan dan pertukaran informasi telah berada di titik lebih dari 550 triliun dokumen dan 7,3 juta halaman baru pada internet setiap harinya. Sementara itu pengklasifikasian kategori pada berita masih dilakukan secara manual sehingga memerlukan waktu yang lama dan secara penataan bahasa masih sulit untuk dikategorikan (Ariadi & Fithriasari, 2015). Umumnya pada portal berita selalu dikelompokkan ke dalam beberapa kategori tertentu (Usmani & Shamsi, 2020). Namun, pengelompokan tersebut masih dilakukan secara manual, artinya pengelompokan tersebut perlu membaca isi berita secara keseluruhan agar pengelompokan dapat dilakukan dengan optimal. Hal tersebut dinilai kurang efektif, terlebih jika data berjumlah banyak.

Penelitian oleh Findra Kartika dan Tri Purnomo yang mengklasifikasi Berita dengan Metode Multinomial Naïve Bayes menggunakan Pembobotan kata TF-IDF dengan 7 kategori dan 10.500 berita menghasilkan akurasi, presisi, recall dan f1-score sebesar 96%, peneliti menyebutkan bahwa sampel sangat sedikit dibandingkan dengan jumlah dataset dan perlu model klasifikasi dan pengujian terbaru. (Kartika & Purnomo, 2021). Kemudian Dio Ariadi dan Kartika Fithriasari dalam penelitiannya menjelaskan klasifikasi dengan metode Naïve Bayes Multinomial dengan menambahkan confix stripping stemmer dengan 1200 berita dan 12 kategori berita menghasilkan performa akurasi, presisi dan recall sebesar

82,2%, 83,9%, dan 82,2% tetapi berbeda saat menggunakan Support Vector Machine yaitu 88,1%, 89,1%, dan 88,1%, pada penelitian ini tidak dilakukan pemilihan atribut sehingga perlu pengurangan dataset (Ariadi & Fithriasari, 2015). Penelitian yang lain oleh Sabrani dkk guna mengklasifikasi 1000 artikel online tentang gempa di Indonesia dengan 6 kategori berbasis Naïve Bayes Multinomial menghasilkan tingkat akurasi *f-measure* sebesar 95.20% dengan 5 kali perulangan *Fold Cross Validation*, namun *stopwords removal* yang diterapkan mengalami penurunan yang signifikan pada pengujian fitur bigramnya. (Sabrani et al., 2020). Dari rangkaian penelitian tersebut, metode Naïve Bayes cukup baik digunakan dalam menghitung nilai akurasi dan dapat digunakan untuk mengklasifikasi pada penelitian ini.

Penelitian yang menggunakan metode *Naïve Bayes* Multinomial terhadap hasil klasifikasi berita masih terbatas, beberapa diantaranya pernah dilakukan oleh peneliti terdahulu. Pada penelitian (Kartika & Purnomo, 2021) dan (Ariadi & Fithriasari, 2015) menjelaskan bahwa nilai akurasi, presisi, recall, dan f1-score menambahkan model pipeline dan SVM berdasarkan studi kasus masing-masing, sehingga memerlukan model yang baru. Oleh karena itu, penelitian ini mengangkat studi kasus pemberitaan menggunakan metode Naïve Bayes Multinomial untuk mendapatkan akurasi dan klasifikasi terbaik dengan 19.200 data dan 14 kategori pemberitaan. Untuk itu, penelitian ini dibuat dengan judul "ANALISIS DATA MINING DENGAN METODE NAÏVE BAYES MULTINOMIAL TERHADAP KLASIFIKASI JUDUL PEMBERITAAN".

1.2 Rumusan Masalah

Sesuai latar belakang, rumusan masalah yang akan dibahas adalah berapakah nilai akurasi, presisi dan recall yang diperoleh metode Multinomial Naïve Bayes dalam mengklasifikasi media pemberitaan?

1.3 Tujuan

Tujuan penelitian berdasarkan rumusan masalah diatas adalah:

1. Mengukur kinerja klasifikasi berupa nilai akurasi, presisi dan recall dengan metode Multinomial Naïve Bayes pada pemberitaan indozone.id
2. Mengetahui pengaruh banyak data terhadap proses klasifikasi.

1.4 Manfaat

Adapun manfaat dari penelitian ini, yaitu:

1. Bagi masyarakat umum, penelitian ini dapat mempermudah dalam menentukan topik berita yang akan diterbitkan.
2. Diharapkan menambahkan variasi dalam penelitian dan teknik teks mining pada media pemberitaan.
3. Hasil penelitian dapat menjadi referensi guna mengukur nilai performa akurasi, presisi, recall dalam pengelompokan dokumen yang menerapkan metode *Naïve Bayes Multinomial* dengan ekstraksi fitur TF-IDF.

1.5 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data yang diambil merupakan data berita dari situs website indozone.id
2. Data terdiri dari 19.200 berita dan 14 kategori berita.
3. Label data yang digunakan adalah Fakta dan Mitos, Film, Game, Kecantikan, Kehidupan, Kesehatan, Kuliner, Musik, News, Olahraga, Otomotif, Selen, Teknologi, Travel.
4. Bahasa yang digunakan adalah bahasa Indonesia.
5. Hasil akhir pengujian berupa nilai akurasi pada metode *Naïve Bayes Multinomial*.
6. Skenario uji menggunakan *K-Fold Cross Validation*
7. Metode yang digunakan *Naïve Bayes Multinomial*.
8. Tools pendukung yang digunakan adalah *Google Collabs*


UNIVERSITAS MUHAMMADIYAH
JEMBER