

BAB 2 METODE PENELITIAN

2.1 Penelitian Terdahulu

Berdasarkan penelitian terdahulu dapat digunakan sebagai acuan dalam mengerjakan penelitian ini.

Tabel 1. penelitian terdahulu

NO	JUDUL PENELITIAN	PENULIS DAN TAHUN	METODE	HASIL AKURASI
1	Deteksi Komentar <i>Cyberbullying</i> Pada Media Sosial Berbahasa Inggris Menggunakan <i>Naïve Bayes Classification</i>	Jasman Pardede, dkk, 2020	<i>Naïve Bayes Classification</i>	Pada penelitian ini menggunakan metode <i>Naïve Bayes</i> dengan menunjukkan hasil akurasi sebesar 80 %, hasil <i>precision</i> sebesar 81% dan hasil <i>recall</i> sebesar 80%.
2	Analisis sentimen maskapai penerbangan menggunakan metode <i>naivebayes</i> dan seleksi fitur informasi gain	Arif Bijaksana Putra Negara dkk, 2021	<i>Naïve Bayes</i> dan Informasi gain	Pada penelitian ini menggunakan metode <i>Naïve Bayes</i> dengan menunjukkan hasil akurasi 81%%. Sedangkan untuk penambahan informasi gain menunjukkan hasil akurasi sebesar 86%.
3	Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia Dengan Metode <i>Naive Bayes Classifier</i>	Fitria Septianingrum dkk, 2021	<i>Naïve Bayes Classifier</i>	Pada penelitian ini menggunakan metode <i>Naïve Bayes</i> dengan menunjukkan hasil 78%.

2.2 Cyberbullying

Cyberbullying merupakan sebuah perilaku perundungan dalam dunia maya, *cyberbullying* dapat juga disebut bentuk intimidasi pelaku yang dilakukan untuk menyakiti korban secara berulang-ulang dengan mengirim dan mengupload sesuatu hal yang dapat berdampak pada perkembangan teknologi terutama pada media social. Pelaku *cyberbullying* melakukan kejahatan tersebut bertujuan untuk mendapatkan kepuasan pribadi, adapun yang dirasakan pada korban berdampak pada kesehatan mental, hilangnya kepercayaan diri dan juga menghambat produktivitas korban.

2.3 Analisis Sentimen

Sentiment analysis atau analisis sentimen dalam bahasa Indonesia adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan mengelompokkan teks ke dalam dokumen untuk menentukan kalimat yang dapat dikategorikan negatif dan positif.

2.4 Preprocessing

Preprocessing merupakan tahap awal yang mempersiapkan data sebelum digunakan dengan bertujuan mendapatkan proses data yang terstruktur dan jelas sehingga menghasilkan informasi dengan kualitas yang baik dan siap digunakan. Pada tahap ini meliputi *case folding*, *tokenizing*, *filtering*, *stopword* dan *stemming*.

a. Case Folding

Pada tahapan ini *case folding* bertugas untuk mengubah semua huruf kapital pada dokumen menjadi huruf kecil.

b. Tokenizing

Pada tahap ini *tokenizing* bertugas untuk memisah kalimat yang terdapat pada dokumen untuk menjadi kata dengan melakukan penghapusan tanda baca, simbol ataupun karakter lainnya yang dapat mempengaruhi proses.

c. Stopword Removal

Pada tahap ini *stopword removal* bertugas untuk menghilangkan kata yang tidak penting pada dokumen, seperti kata sambung, kata ganti dan lain-lainya.

d. Stemming

Pada tahap ini menggunakan *library preprocessing* sastrawi untuk mengubah kata imbuhan menjadi kata dasar. Selanjutnya dilakukan proses untuk menghilangkan term yang mengandung sisipan, akhiran dan awalan [5].

2.5 Pembobotan Kata

Setelah dilakukannya tahapan *preprocessing* dapat dilanjutkan dengan tahapan TF-IDF, yang digunakan untuk pemrosesan pembobotan kata. Oleh karena itu, semakin baik kata yang didapatkan maka semakin penting kata tersebut. Berikut rumus yang digunakan dalam tahapan TF-IDF :

a. TF (*Term Frequency*)

TF (*Term frequency*) adalah metode pembobotan term. *Term frequency* menyatakan jumlah berapa banyak keberadaan suatu term dalam satu dokumen[6]. Berikut persamaan pembobotan TF :

$$Tf_{ef} = \frac{f d (e)}{\max f d (f)} \quad (1)$$

Keterangan :

$f d (e)$ = Frekuensi term e pada dokumen f

$\max f d (f)$ = Hasil keseluruhan pada dokumen f

b. IDF (*Invers Document Frequency*)

IDF (*Invers Document Frequency*) berfungsi untuk mengurangi suatu term atau dokumen yang bermunculan secara banyak. Berikut persamaan pembobotan IDF :

$$Idf (c,d) = \log \left(\frac{N}{df(c) + 1} \right) \quad (2)$$

Keterangan :

N = Nilai keseluruhan dokumen

$df(c) + 1$ = Total dokumen yang memuat c

2.6 Gain Ratio

Setelah dilakukan tahap preprocessing maka selanjutnya di lakukan seleksi fitur *gain ratio*. *Gain ratio* merupakan teknik pembobotan atribut yang pada umumnya digunakan pada metode *Decision Trees* [7]. Seleksi fitur ini berpengaruh besar pada proses prediksi dan fitur ini juga menyeleksi berdasarkan rangking nilai *gain rasio*. Maka semakin besar nilai yang dimiliki *gain ratio* dapat menjadi pengaruh besar terhadap proses presiksi [8]. Dalam

pembobotan *gain ratio* digunakan untuk mengoptimalkan hasil akurasi pada metode *naïve bayes*. Selanjutnya *gain rasio* mendapatkan beberapa persamaan sebagai berikut :

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \quad (3)$$

$$E(A) = \sum_{i=1}^v \frac{p_i+n_i}{p+n} I(p_i, n_i) \quad (4)$$

$$\text{Gain}(A) = I(p, n) - E(A) \quad (5)$$

Dimana $I(p, n)$ adalah nilai split info dengan nilai p menyatakan kelas positif dan nilai n menyatakan banyaknya jumlah kelas negatif, $E(A)$ adalah nilai *entropy* atribut, $\text{Gain}(A)$ adalah nilai gain dari atribut, $I(p_i, n_i)$ adalah nilai split info dari atribut.

$$IV(A) = -\sum_{i=1}^v \frac{p_i+n_i}{p+n} \log_2 \frac{p_i+n_i}{p+n} \quad (6)$$

$$\text{Gain rasio}(A) = \text{gain}(A) / IV(A) \quad (7)$$

Hasil dari menentukan nilai bobot dari setiap atribut, maka diperoleh nilai dari perhitungan bobot dari setiap atribut :

$$w_i = \frac{\text{Gain rasio}(A)}{\frac{1}{a} \sum_{i=1}^v \text{Gain rasio}(A)} \quad (8)$$

Selanjutnya dilakukannya proses perbandingan probabilitas kelas pertama lebih besar dari probabilitas kelas lainnya, jika kelas tersebut lebih besar dari kelas lainnya maka termasuk dalam kelas pertama. Berikut persamaan dari perbandingan probabilitas kelas pertama dan kelas lainnya :

$$P(Y|X) = P(Y) \prod_{i=1}^a P(Y|X) w_i \quad (10)$$

$P(Y|X)$ = menunjukkan hasil klasifikasi dari persamaan *gain rasio*.

2.7 Naïve Bayes

Naïve bayes adalah merupakan salah satu metode pembelajaran mesin yang memanfaatkan perhitungan probabilitas dan statistic yang dikemukakan oleh ilmuan (Inggris Thomas Bayes)[9]. Konsep dasar yang digunakan pada *Naïve bayes* adalah *Teorema Bayes*, dalam penggunaannya dalam statistika untuk menghitung suatu peluang, proses pengelompokan atau klasifikasinya dibagi menjadi dua fase yaitu *learning/training* dan *testing/classify*[10]. Metode dipergunakan dapat mencapai probabilitas dan dapat diformulasikan menjadi persamaan :

$$P(c | Y) = \frac{P(Y|C)P(c)}{P(Y)} \quad (11)$$

2.8 Evaluasi Pengujian

Pada tahap ini dilakukannya pengujian metode dengan melakukan pengukuran pada *accuracy*, *precision*, dan *recall*. Dalam pengujian tersebut dilakukannya 2 pembagian yaitu metode *Naive bayes* dengan pembobotan *gain ratio* dan metode *Naive bayes*, dari data latih tersebut akan dilakukan split data dan dilakukan pengujian dengan *confusion matrix*.

