Text Mining: Sistem Prediksi *Cyberbullying* pada Platform Twitter Menggunakan Perbandingan *Logistic Regression, KNN*, dan *Naive Bayes*

Aditya Erlangga Wibowo[1], Alpian Khairi[2], Hildiana Humairoh[1], M Irvin Fadillah[1], M. Jordy Dwi Hartawan[1]
[1]Teknik Elektro, Fakultas Teknik, [2]Teknik Informatika, Fakultas Ilmu Komputer,
Universitas Sriwijaya
Palembang, Indonesia
Penulis korespondensi: adityaerlangga2003@gmail.com

Abstrak— Kemajuan pesat teknologi dan sosial media bisa memudahkan orang-orang untuk berkomunikasi dan juga memberikan informasi. Namun sosial media dapat membagikan efek negatif melalui cara membuat ketikan negatif atau komentar yang semaunya yang bertujuan untuk merendahkan bahkan menjatuhkan seseorang tanpa melihat perasaan orang tersebut. Hal itulah yang membuat terjadinya aktivitas kekerasan di dalam ruang siber (Cyberbullying). Beranjak dari permasalahan tersebut tentunya diperlukan sebuah sistem yang dapat mengklasifikasi komentar jahat atau bullying di media sosial Twitter. Dengan menggunakan tiga algoritma yaitu Logistic Regression, K-Nearest Neighbor (KNN), dan Naive Bayes classification, maka kita dapat menentukan apakah komentar tersebut mengandung makna bullying atau tidak. Jumlah data yang digunakan sebanyak 10535 data, dimana 6114 terdefinisi sebagai teks non-bullying dan 4421 data terdefinisi mengandung makna bullying. Prosedurnya adalah dengan mengurangi peluang dari setiap kata baru berdasarkan class dan perkalian class conditional probability. Dari hasil pengujian memakai dataset "komentar cyberbullying" yang diperoleh dari Algoritma Data Science School. Hasil terbaik dari ketiga algoritma tersebut diperoleh dengan menggunakan metode Naive Bayes, dengan accuracy sebesar 80,73%, precision 77,55%, dan recall sebesar 85,07%.

Kata Kunci— Cyberbullying, Twitter, Dataset, Logistic Regression, KNN, Naive Bayes.

Abstract— Rapid advances in technology and social media can make it easier for people to communicate and also provide information. However, social media can share negative effects through making negative typing or comments that aim to demean or even bring someone down without seeing that person's feelings. This is what makes violent activity in cyberspace (Cyberbullying). Moving on from these problems, of course, we need a system that can classify malicious comments or bullying on Twitter social media. By using three algorithms, namely Logistic Regression, K-Nearest Neighbor (KNN), and Naive Bayes classification, we can determine whether the comment contains the meaning of bullying or not. The amount of data used is 10535 data, of which 6114 are defined as non-bullying texts and 4421 data are defined as bullying. The procedure is to reduce the probability of each new word based on the class and the class conditional probability multiplication. From the test results using the "cyberbullying comments" dataset obtained from the Data Science School Algorithm. The best results from the three algorithms were obtained using the Naive Bayes method, with an accuracy of 80.73%, precision 77.55%, and recall of 85.07%.

Keywords— Cyberbullying, Twitter, Dataset, Logistic Regression, KNN, Naive Bayes.

I. PENDAHULUAN

Media sosial saat ini banyak digunakan baik oleh orang tua maupun anak-anak. Dengan bantuan teknologi internet, masyarakat umum dapat menemukan dan menerima informasi dengan lebih mudah. Media sosial menjadi alat yang mempermudah adanya komunikasi. Salah satu media sosial yang banyak dipergunakan terkhususnya di negara Indonesia adalah Twitter. Pengguna Twitter di Indonesia berjumlah 18,45 juta pada tahun 2022. Jumlah tersebut berarti 4,23% dari 436 juta pengguna aktif Twitter di seluruh dunia.[1] Namun tidak hanya memiliki dampak positif, dalam media Twitter ini juga mengandung dampak negatif, seperti adanya gambar atau tulisan kasar yang dapat diunggah oleh seseorang kepada orang lain (korban) dengan berbagai maksud dan tujuan yang tidak baik [2]. Terkadang banyak orang menjadikan Twitter sebagai akun alter ego (konteks alter ego berbentuk akun dengan perilaku tidak baik) kepada seseorang dengan maksud seperti mencemarkan nama baik korban dan menakut-nakuti korban sehingga korban dapat merasa malu dan tersakiti secara tidak langsung. Bentuk tindakan tersebut merupakan salah satu bentuk bullying online.

Cyberbullying adalah bentuk perilaku tidak menyenangkan yang dapat dilakukan seseorang dalam dunia maya. Cyberbullying dilakukan dengan sengaja oleh pelaku terhadap korban dengan mencapai maksud dan tujuan pelaku. Serangan yang dilakukan pelaku terus-menerus dapat dilayangkan kepada korban karena korban tidak dapat melindungi diri mereka dalam konteks elektronik (seperti email, blog, obrolan instan, dan pesan teks) karena tidak adanya larangan untuk meninggalkan komentar di kolom komentar. Korban dari tindak Cyberbullying akan dapat merasa tersakiti akibat perbuatan pelaku secara mental. Dari permasalahan ini, dilakukan penelitian agar dapat membangun sebuah sistem yang dapat mencegah dan memprediksi adanya tindakan Cyberbullying

di Twitter. Pendekatan supervised Machine Learning (ML) digunakan untuk membandingkan ketiga metode Logistic Regression, KNN, dan Naive Bayes yang dapat digunakan untuk membantu pengklasifikasian teks atau komentar tidak menyenangkan di media sosial dengan berdasarkan data yang diambil berupa tweet bullying di Twitter.

Media sosial memiliki efek negatif, seperti memposting tulisan kejam atau mengunggah foto terkait individu lain dengan tujuan menakut-nakuti dan merusak citra baik korban sedemikian rupa sehingga korban merasa terluka dan malu. Penelitian sebelumnya membahas mengenai komentar bullying pada platform instagram dan tidak membandingkan algoritma lainnya sedangkan dalam penelitian ini dilakukan pada platform twitter yang notabene komentarnya lebih frontal dan dilakukan perbandingan dengan menggunakan tiga algoritma model [3]. Selanjutnya, [4] membahas mengenai komentar bullying pada platform facebook dan tidak membandingkan algoritma lainnya sedangkan dalam penelitian ini dilakukan pada platform twitter yang notabene komentarnya lebih frontal dan dilakukan perbandingan dengan menggunakan tiga algoritma model.

Penelitian ini dilakukan bertujuan untuk melakukan perbandingan tingkat akurasi tiap metode untuk menemukan kata atau gabungan kata yang paling berpotensi digunakan untuk melakukan Cyberbullying pada media sosial khususnya Twitter menggunakan perbandingan dari metode Logistic Regression, KNN, dan Naive Bayes. Kemudian dapat dilihat tingkat akurasi dari ketiga metode tersebut sehingga mempermudah pengembang untuk membuat sistem atau perangkat lunak yang dapat melakukan deteksi dini terhadap terjadinya Cyberbullying di media sosial.

II. KAJIAN LITERATUR

A. Cyberbullying

Kejahatan *cyber bullying* merupakan peristiwa baru yang muncul ketika internet berkembang di dunia khususnya di Indonesia. Sebenarnya, ada berbagai bentuk kejahatan *cyber bullying* yang terjadi di Indonesia. Dengan demikian, kejahatan *cyber bullying* dapat dibagi menjadi empat bentuk. Pertama, pertengkaran dilakukan secara online. Biasanya, menggunakan kalimat yang mengandung kemarahan dan kebencian. Hal ini sangat sering terjadi di media sosial. Akibatnya, pertengkaran yang dilakukan secara online adalah hal biasa. Kedua, tindak pidana pelecehan yang dilakukan di media sosial. penggunaan dari kata-kata kasar, dan kasar adalah ciri dari kejahatan ini. Ketiga, pencemaran nama baik merupakan kejahatan *cyber bullying* yang dapat dilakukan oleh mengunggah data, memberikan komentar negatif, menyebarkan gosip, dan rumor tentang sesuatu yang buruk dari seseorang untuk memberikan kesan negatif untuk orang itu [4]. Keempat, tindakan mengucilkan seseorang dari media massa. Sebagai demikian, peristiwa tersebut sering terjadi pada kelompok sosial pada umumnya.

B. Logistic Regression

Logistic Regression atau bisa juga disebut Distribution Model adalah model prediksi pendekatan yang sama halnya seperti regresi linear atau sering disebut OLS atau Ordinary Least Square regression [5]. Perbedaannya berada pada variabel yang diteliti berskala dikotomi, regresi logistik, skala data numerik dengan kedua kategori, misalnya Benar atau Salah, Hangat atau Dingin, dan Terindikasi atau Tidak Terindikasi.

$$log\frac{p(x)}{1-p(x)} = a_0 + a.x \tag{1}$$

C. KNN Model

K-Nearest Neighbor (KNN) adalah algoritma yang mengambil hasil query baru untuk diklasifikasikan. KNN berfungsi untuk mencari jarak data estimasi terpendek antara nilai k-neighbor dari data latih [6]. Tujuan model ini untuk mengklasifikasikan objek baru yang diambil dari atribut dan pola latihan. Hasil uji sampel diambil dan diklasifikasikan dari kategori KNN [7].

$$P(X = x) = \frac{1}{k}\sum_{i \in A} I(y^{(i)} = j) \tag{2}$$

D. Naive Bayes

Naive Bayes adalah algoritma klasifikasi yang didasari oleh Bayes' Theorem of Probability. Faktanya, teorema Bayes sering digunakan pada kehidupan sehari-hari. Naive Bayes juga merupakan prosedur pemecahan pembelajaran yang memakai rule Bayes dan dugaan kuat bahwa aksesoris atau atribut dari kelasnya merupakan bentuk independen bersyarat. Meskipun dugaan independensi yang diuji seringkali diabaikan dalam praktiknya, model Bayes ini masih membagikan tingkat keakuratan klasifikasi yang bersaing. Digabungkan dengan keefisienan komputasi dan banyak properti yang lebih gemar, ini membuat Naive Bayes tersebar luas [8].

Teorema Bayes memiliki cara perhitungan kemungkinan posterior P (c| x) untuk P (c), P (x), juga P (x|c). Pada klasifikasi Naive Bayes, model menganggap efek hanya dari nilai prediktor (x) kelas tertentu (c) tanpa bergantung pada value dari prediktor yang lain [7]. Bentuk persamaan algoritma Naive Bayes dapat dituliskan seperti yang ada di bawah ini.

(3)

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)}$$

Keterangan:

P(c|x) merupakan kemungkinan posterior kelas (target) pada prediktor.

P(c) merupakan kemungkinan kelas yang sebelumnya.

P(x|c) merupakan kemungkinan peluang kelas yang diberikan oleh prediktornya.

P(x) merupakan kemungkinan dari prediksi sebelumnya.

III. METODOLOGI PENELITIAN

A. *Metode Penelitian*

Metode yang digunakan merupakan metode empiris. Dimana percobaan yang dilakukan berdasarkan pengamatan terhadap peristiwa atau kejadian nyata yang dialami. suDalam kasus ini melakukan percobaan dan pengamatan dengan didukung beberapa sumber artikel yang relevan antara lain: ScienceDirect, SpringerLink, IEEE Xplore, Jurnal mengenai *Machine Learning* dan beberapa jurnal terbitan dari JSS (*Journal Smart System*). Artikel yang dipilih memenuhi kriteria inklusi, memiliki sifat empiris terhadap gambaran prediksi Cyberbullying.

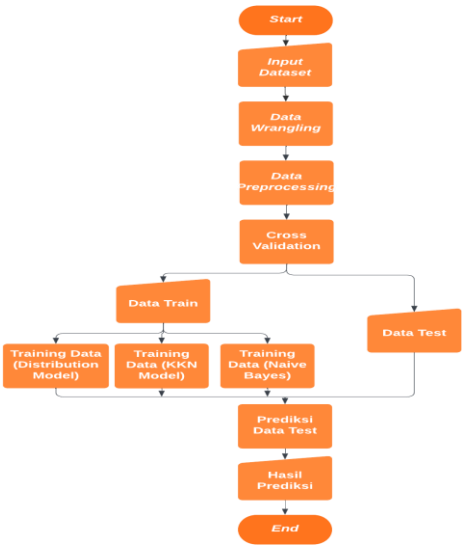
B. *Dataset*

Dataset yang digunakan dalam penelitian ini adalah dataset Data Science School Algorithm. Ini memiliki dua atribut fitur dengan deskripsi dan nilai ditampilkan pada Tabel 1. Dataset ini digunakan sebagai data training untuk bahasa pemrograman Python dan menggunakan confusion matrix untuk memvisualisasikan hasil pengklasifikasian objek berupa prediksi bahwa sebuah tweet memiliki referensi negatif menggunakan pendekatan tiga model, yaitu Distribution Model, KNN Model, dan Naive Bayes. Dilakukan pengujian keakuratan sistem prediksi hasil yang dibuat dengan menghitung persentase benar atau salah.

TABLE I. FITUR DATASET

No	Atribut	Deskripsi	Nilai
1	***bully***	Tweet terindikasi bullying atau tidak	0 = Tidak terindikasi 1 = Terindikasi
2	***tweet***	Tweet atau text yang dikirim oleh pengguna twitter	teks

C. *Alur Prosedur Sistem*



Gambar 1. Alur diagram memprediksi *cyberbullying*.

Pada penelitian ini, dirancang skema sistem prediksi yang terlihat pada Gambar 1 yang memiliki tujuan untuk mendapatkan suatu sistem prediksi bullying [9], alhasil bisa bermanfaat untuk memprediksi adanya kata yang berindikasi makna bully. Adapun tahapan untuk mencari model yang paling baik untuk sistem prediksi text bullying ini sebagai berikut :

- Pertama data akan dibaca sebagai data tabel menggunakan *tools* bahasa pemrograman R
- Selanjutnya pada tahapan *data wrangling* data akan dimodifikasi bentuknya kedalam bentuk seharusnya agar data dapat diproses dengan lebih mudah.
- Selanjutnya pada data *preprocessing* data yang masih mentah akan dilakukan pengolahan lebih lanjut agar data dapat diproses dengan baik. Data akan diubah kedalam format yang sesuai kemudian dilakukan pembersihan dan penyaringan agar informasi yang ada pada data akan lebih mudah ditafsirkan.
- Tahapan selanjutnya adalah *cross validation,* pada tahapan ini data akan dibagi menjadi 2 bagian yakni 80% data untuk proses *training* dan 20% data untuk memvalidasi kebaikan model. Hal ini diperlukan untuk melihat seberapa baik model *machine learning* yang telah dibuat dalam mengklasifikasikan data, sehingga jika model masih buruk dapat dilakukan tindakan lanjutan untuk memperbaiki atau mencari model yang lebih baik.
- Selanjutnya setelah data telah dilakukan pengolahan kedalam bentuk yang sesuai, maka tahapan selanjutnya adalah *modelling.* Pada tahapan ini *data train* akan digunakan untuk melatih model *machine learning* menggunakan algoritma *machine learning* tertentu
- Setelah model *machine learning* didapatkan, hal yang akan dilakukan selanjutnya adalah memprediksi *data test* yang ada dan kemudian dilakukan pengukuran kebaikan model menggunakan *confusion matrix* antara hasil prediksi dan data asli untuk melihat seberapa baik model yang telah dibuat. Jika model masih kurang baik maka tahapan selanjutnya adalah melakukan perbaikan model atau mencari metode *machine learning* lainnya sampai menemukan model paling baik.

IV. HASIL DAN PEMBAHASAN

Dilakukan eksperimen dengan 3 algoritma yaitu Distribution, KNN, dan Naive Bayes. Hal ini memiliki tujuan untuk melihat classifier mana yang paling cocok untuk mengklasifikasikan data sebelum diproses (preprocessing), melatih data, dan membuat prediksi data dengan menggunakan ketiga model yang dilatih dari proses pelatihan.

A. Data Preprocessing

Dalam tahap Data Preprocessing, dataset yang digunakan akan melewati dua tahapan penting agar semua data, baik data pengujian dan data testing memiliki jenis yang sama [10]. Adapun tahap-tahap tersebut antara lain:

1) *Text to Corpus,* tahapan memindahkan dataset ke dalam bentuk Korpus.
2) *Text Cleansing,* tahapan ini berguna untuk memastikan bahwa data tweet yang dilakukan dalam penelitian tidak mengandung karakter-karakter yang tidak dibutuhkan seperti spesial karakter (@,#,$) dan emoji. Dalam tahap *Text Cleansing*, memiliki beberapa tahapan. Hal yang pertama harus dilakukan yaitu:
    - *Case-folding*, memastikan semua text dalam bentuk *lowercase*.
    - Remove number, menghapus semua angka yang ada pada tweet.
    - *Remove stopwords*, memastikan semua kata-kata yang tidak penting akan diabaikan selama training model dan pengujian.
    - *Remove all punctuation*, menghapus semua karakter spesial seperti @, #, dan $. Setelah semua karakter spesial dihapus.
    - *Text Stemming*, mereduksi setiap kata menjadi sebuah kata inti saja.
    - *Correction Spelling,* melakukan koreksi ejaan kata-kata.
    - *Remove Emoticons and Emojis* menghapus emoticon dan emoji yang terdapat dalam tweet.
    - *Remove White Space*, menghapus tab, spasi, dan baris kosong yang tidak diperlukan

B. Pemisahan Data

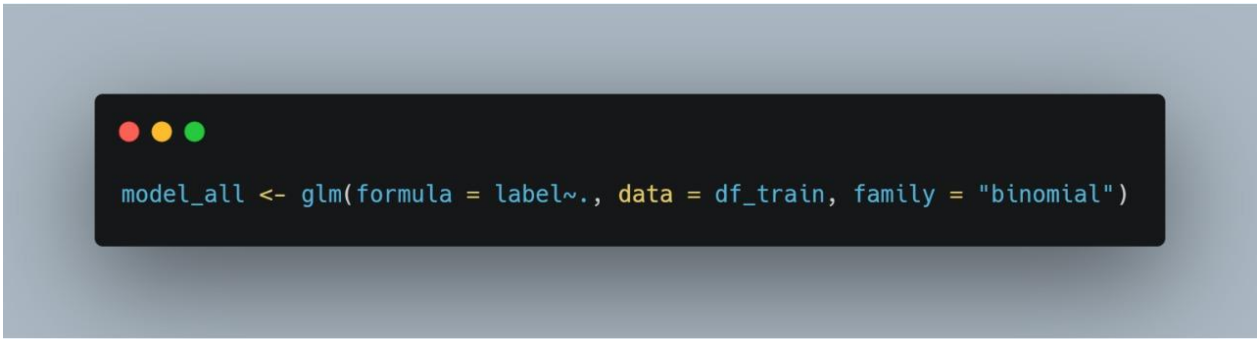
Dataset yang digunakan dipisah menjadi data latih dan data test yang ditampilkan pada Tabel 2.

TABLE II. PERBANDINGAN DATA LATIH DAN DATA TES

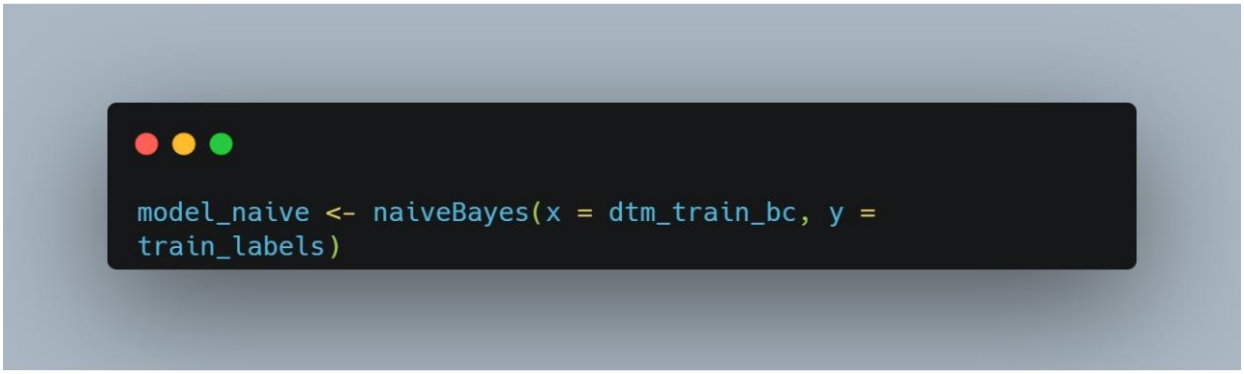
	Data Test	Data Latih	Total
Persentase	20%	80%	100%
Banyaknya	2107	8428	10535

C. Training Data dan Testing Data

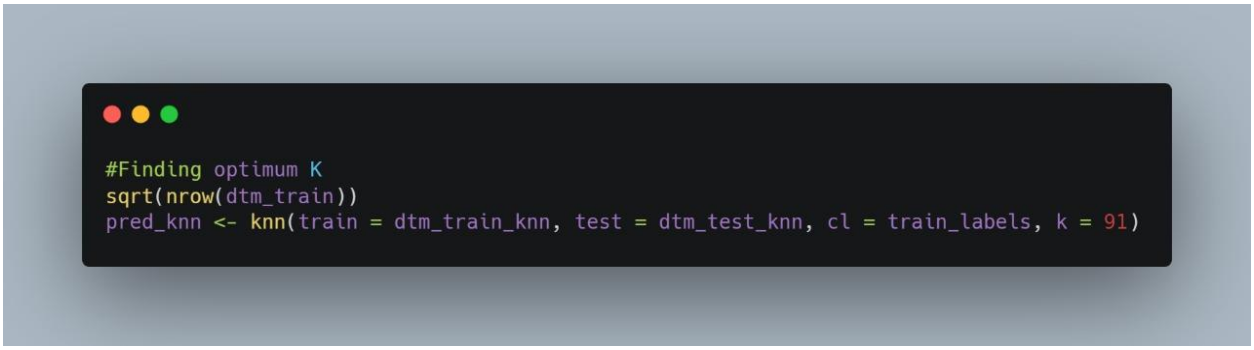
Data latih adalah fase data yang dipakai untuk memberi latihan kepada program yang kemudian dijalankan untuk membuat prediksi sehingga dapat menemukan hubungan antara data yang diberikan dengan data yang akan dites [7]. Data yang dipakai sebagai train data adalah 80 persen dari total data, yaitu 10535 data, sehingga data yang dipakai sebagai train data adalah 8428 data, dan sisanya digunakan untuk pengujian. Berikut adalah kode program data testing dan training data untuk Logistic Regression, Naive Bayes, dan KNN.



Gambar 2. Training data Logistic Regression Model



Gambar 3. Training data Naive Bayes Model



Gambar 4. Training data KNN Model

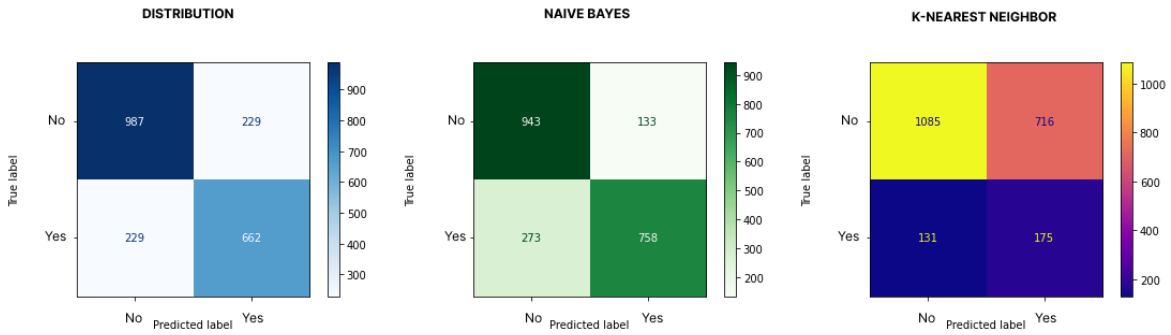
D. Hasil Prediksi dan Confusion Matrix

Ketiga model dilakukan pengujian prediksi terhadap data train sebelumnya menghasilkan prediksi yang berbeda-beda. Model yang paling banyak melakukan prediksi dengan benar adalah Naive Bayes dengan total prediksi benar sebesar 1701 dari 2107 total data test. Sedangkan model yang paling banyak melakukan prediksi yang salah adalah KNN dengan total prediksi salah sebesar 847 dari 2107 total data test. Berikut adalah visualisasi dari hasil prediksi ketiga model yang diuji.

$$Akurasi = \frac{Jumlah\ Data\ Prediksi\ Benar\ (True\ Positive + True\ Negative)}{Jumlah\ seluruh\ data} \quad (4)$$

$$Recall = \frac{True\ Positif}{(True\ Positive + False\ Negatif)} \quad (5)$$

$$Precision = \frac{True\ Positif}{(True\ Positive + False\ Positive)} \quad (6)$$



Gambar 5. True label dan predicted label

E. Tingkat Akurasi, Recall, Precision

Tingkat akurasi dari ketiga model didapatkan dengan menggunakan rumus total prediksi benar (true positive + true negative) dibagi dengan total seluruh data test yang digunakan. Berikut adalah tingkat akurasi ketiga model yang diuji.

TABLE III. PERBANDINGAN DISTRIBUTION MODEL, NAIVE BAYES, DAN K-NEAREST NEIGHBOR

	Distribution Model	**Naive Bayes**	**K-Nearest Neighbor**
Akurasi	78.26%	80,73%	59,80%
Recall atau Sensitivitas	74,30%	85,07%	19,64%
Precision	81,17%	77,55%	89,22%

Pada Tabel 3 model dengan akurasi tertinggi adalah model Naive Bayes dengan akurasi sebesar 80,73%, diikuti oleh model distribution sebesar 78,26% dan kemudian model dengan akurasi terendah adalah model K-Nearest Neighbor dengan akurasi sebesar 59,90%. Sedangkan model dengan tingkat recall tertinggi kembali didapatkan oleh model Naive Bayes dengan recall sebesar 85,07%, diikuti dengan model distribution dengan recall sebesar 74,30% dan model dengan recall terkecil adalah model K-Nearest Neighbor dengan recall sebesar 19,64%. Adapun untuk model dengan precision terbaik adalah model K-Nearest Neighbor dengan presisi sebesar 89,22%, urutan kedua adalah model distribution dengan nilai precision sebesar 81,17% dan model yang memiliki precision paling kecil adalah model Naive Bayes dengan precision sebesar 77,55%. Pada kasus ini ukuran kebaikan model yang paling diutamakan adalah nilai recall untuk meminimalisir nilai false negative, sehingga model yang terbaik adalah model Naive Bayes dengan nilai recall sebesar 85,07%.

V. KESIMPULAN

Pengujian klasifikasi data Cyberbullying dengan perbandingan ketiga klasifikasi tersebut dapat disimpulkan bahwa Distribution Model, Naive Bayes, dan KNN Model mempunyai level keakurasian yang berbeda-beda. Model Naive bayes dengan akurasi sebesar 80.73%. Distribution model dengan hasil akurasi klasifikasi sebesar 78.26%. Sedangkan KNN memiliki akurasi sebesar 59.80%. Dari ketiga data tersebut, model yang memiliki tingkat akurasi terbaik adalah Naive Bayes lalu diikuti oleh Distribution Model. Sedangkan KNN Model memiliki akurasi yang rendah dan jauh dari kedua model lainnya. Jumlah data yang beragam mempengaruhi kinerja dari tiga model yang digunakan, terutama performa dari model KNN yang lebih fokus terhadap data neighbor (tetangga) membuat performanya menjadi lebih kecil. Pada saat melakukan training dan test hasil menunjukan bahwa pengklasifikasian model Naive Bayes lebih cepat dibandingkan kedua model yang lain.

DAFTAR PUSTAKA

[1] R. Watrianthos, M. Giatman, W. Simatupang, R. Syafriyeti, and N. K. Daulay, “Analisis Sentimen Pembelajaran Campuran Menggunakan Twitter Data,” J. Media Inform. Budidarma, vol. 6, no. 1, p. 166, 2022, doi: 10.30865/mib.v6i1.3383.

[2] N. Istiani and A. Islamy, “Fikih Media Sosial Di Indonesia,” Asy Syar’Iyyah J. Ilmu Syari’Ah Dan Perbank. Islam, vol. 5, no. 2, pp. 202–225, 2020, doi: 10.32923/asy.v5i2.1586.

[3] R. M. Candra and A. Nanda Rozana, “Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor,” IT J. Res. Dev., vol. 5, no. 1, pp. 45–52, 2020, doi: 10.25299/itjrd.2020.vol5(1).4962.

[4] N. F. Hasan, “Deteksi Cyberbullying pada Facebook Menggunakan Algoritma K-Nearest Neighbor,” J. Smart Syst., vol. 1, no. 1, pp. 35–44, 2021, doi: 10.36728/jss.v1i1.1605.

[5] Y. Tai, “A Survey Of Regression Algorithms And Connections With Deep Learning,” pp. 1–12, 2021, [Online]. Available: http://arxiv.org/abs/2104.12647.

[6] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2888, no. August, pp. 986–996, 2003, doi: 10.1007/978-3-540-39964-3_62.

[7] J. Pardede, “Deteksi Komentar Cyberbullying Pada Media Sosial Berbahasa Inggris Menggunakan Naïve Bayes Classification,” J. Inform., vol. 7, no. 1, pp. 46–54, 2020, doi: 10.31311/ji.v7i1.6920.

[8] E. Azeraf, E. Monfrini, and W. Pieczynski, Using the Naive Bayes as a discriminative classifier. 2020.

[9] S. Shalev-Shwartz and S. Ben-David, Understanding machine learning: From theory to algorithms, vol. 9781107057. 2013.

[10] M. Rezwanul, A. Ali, and A. Rahman, “Sentiment Analysis on Twitter Data using KNN and SVM,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 6, pp. 19–25, 2017, doi: 10.14569/ijacsa.2017.080603.