Jurnal TICOM: Technology of Information and Communication
Volume 12, Nomor 1, September 2023 Hal: 14-19 E-ISSN: 2962-7982

Aplikasi Pendeteksi Kalimat Kasar Bahasa Indonesia Pada File Audio Menggunakan Jaccard Similarity Dan N-Gram

Muhammad Farras Majid[1*], Achmad Solichin[2]

[1*,2,3,4]Fakultas Teknologi Informasi, Teknik Informatika, Universitas Budi Luhur, Jakarta, Indonesia
Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan, DKI Jakarta, 12260
E-mail: [1*]farrasmajid10@gmail.com, [2] achmad.solichin@budiluhur.ac.id
(*: corresponding author)

Abstrak— **Di Indonesia, ujaran kebencian (*hate speech)* banyak sekali ditemukan di berbagai aplikasi media sosial. Bentuk ujaran kebencian dapat berupa tulisan, suara (audio), dan video. Salah satu ciri ujaran kebencian adalah keberadaan kata-kata kasar, baik yang terucap maupun tertulis. Selain berpotensi menimbulkan kebencian atau konflik, keberadaan kata kasar dapat menimbulkan dampak negatif bagi masyarakat, terutama anak-anak. Keterbukaan akses informasi bagi anak-anak melalui berbagai media sosial mengakibatkan dampak negatif jika anak-anak sering mendengar kata kasar, terutama dalam bentuk audio dan video. Hal tersebut dapat dianggap sebagai suatu kewajaran. Oleh karena itu, deteksi keberadaan kata kasar terutama pada media suara (audio) sangat penting untuk dilakukan. Pada penelitian ini, dikembangkan sebuah aplikasi yang dapat mendeteksi kata atau kalimat kasar dalam Bahasa Indonesia. Aplikasi tersebut dapat digunakan untuk memfilter konten-konten media sosial media. Pada penelitian ini digunakan metode *Jaccard Similarity* dan N-Gram untuk mendeteksi kata atau kalimat kasar pada sebuah *file* audio. Hasil penelitian menunjukkan bahwa penggunaan metode *Jaccard Similarity* dan N-Gram dapat diterapkan dengan baik untuk mendeteksi kata atau kalimat kasar dengan nilai akurasi sebesar 73,4% dan presisi sebesar 88,9%. Aplikasi yang dikembangkan dapat bermanfaat untuk masyarakat dalam mendeteksi dan menyaring kata atau kalimat kasar pada berbagai media, terutama media suara (audio).**

Kata Kunci— ***hate speech, jaccard similarity, n-gram***

Abstract— ***In Indonesia, hate speech can be found in various social media applications. Forms of hate speech can be in the form of writing, sound (audio), and video. One characteristic of hate speech is the existence of harsh words, both spoken and written. Apart from potentially causing hatred or conflict, the presence of harsh words can have a negative impact on society, especially children. Open access to information for children through various social media has a negative impact if children often hear harsh words, especially in the form of audio and video. This can be considered as a fairness. Therefore, detecting the presence of offensive words, especially in audio media, is very important to do. In this research, an application was developed that can detect harsh words or sentences in Indonesian. This application can be used to filter social media content. In this study, the Jaccard Similarity and N-Gram methods were used to detect harsh words or sentences in an audio file. The results showed that the use of the Jaccard Similarity and N-Gram methods can be applied well to detect harsh words or sentences with an accuracy value of 73.4% and a precision of 88.9%. The developed application can be useful for the community in detecting and filtering harsh words or sentences in various media, especially audio media.***

Keyword— ***hate speech, jaccard similarity, n-gram***

I. PENDAHULUAN

Ujaran kebencian (*hate speech*) adalah tindakan komunikasi yang dilakukan oleh suatu individu atau kelompok dalam bentuk provokasi, hasutan, ataupun hinaan kepada individu atau kelompok lain dalam hal berbagai aspek seperti ras, warna kulit, gender, cacat, orientasi seksual, kewarganegaraan, agama dan lain-lain. Ujaran kebencian yang dilayangkan kepada seseorang atau kelompok orang tertentu banyak mencuri perhatian akhir-akhir ini. Melalui postingan di media sosial dengan ujaran kebencian semakin marak diperbincangkan. Banyak pengguna internet (*netizen*) menyebarluaskan suatu postingan (gambar, foto, video, suara, dan kata-kata) dengan ujaran kebencian yang menimbulkan penghinaan, pencemaran nama baik, penistaan agama, dan lain sebagainya [1]–[3].

Dalam Bahasa Indonesia, kata kasar dapat diungkapkan salah satunya dengan menyebutkan jenis hewan tertentu, seperti anjing, monyet, dan sebagainya. Namun demikian, tidak semua kalimat yang memuat jenis hewan seperti contoh tersebut merupakan bentuk kalimat yang bersifat ofensif. Oleh karena itu, untuk mengidentifikasi apakah suatu kata dianggap sebagai kata yang kasar perlu melihat konteks kalimatnya secara menyeluruh [4].

Di Indonesia ujaran kebencian (*hate speech*) ini banyak sekali ditemukan di dalam sosial media khususnya konteks audio, dan ujaran kebencian yang dilakukan di dalam sosial media ini sudah seperti hal yang lumrah untuk dilakukan padahal ujaran kebencian tersebut dapat menimbulkan dampak yang sangat bahaya apalagi jika didengar dan diikuti oleh anak-anak di bawah umur. Dan jika budaya seperti ini terus berlanjut maka ke depannya Indonesia bisa memiliki generasi muda yang dapat dengan mudah menggunakan ujaran kebencian *(hate speech*), dampak yang bisa ditimbulkan dari *hate speech* antara lain adalah, dapat membuat perselisihan dan konflik, dapat

membuat orang lain yang tidak terima terkena gangguan mental, dan masih banyak lagi [1], [2], [5], [6].

Oleh karena itu, pada penelitian ini dikembangkan sebuah aplikasi berbasis *website* yang dapat mendeteksi kalimat kasar terutama pada sebuah *file* audio, agar pengguna internet di Indonesia bisa merasa aman dan nyaman saat menggunakan internet ini, tanpa perlu takut adanya hate speech.

Pada penelitian kali ini teknologi yang mendukung pengembangan aplikasi ini, yaitu *Natural Language Processing* (NLP). NLP merupakan cabang dari kecerdasan buatan yang berkaitan dengan kemampuan komputer untuk memahami, menganalisis, dan memanipulasi bahasa manusia secara alami. Tujuan utama dari NLP adalah memungkinkan komputer untuk berinteraksi dengan manusia dengan menggunakan bahasa manusia secara efektif[7].

Pada penelitian kali ini algoritma yang akan digunakan adalah *Jaccard similarity. Jaccard similarity* adalah suatu metode yang digunakan untuk mengukur kesamaan atau *similarity* antara dua set data. Metode ini berguna dalam berbagai bidang, seperti analisis teks, data *mining*, pengenalan pola, dan sistem rekomendasi. *Jaccard similarity* mengukur kesamaan berdasarkan elemen-elemen yang ada dalam kedua set, tanpa memperhatikan urutan elemen atau frekuensi masing-masing elemen[8].

Penelitian ini merujuk kepada penelitian dengan judul "Deteksi Berita Palsu Tentang Vaksinasi Covid-19 Dengan Menggunakan *Text Mining* Dan Algoritma *Cosine Similarity*" dengan hasil analisis terhadap data dengan menerapkan algoritma *text mining* dan *cosine similarity* diperoleh hasil yang menunjukkan berita yang diidentifikasi dinyatakan *hoax* sesuai dengan data yang dimiliki dengan presentasi berita *hoax* 100% berdasarkan perhitungan nilai probabilitas kemunculan berita. Berdasarkan hasil penelitian langkah cara yang diterapkan mendeteksi berita *hoax* dengan menerapkan algoritma *text mining* dan *cosine similarity* ini merupakan urutan langkah yang sesuai karena hasil *output* sesuai dengan data yang dikumpulkan. Aplikasi deteksi berita *hoax* yang dirancang dan dibangun dapat membantu pihak yang membutuhkan untuk mendeteksi berita *hoax* vaksinasi COVID-19.

Tujuan dari penulis membuat penelitian adalah untuk melihat hasil pendeteksian menggunakan algoritma *similarity* lain selain *cosine* yaitu *jaccard*. agar bisa membandingkan algoritma *similarity* yang lebih baik untuk mendeteksi sebuah *hate speech*.

Manfaat dari aplikasi yang dibuat oleh penulis kali ini adalah antara lain, membantu mencegah penyebaran konten yang tidak pantas dan mempromosikan komunikasi yang lebih sopan di berbagai platform digital, melindungi pengguna dari bahasa kasar atau pelecehan verbal dalam konteks audio yang diakses oleh masyarakat, mendukung keamanan dan privasi mengingat konten yang mengandung kata – kata kasar dapat merugikan individu atau organisasi, mempercepat proses moderasi konten dalam platform digital, mengurangi beban kerja manual yang dibutuhkan manusia untuk mendeteksi dan menghapus konten tidak pantas, memberikan *control* dan pilihan kepada pengguna internet untuk memfilter atau menghindari konten yang mengandung kata-kata kasar.

II. METODE PENELITIAN

A. Tahapan Penelitian

Tahapan penelitian meliputi pengumpulan data pengolahan data, *preprocessing data*, dan implementasi jaccard similarity dan N-Gram

1) Pengumpulan Data

Sebelum melakukan *preprocessing* Data pada penelitian ini, Langkah pertama dilakukan pengumpulan data terlebih dahulu. Pengumpulan data dilakukan dengan mencari data yang berhubungan dengan topik penelitian yaitu kalimat kasar Bahasa Indonesia. Tahap pengumpulan data ini tidak boleh dilewatkan karena jika tidak ada data yang terkumpul maka tahapan pengolahan data tidak dapat dilakukan. Proses pengumpulan data dilakukan dengan mencari kalimat kasar Bahasa Indonesia melalui mesin pencarian *google* dari *website* Kaggel.com. ini merupakan *website* untuk mencari sumber *dataset* yang sudah dikumpulkan.

2) Preprocessing Data

Preprocessing Data adalah merupakan tahapan lanjutan dari pengumpulan data. Pengolahan data dilakukan dengan menerapkan algoritma text mining sebagai *text processing.* Tahapan dari text mining yaitu *case folding, tokenizing, stemming.*

B. Data Penelitian

Data yang akan digunakan dalam penelitian ini adalah data yang berasal dari *website Kaggle*. Bentuk data ini adalah dataset kalimat kasar bahasa Indonesia. Data diperoleh secara manual dengan mengakses *website* https://www.kaggle.com/datasets/tsqfnfl/kalimat-kasar-bahasa-indonesia, dan keluaran data yang muncul pada *website* tersebut diambil secara manual dan dimasukkan ke dalam dokumen teks yang nantinya akan dilakukan aksi *preprocessing* data.

Data uji suara yang digunakan pada penelitian ini dibuat melalui *website SoundofText* dengan teknik *text-to-speech*, dalam kurun waktu Mei – Juni 2023, dan untuk datasetnya diambil pada bulan Mei 2023. Tabel 1 menyajikan *dataset* yang digunakan dalam penelitian ini.

TABEL I
DATASET

Nama	Kategori
Jual makanan anjing dog food happy dog murah harga promo	Halus
Jual grosir makanan anjing makanan kucing untuk petshop	Halus
Pagi ini Cuma mau panggil anjing aja buat elo yang naik motornya ugal ugalan	Kasar
males itu kalo kerja pagi trus gak ada yg nganter anjing	Kasar
cokelat merupakan racun bagi anjing dan kucing	Halus
anjing jg lo ler	Kasar
so sweet pemain ini gendong anjing pengganggu laga	Halus
kesel deh pagi anjing ku masa mau gigit perban kakiku	Kasar
lancau anjing stressnya aku	Kasar
nyepam bgt anjing	Kasar

TABEL II
DATA SUARA

Nama	Isi	Kategori
Halus 1.WAV	Farras memiliki hewan peliharaan yaitu anjing	Halus
Halus 2.WAV	Sesama anjing putih berantem	Halus
Halus 3.WAV	Mana yang lebih lucu anak anjing atau anak kucing	Halus
Halus 4.WAV	Anjing laut jalannya pake perut	Halus
Halus 5.WAV	Dikejar anjing pagi pagi	Halus
Kasar 1.WAV	Woy anjing tolol mukamu mirip kontolku	Kasar
Kasar 2.WAV	Nyawa lebih penting dari pada kaca anjing	Kasar
Kasar 3.WAV	Anjing bangsat tai doang lah	Kasar
Kasar 4.WAV	Sejenis anjing lah kau	Kasar
Kasar 5.WAV	Berisik anjing	Kasar

C. Data Dokumen

Data yang dikumpulkan adalah data yang berasal dari website Kaggle (https://www.kaggle.com/datasets/tsqfnfl/kalimat-kasar-bahasa-indonesia). Data yang diambil dari website Kaggle adalah data kalimat kasar bahasa Indonesia

Kalimat yang ada di dalam data tersebut berisi kumpulan kalimat kasar bahasa Indonesia. Dan data diperoleh secara manual yang nantinya akan dilakukan aksi pre-processing karena data yang diambil melalui Kaggle masih merupakan data mentah yang disimpan dalam format .csv

Data kedua yang dikumpulkan adalah data suara yang berasal dari *website soundoftext* dengan fitur *text-to-speech*. Data yang diambil dari *website soundoftext* adalah suara yang mengandung kalimat kasar dan tidak mengandung kalimat kasar bahasa Indonesia

D. Pre-Processing

Aksi *preprocessing* ini bertujuan untuk merapikan isi *dataset* sehingga menjadi lebih mudah untuk diproses dan dipanggil ke dalam aplikasi pendeteksi kalimat kasar. Pada tahapan ini akan dilakukan beberapa proses untuk merapikan *dataset*

Sebelum data diproses, data harus melalui tahap pra proses data. Sebab data masih berbentuk kalimat atau paragraf utuh. Proses komputasi tidak secara langsung dapat mengenali data berupa teks sehingga sangat sulit dilakukan tahapan pra proses yang dilakukan adalah:

1) *Case Folding*

Case folding merupakan suatu proses penyeragaman kata pada sebuah komentar, dalam hal ini huruf yang digunakan dalam kata adalah huruf kecil (*lowercase)*[9].

2) *Text Cleaning*

Text cleaning adalah proses untuk membersihkan *dataset* dari hal yang tidak diperlukan seperti tanda baca, normalisasi *Unicode*, dan sebagainya. Dalam melakukan proses *cleaning* tersebut dilakukan 4 tahapan untuk mendapatkan hasil yang maksimal, diantaranya ialah:

a. Menghapus Tanda baca
b. Menghapus angka
c. Menghapus kelebihan spasi

3) *Tokenizing*

Tokenizing adalah proses untuk pemotongan kata dalam suatu kalimat kedalam bentuk token, dimana tiap kata dalam suatu kalimat dipisahkan oleh spasi. [9]

4) *Stopword Removal*

Stopword removal adalah proses penghapusan kata yang tidak mempengaruhi dalam proses penghapusan kata yang tidak mempengaruhi dalam proses klasifikasi contoh: dan, ke, atau, dari, yang, dll.

5) *Stemming*

Stemming adalah proses seleksi kata yang memiliki kata sambung, kata imbuhan, kata ganti, dan kata kerja menjadi kata dasar, dengan menghapus awalan atau akhiran

E. N-Gram

N-gram adalah *substring* penggabungan karakter sejumlah K pada *text* dokumen. Dalam menentukan hasil deteksi kalimat kasar dengan menggunakan metode n-gram, *dimana* kalimat akan diproses dan akan dibentuk sebanyak n-gram atau memisahkan *string* sepanjang n yang akan dihitung pergeserannya secara terus menerus ke depan sejumlah n sampai akhir kalimat[4], [5], [8], [10]. Metode N-Gram sendiri cukup populer dan sering digunakan dalam mengolah data teks untuk berbagai keperluan, terutama untuk pencarian dan menemukan *similaritas* teks [11], [12].

F. Jaccard Similarity

Pada tahapan ini untuk mencari nilai *similarity* pada sebuah dokumen dengan rumus *jaccard similarity* dengan melihat irisan dan union antara dua dokumen. Beberapa penelitian yang menggunakan metode *Jaccard similarity* antara lain untuk mencari kemiripan abstrak tugas akhir [12], deteksi plagiarisme [11], dan pencarian artikel [13]. Persamaan 1 digunakan untuk menghitung *similaritas Jaccard* [8].

$$Similarity\ (X,Y) = \frac{|x \cap y|}{|x \cup y|} \times 100\% \quad (1)$$

III. HASIL DAN PEMBAHASAN

Pada bagian ini berisi analisis, hasil implementasi ataupun pengujian serta pembahasan dari topik penelitian, yang bisa dibuat terlebih dahulu metodologi penelitian. Bagian ini juga merepresentasikan penjelasan yang berupa penjelasan, gambar, tabel dan lainnya.

A. Tahapan Pre-Processing Data

1) *Text Cleaning*

Proses *cleaning* pada tahap ini bertujuan untuk membersihkan data kalimat kasar bahasa Indonesia dari tanda baca, dan sebagainya. Dalam melakukan proses *cleaning*

tersebut dilakukan 4 tahapan untuk mendapatkan hasil yang maksimal, diantaranya adalah:

TABEL III
TEXT CLEANING

Keterangan	Data Awal	Hasil *Text Cleaning*
Dataset	JANGAN MENTANG-MENTANG KAMU ANAK GAUL. PAS DIKEJAR ANJING BUKANNYA LARI MALAH BILANG ‘TERUS GUE HARUS KABUR SAMBIL BILANG WOW GITU?’	jangan mentang kamu anak gaul. Pas dikejar anjing bukannya lari malah bilang terus gue harus kabur sambil bilang wow gitu

2) *Tokenizing*

Pada tahap ini akan dilakukan pemisahan per kata pada sebuah kalimat yang akan ditampilkan pada tabel di bawah.

TABEL IV
TOKENIZING

Keterangan	Data setelah *text cleaning*	Hasil *Tokenizing*
Dataset	jangan mentang kamu anak gaul. Pas dikejar anjing bukannya lari malah bilang terus gue harus kabur sambil bilang wow gitu	(jangan) (mentang) (kamu) (anak) (gaul) (pas) (dikejar) (anjing) (bukannya) (lari) (malah) (bilang) (terus) (gue) (harus) (kabur) (sambal) (bilang) (wow) (gitu)

3) *Stopword Removal*

Pada tahap ini akan dilakukan penghapusan kata yang tidak penting, tahapan ini akan ditampilkan pada tabel di bawah.

TABEL V
\STOWORD REMOVAL

Keterangan	Data setelah *text cleaning*	Hasil *stopword removal*
Dataset	jangan mentang kamu anak gaul. Pas dikejar anjing bukannya lari malah bilang terus gue harus kabur sambil bilang wow gitu	Jangan mentang kamu anak gaul kejar anjing bukan lari malah bilang terus gue harus kabur sambil bilang wow

4) *Stemming*

Pada tahap ini akan dilakukan penghapusan kata sambung, kata imbuhan, kata ganti dan kata kerja menjadi kata dasar, dengan menghapus awalan atau akhiran, contoh akan ditampilkan pada tabel di bawah.

TABEL VI
STEMMING

Keterangan	Data setelah *text cleaning*	Hasil *stemming*
Dataset	jangan mentang kamu anak gaul. Pas dikejar anjing bukannya lari malah bilang terus gue harus kabur sambil bilang wow gitu	Jangan mentang kamu anak gaul kejar anjing bukan lari malah bilang terus saya kabur sambil bilang wow

B. N-Gram

Tujuan digunakan N-gram dikarenakan dalam bahasa Indonesia banyak frase yang tidak hanya terdiri dari satu kata. Dengan N-Gram kosakata menjadi lebih dapat terlihat nilainya, contohnya akan dijelaskan di bawah.

TABEL VII.
N-GRAM

Contoh Kalimat	N-Gram	Hasil Ekstraksi N-Gram
Dasar lu anjing	Unigram	Dasar, lu, anjing
	Bigram	Dasar lu, lu anjing
	Trigram	Dasar lu anjing

C. Jaccard Similarity

Pada tahap ini akan menjelaskan penghitungan *jaccard similarity* untuk menghitung tingkat persamaan dari kalimat masukan dengan kalimat ada pada dataset, contohnya akan dijelaskan di bawah.

$$Similarity\ (X,Y) = \frac{|x \cap y|}{|x \cup y|} \times 100\% \quad (2)$$

$$Similarity\ (X,Y) = \frac{|PADAHAL\ KELIATANNYA\ ORANG\ BAIK,TAPI\ TERNYATA\ BUSUK|}{|PADAHAL\ ORANG\ BAIK\ TAPI\ TERNYATA\ KELAKUANNYA\ DIBELAKANG\ BUSUK|} \times 100\% \quad (3)$$

$$Similarity\ (X,Y) = \frac{|7|}{|8|} \times 100\% \quad (4)$$

$$Similarity\ (X,Y) = 0{,}875 \times 100\% = 8{,}75\ \% \quad (5)$$

D. Hasil Pengujian

Pada sub-bab ini, aplikasi akan dilakukan suatu pengujian tingkat akurasi dan presisi berdasarkan keberhasilan setiap kalimat hasil transkrip dari *file* audio, yang dimasukkan ke dalam program. Pengujian ini dilakukan dengan cara menyisipkan *file* audio yang memiliki kalimat yang mengandung kalimat kasar atau tidak yang nantinya akan ditranskripsi lalu dibandingkan kemiripannya dengan *dataset* menggunakan *jaccard similarity*.

TABEL VIII
KALIMAT UNTUK PENGUJIAN

Kalimat	Target pendeteksian Kalimat	Target Pendeteksian Makna
Anjing ni org mabok atau stress	anjing ni org mabok atau stress	Kasar
Anjing bangsat tai doang lah	anjing bangsat tai doang lah	Kasar
ya biasa lah otak goblok, apa2 dikaitin sama agama, tp dia sendiri	ya biasa lah otak goblok apa2 dikaitin sama agama tp dia sendiri berdosa	Kasar
Bangsat polisi anjing . Ya Allah tolonglah kami yg teraniaya	bangsat polisi anjing ya allah tolonglah kami yang teraniaya	Kasar
Ga usah sok ngartis sih anjing, jangan sok cantik banyak gaya lu	ga usah sok ngartis sih anjing jangan sok cantik banyak gaya lu	Kasar
kalo temen yang anjing gimana? boleh?	kalo temen yang gimana boleh	Kasar
tapi kamu mukanya kaya anjing	tapi kamu mukanya kaya anjing	Kasar
sumpah lupa kemarin ngapain aja anjing	sumpah lupa kemarin ngapain aja anjing	Kasar
Kenapasih org org rumah kek anjing bgt hari ini	Kenapasih org org rumah kek anjing bgt hari ini	Kasar
Anjing pantesan barang yg dikirim ke gua rusak	Anjing pantesan barang yg dikirim ke gua rusak	Kasar
Dari semua binatang, hanya anjing dan gajah yang pintar memahami gerakan tangan manusia	Dari semua binatang, hanya anjing dan gajah yang pintar memahami gerakan tangan manusia	Halus
Shio mu apa? Babi? Atau monyet	Shio mu apa? Babi? Atau monyet	Halus
Pada saat ini, para ilmuwan di AS tengah mempelajari bagaimana otak seekor monyet dapat mengontrol tubuh monyet lainnya	Pada saat ini, para ilmuwan di AS tengah mempelajari bagaimana otak seekor monyet dapat mengontrol tubuh monyet lainnya	Halus
Ada anjing yang perilakunya mirip manusia	Ada anjing yang perilakunya mirip manusia	Halus
Emang anjing doang yang bisa ngertiin aku	Emang anjing doang yang bisa ngertiin aku	Halus

TABEL IX
HASIL PENGUJIAN

Kalimat	Target pendeteksian Kalimat	Target Pendeteksian Makna	Label
Anjing ni org mabok atau stress	anjing ni org mabok atau stress	Kasar	TP
Anjing bangsat tai doang lah	anjing bangsat tai doang lah	Kasar	TP
ya biasa lah otak goblok, apa2 dikaitin sama agama, tp dia sendiri	ya biasa lah otak goblok apa2 dikaitin sama agama tp dia sendiri berdosa	Kasar	TP
Bangsat polisi anjing . Ya Allah tolonglah kami yg teraniaya	bangsat polisi anjing ya allah tolonglah kami yang teraniaya	Kasar	TN
Ga usah sok ngartis sih anjing, jangan sok cantik banyak gaya lu	ga usah sok ngartis sih anjing jangan sok cantik banyak gaya lu	Kasar	TP
kalo temen yang anjing gimana? boleh?	kalo temen yang gimana boleh	Kasar	TN
tapi kamu mukanya kaya anjing	tapi kamu mukanya kaya anjing	Kasar	TP
sumpah lupa kemarin ngapain aja anjing	sumpah lupa kemarin ngapain aja anjing	Kasar	TP
Kenapasih org org rumah kek anjing bgt hari ini	Kenapasih org org rumah kek anjing bgt hari ini	Kasar	FP
Anjing pantesan barang yg dikirim ke gua rusak	Anjing pantesan barang yg dikirim ke gua rusak	Kasar	TP
Dari semua binatang, hanya anjing dan gajah yang pintar memahami gerakan tangan manusia	Dari semua binatang, hanya anjing dan gajah yang pintar memahami gerakan tangan manusia	Halus	TN
Shio mu apa? Babi? Atau monyet	Shio mu apa? Babi? Atau monyet	Halus	FN
Pada saat ini, para ilmuwan di AS tengah mempelajari bagaimana otak seekor monyet dapat mengontrol tubuh monyet lainnya	Pada saat ini, para ilmuwan di AS tengah mempelajari bagaimana otak seekor monyet dapat mengontrol tubuh monyet lainnya	Halus	TP
Ada anjing yang perilakunya mirip manusia	Ada anjing yang perilakunya mirip manusia	Halus	FN
Emang anjing doang yang bisa ngertiin aku	Emang anjing doang yang bisa ngertiin aku	Halus	FN

Dari kalimat-kalimat yang akan diuji di atas dengan target pendeteksiannya sebanyak 15 kalimat pengujian. Setiap kalimatnya akan diuji pada aplikasi dan dilakukan pengecekan apakah kalimat yang ada di dalam *file* audio bisa di transkrip secara benar dan bisa dideteksi apakah termasuk kalimat kasar atau bukan. Setiap kalimatnya akan ditandakan apakah hasil uji tersebut merupakan:

a. *True Positive* (TP), yaitu target pendeteksian dan hasil deteksi yang dihasilkan memiliki hasil yang sama dan bisa dijalankan
b. *True negative* (TN), yaitu target pendeteksian dan hasil deteksi yang dihasilkan memiliki hasil yang sama tapi tidak bisa dideteksi maknanya
c. *False positive* (FP), yaitu target pendeteksian dan hasil deteksi yang dihasilkan tidak sesuai dengan apa yang dimasukan, namun pendeteksian masih tetap bisa dijalankan
d. *False Negative* (FN), yaitu target pendeteksian dari hasil deteksi tidak sesuai dengan apa yang dimasukan, dan aplikasi tidak bisa mendeteksi makna kalimat tersebut

Berdasarkan hasil dari pengujian tersebut, maka dapat dihitung nilai akurasi dan presisinya berdasarkan label yang telah diberikan. Berikut adalah perhitungannya menggunakan *confusion matrix*

$$Akurasi = TP + TN / TP + TN + FP + FN = 8 + 3 / 8 + 3 + 1 + 3 = 11 / 15 = 73{,}4\ \% \quad (6)$$

$$Presisi = TP / TP + FP = 8 / 8 + 1 = 8 / 9 = 88{,}9\ \% \quad (7)$$

Hasil penelitian memiliki akurasi sebesar 73,4 % dan tingkat presisi sebesar 88,9 %. Ini dapat diartikan bahwa hasil pendeteksian menggunakan *jaccard similarity* dapat menghasilkan 73,4 %. Dengan nilai presisi tersebut, maka sudah bisa menjalankan pendeteksian dengan baik dan masih ada sedikit yang kurang tepat namun masih dapat dijalankan

IV. PENUTUP

Berdasarkan hasil penelitian dan pembahasan yang dilakukan, dapat disimpulkan bahwa aplikasi pendeteksi kalimat kasar bahasa Indonesia menggunakan *Jaccard similarity* dan N-gram dapat diimplementasikan untuk mendeteksi kalimat kasar, serta tingkat akurasi dan presisi yang didapatkan juga cukup baik, untuk akurasi mendapatkan nilai 73,4 % untuk mentranskripsi dan mendeteksi kalimat kasar, serta nilai presisi 88,9 % ini membuktikan bahwa aplikasi ini sudah sangat presisi dalam mendeteksi kalimat kasar bahasa Indonesia yang menggunakan *Jaccard similarity* dan N-gram.

REFERENSI

[1] zulkarnain, "STUDIA SOSIA RELIGIA," 2020. [Daring]. Tersedia pada: http://jurnal.uinsu.ac.id/index.php/ssr
[2] A. F. Hidayatullah, A. Aulia, F. Yusuf, K. P. Juwairi, R. Abida, dan N. Nayoan, "Identifikasi Konten Kasar pada Tweet Bahasa Indonesia," 2019. [Daring]. Tersedia pada: https://t.co/YQCC0CM4gG
[3] B. Wijaya dan V. C. Mawardi, "PENDETEKSI UJARAN KEBENCIAN PADA PLATFORM MEDIA SOSIAL TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE," *Teknik dan Kedokteran*, vol. 1, no. 1, hlm. 11–17, 2023, doi: 10.24912/jsstk.v1i1.22746.
[4] I. And dan D. Expert, "Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia INFORMASI ARTIKEL ABSTRAK," 2022. [Daring]. Tersedia pada: http://index.unper.ac.id
[5] M. Hakiem dan M. Ali Fauzi, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain," 2019. [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id
[6] D. K. Teologi, "STUDIA SOSIA RELIGIA." [Daring]. Tersedia pada: http://jurnal.uinsu.ac.id/index.php/ssr
[7] D. Marta, G. Leonarde Ginting, dan A. Hatuaon Sihite, "Deteksi Berita Palsu Tentang Vaksinasi Covid-19 Dengan Menggunakan Text Mining Dan Algoritma Cosine Similarity," *Nasional Teknologi Informasi dan Komputer)*, vol. 6, no. 1, 2022, doi: 10.30865/komik.v6i1.5738.
[8] "4 Jaccard Similarity and k-Grams."
[9] P. Sulistiyawati *dkk.*, "PREDIKSI KATA KASAR BERBAHASA INDONESIA MENGGUNAKAN MACHINE LEARNING BERBASIS MOBILE INFRASTRUCTURE," *Transmisi*, vol. 24, no. 2, hlm. 55–61, Mei 2022, doi: 10.14710/transmisi.24.2.55-61.
[10] I. And dan D. Expert, "Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia INFORMASI ARTIKEL ABSTRAK," 2022. [Daring]. Tersedia pada: http://index.unper.ac.id
[11] Sunardi, A. Yudhana, dan I. A. Mukaromah, "Implementasi Deteksi Plagiarisme Menggunakan Metode N-Gram dan Jaccard Similarity Terhadap Algoritma Winnowing," *TRANSMISI*, vol. 20, no. 3, hlm. 105–110, 2018.
[12] W. Desena dan A. Solichin, "Pencarian Abstrak Tugas Akhir Mahasiswa Berdasarkan Tingkat Kemiripan Menggunakan Algoritma Winnowing dan Jaccard Similarity pada Universitas Budi Luhur," *Jurnal Informatik*, vol. 17, no. 2, hlm. 112–122, 2021.
[13] K. Rinartha, "Simple Query Suggestion untuk Pencarian Artikel Menggunakan Jaccard Similarity," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 3, no. 1, hlm. 30–34, 2017.