SKRIPSI

ANALISIS SENTIMEN TERHADAP VIDEO BERDASARKAN PENGENALAN EMOSI PADA DATA AUDIO

Disusun dan diajukan oleh:

ANDI MUH. RIZKY
D121 17 1323

PROGRAM STUDI SARJANA
TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS HASANUDDIN
GOWA
2023

LEMBAR PENGESAHAN SKRIPSI

ANALISIS SENTIMEN TERHADAP VIDEO BERDASARKAN PENGENALAN EMOSI PADA DATA AUDIO

Disusun dan diajukan oleh

ANDI MUH. RIZKY

D121171323

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 29 November 2023 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Anugrayani Bustamin, S.T., M.T.	Elly Warni, ST., MT.
Nip. 19901201 2018074 001	Nip. 19820216 2008122 001

Ketua Program Studi,

Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng.
Nip. 19750716 2002121 004

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini ;

Nama	: Andi Muh. Rizky
NIM	: D121 17 1323
Program Studi	: Teknik Informatika
Jenjang	: S1

Menyatakan dengan ini bahwa karya tulisan saya berjudul

Analisis Sentimen Terhadap Video Berdasarkan Pengenalan Emosi pada Data Audio

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan tulisan orang lain dan bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Semua informasi yang ditulis dalam skripsi yang berasal dari penulis lain telah diberi penghargaan, yakni dengan mengutip sumber dan tahun penerbitannya. Oleh karena itu semua tulisan dalam skripsi ini sepenuhnya menjadi tanggung jawab penulis. Apabila ada pihak manapun yang merasa ada kesamaan judul dan atau hasil temuan dalam skripsi ini, maka penulis siap untuk diklarifikasi dan mempertanggungjawabkan segala resiko.

Segala data dan informasi yang diperoleh selama proses pembuatan skripsi, yang akan dipublikasi oleh Penulis di masa depan harus mendapat persetujuan dari Dosen Pembimbing.

Apabila dikemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan isi skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Gowa, 22 November 2023

Yang Menyatakan

Andi Muh. Rizky

ABSTRAK

ANDI MUH. RIZKY. *Analisis Sentimen Terhadap Video Berdasarkan Pengenalan Emosi pada Data Audio* (dibimbing oleh Anugrayani Bustamin dan Elly Warni

YouTube menjadi salah satu *platform streaming* video yang populer dalam mencari suatu informasi khususnya video ulasan produk. Meskipun sebuah video ulasan semakin dicari oleh pengguna YouTube namun kebanyakan video memiliki durasi yang panjang dan terdapat perbedaan secara verbal dengan emosi suara yang disampaikan, sehingga perlu metode yang cepat dan akurat untuk memahami emosi suara pada video.

Penelitian ini menggunakan data primer dan data sekunder. Data primer merupakan kumpulan rekaman suara dari 10 responden dan data sekunder merupakan kumpulan dari 60 sampel video YouTube pada kanal GadgetIn. Penelitian ini juga membangun sebuah model yang dapat mengenali dan mengklasifikasi emosi suara menggunakan arsitektur *Convolutional Neural Network* (CNN) serta dilakukan evaluasi performa model pengenalan emosi suara.

Arsitektur CNN adalah jenis jaringan saraf yang terdiri dari lapisan konvolusi, aktivasi, *pooling*, dan *fully connected* yang mampu mengekstrak fitur-fitur penting dari data audio secara berurutan. Augmentasi data dilakukan dengan melakukan metode *noise*, *time stretch*, dan *pitch shift*. Pelatihan data menggunakan 6 skenario yaitu skenario pelatihan data primer sejumlah 300 data, skenario pelatihan data primer dengan augmentasi sejumlah 900 data, skenario pelatihan data sekunder sejumlah 500 data, skenario pelatihan data sekunder dengan augmentasi sejumlah 1497 data, skenario pelatihan data gabungan sejumlah 800 data, dan skenario pelatihan data gabungan dengan augmentasi sejumlah 2397 data.

Hasil analisis dan pengujian performa model berdasarkan klasifikasi 3 kelas sentimen (positif, negatif, netral) menghasilkan 6 model dengan nilai akurasi yang cukup tinggi yaitu model data primer dengan augmentasi sebesar 94%, model data primer sebesar 92%, model data gabungan dengan augmentasi sebesar 90%, model data sekunder dengan augmentasi sebesar 85%, dan model data sekunder serta model data gabungan sebesar 81%.

Kata Kunci: CNN, Emosi, Klasifikasi, Sentimen, Video

ABSTRACT

ANDI MUH. RIZKY. *Sentiment Analysis of Videos Based on Emotion Recognition in Audio Data* (supervised by Anugrayani Bustamin and Elly Warni)

YouTube has become a popular video streaming platform for finding information, especially product review videos. Although YouTube users increasingly seek video reviews, most videos have a long duration, and there are differences verbally with the sound emotion conveyed, so a fast and accurate method is needed to understand the sound emotion in the video.

This research uses primary data and secondary data. Primary data is a collection of voice recordings from 10 respondents, and secondary data is a collection of 60 YouTube video samples on the GadgetIn channel. This research builds a model that can recognize and classify speech emotions using the Convolutional Neural Network (CNN) architecture and evaluates the speech emotion recognition model.

CNN architecture is a type of neural network consisting of convolution, activation, pooling, and fully connected layers that are able to extract important features from audio data sequentially. Data augmentation uses noise, time stretch, and pitch shift methods. Data training involved six scenarios: the primary data training with 300 data, the primary data training using augmentation with 900 data, the secondary data training with 500 data, the secondary data training using augmentation with 1497 data, the combined data training with 800 data, and the combined data training using augmentation with 2397 data.

The results of analysis and testing of model performance based on the classification of 3 sentiment classes (positive, negative, neutral) produced 6 models with quite high accuracy values, namely the primary data model with augmentation of 94%, the primary data model of 92%, the combined data model with augmentation of 90%, secondary data model with augmentation at 85%, secondary data model and combined data model at 81%.

Keywords: CNN, Emotion, Classification, Sentiment, Videos

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN DAN ARTI SIMBOL

Lambang/Singkatan	Arti dan Keterangan
SER	*Speech Emotion Recognition*
CNN	*Convolutional Neural Network*
TP	*True Positive*
TN	*True Negative*
FP	*False Positive*
FN	*False Negative*

DAFTAR LAMPIRAN

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan atas kehadirat Allah *subhanahu wa ta'ala* yang tiada batas memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan tugas akhir dengan judul "Analisis Sentimen Terhadap Video Berdasarkan Pengenalan Emosi pada Data Audio" sebagai syarat dalam menyelesaikan studi jenjang Strata-1 di Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin.

Penulis menyadari bahwa dalam penyusunan dan penulisan laporan penelitian ini tidak terlepas dari doa, usaha, dan bantuan dari berbagai pihak, mulai masa perkuliahan hingga masa penyusunan tugas akhir ini. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih dengan bangga kepada:

1. Keluarga tercinta terkhusus kepada kedua orang tua penulis, Bapak Andi Tanjong dan Ibu Herawati yang selalu memberikan pikiran, dukungan, doa, dan bantuannya yang tak terbatas.
2. Pembimbing utama, Ibu Anugrayani Bustamin, S.T., M.T. dan pembimbing pendamping, Ibu Elly Warni, S.T., M.T. atas waktu, pikiran, perhatian, tenaga, arahan dan bantuannya yang luar biasa hingga saat ini penulis dapat menyelesaikan penelitian ini.
3. Ketua Departemen Teknik Informatika Fakultas Teknik Universitas Hasanuddin, Bapak Prof. Dr. Ir. Indrabayu, S.T., M.T., M.Bus.Sys., IPM, ASEAN. Eng. atas ilmu, wawasan, dan pengalaman yang diberikan kepada penulis.
4. AIMP Research Group, yang selalu memberikan ruang diskusi dan memberikan banyak masukan kepada penulis.
5. Rekan tim penelitian Multimodal, saudara Amiruddin dan Muh. Ikbal yang banyak memberikan dukungan dan pemikiran-pemikiran dalam penelitian penulis.
6. Responden dari penyiar radio PRO2 Makassar dan Prambors Makassar, Muh. Ibrahim, Dischidia Darajat, Anggi Athifah, kak Nindi, kak Papam, Ocang, Enggi karena bantuan mereka penelitian ini dapat terselesaikan dengan baik.
7. Komisaris dan Direktur PT. Maroa Tungga Abadi, Ir. Danto Joro dan saudara Muh. Safar Alfarel S.T. yang telah memberikan dukuran moral dan materi kepada penulis.
8. DOMI09, sebagai rumah kedua dan tempat cerita penulis.
9. Orang-orang hebat dan spesial sebagai sosok yang telah hadir dalam kehidupan saat senang dan sulit penulis.
10. Teknik Informatika angkatan 2017 (RECOG17NER), saudara seperjuangan yang menyimpan banyak kenangan dan pengalaman tak terlupakan karena "takkan ada kisah tanpa ada kamu, aku, dan kalian".
11. Last but no least, terima kasih yang luar biasa kepada diri sendiri yang pantang menyerah telah berjuang hingga di titik ini "KEEP ON FIGHTING TILL THE END".

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, media sosial telah menjadi *platform* yang diminati untuk mencari dan berbagi informasi. Salah satu media sosial dan layanan *streaming* video populer adalah YouTube. YouTube memungkinkan pengguna untuk mengunggah, menonton, dan berbagi video (Burgess & Green, 2018).

Beragamnya video yang disajikan dalam *platform* YouTube sangat membantu pengguna untuk mencari dan mendapatkan informasi yang dibutuhkan diantara lain video ulasan produk. Video tentang mengulas sebuah produk menjadi salah satu konten yang banyak dicari oleh pengguna YouTube untuk membantu mereka dalam membeli sebuah barang atau produk (Huang et al., 2022). Berdasarkan hasil survei pada tahun 2019 yang dilakukan oleh platform *bright local* sebanyak 82% orang mempercayai *online review* untuk mempercayai suatu produk (Murphy, 2019). Namun, dengan durasi video yang terkadang panjang dan terdapat perbedaan secara verbal dan emosi yang disampaikan sehingga pengguna membutuhkan metode yang cepat dan akurat untuk memahami emosi suara pada konten video.

Memahami emosi manusia adalah aspek penting dalam komunikasi antar manusia (El Ayadi et al., 2011). Mengeksplorasi bagaimana emosi yang diungkapkan dalam ulasan produk mempengaruhi persepsi manfaat dari ulasan tersebut (Ahmad & Laroche, 2015). Teknologi pengenalan emosi suara dapat memiliki peran penting dalam berbagai bidang, termasuk interaksi manusia dan komputer, layanan pelanggan, dan dalam penelitian ini yaitu analisis video ulasan produk. Analisis sentimen berdasarkan pengenalan emosi dapat menjadi teknologi yang bermanfaat (Pang & Lee, 2008).

Pengenalan emosi suara merupakan teknologi yang mengidentifikasi emosi dari suara atau ucapan manusia. Ada beberapa metode atau arsitektur yang dapat
n dalam mengidentifikasi emosi suara, salah satunya adalah *Convolutional*
etwork (CNN). CNN adalah jenis jaringan saraf tiruan yang telah terbukti
alam berbagai tugas pengenalan pola, termasuk pengenalan suara dan

pengenalan emosi (Lecun et al., 2015). CNN merupakan metode yang potensial untuk mengidentifikasi emosi dari sinyal suara dan memiliki kemampuan unik untuk belajar fitur yang paling relevan dari data mentah yang membuatnya sangat cocok untuk tugas yang memerlukan pemrosesan data berdimensi tinggi seperti pengenalan suara dan pengenalan emosi (Lecun et al., 2015). Oleh karena itu, penelitian ini menggunakan arsitektur CNN untuk merancang dan membangun sebuah model yang nantinya dapat digunakan untuk mengenali emosi suara dan mengetahui tingkat sentimen pada video ulasan produk.

Sampai saat ini penelitan *Speech Emotion Recognition* (SER) yang berbasis bahasa Indonesia masih sangat sedikit. Hal ini disebabkan keterbatasan data atau korpus berbahasa indonesia untuk pengenalan emosi suara (Khoirotul Aini et al., 2021). Oleh karena itu, penulis mengembangkan data atau korpus berbahasa indonesia yang terbagi menjadi data primer dan data sekunder. Korpus atau data ini akan diberikan pelatihan yang beragam dan validasi data pada model yang telah dibangun.

Berdasarkan latar belakang tersebut, penulis mengambil penelitian dengan judul **"Analisis Sentimen Terhadap Video Berdasarkan Pengenalan Emosi pada Data Audio"**. Penelitian ini sebagai bagian atau langkah awal dari sistem multimodal yang tidak melibatkan hanya satu aspek melainkan berbagai aspek seperti teks, dan visual pada video.

Penelitian ini diharapkan dapat memberikan kontribusi penting dalam pengembangan teknologi analisis sentimen khususnya dalam konteks video tentang ulasan produk. Hasil dari penelitian ini diharapkan dapat membantu pengguna dalam mengevaluasi konten video dengan lebih efisien dan akurat.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka rumusan masalah dari penelitian ini adalah:

a. Bagaimana mengenali fitur dan mengklasifikasi emosi suara menggunakan CNN?

agaimana performa model pengenalan emosi suara menggunakan CNN?

1.3 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

a. Untuk mengenali dan mengklasifikasi emosi suara menggunakan CNN.
b. Untuk mengetahui performa sistem pengenalan emosi suara menggunakan CNN.

1.4 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah:

a. Untuk memberikan informasi analisis sentimen pada video ulasan produk yang dapat menjadi referensi konsep multimodal pada analisis sentimen yang lebih kompleks.
b. Menjadi referensi untuk pengembangan penelitian mendatang yang relevan.

1.5 Ruang Lingkup

Ruang lingkup dalam penelitian ini adalah:

a. Data primer merupakan kumpulan rekaman suara berbahasa Indonesia dari 10 responden (5 laki-laki dan 5 perempuan) sejumlah 300 file audio.
b. Data sekunder merupakan kumpulan video YouTube kanal GadgetIn sejumlah 60 file video.
c. Terdiri dari 5 kelas emosi (*neutral, happy, surprise, disgust, disappointed*) yang nantinya akan diklasifikasi secara general dalam 3 kelas sentimen (positif, negatif, dan netral).
d. Pengujian menggunakan 6 skenario pelatihan (pelatihan data primer, pelatihan data primer dengan augmentasi, pelatihan data sekunder, pelatihan data sekunder dengan augmentasi, pelatihan data gabungan, pelatihan data gabungan dengan augmentasi).
e. Pemodelan data latih menggunakan arsitektur *Convolutional Neural Network* (CNN).
f. Evaluasi dan pengukuran performa model menggunakan *confusion matrix* dan metrik evaluasi (*Precision, Recall, Accuracy*).
g. Pengujian sistem menggunakan *input* data audio.

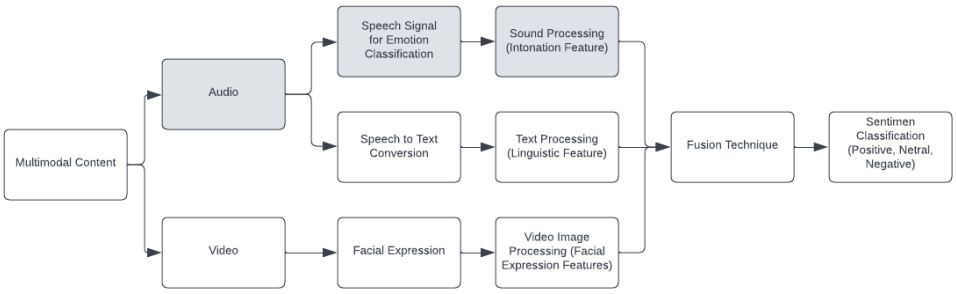
BAB 2
TINJAUAN PUSTAKA

2.1 Multimodal

Multimodal adalah sebutan yang mengacu pada metode komunikasi seseorang yang menggunakan berbagai aspek secara bersamaan (Kress & Leeuwen, 1996), dan diartikan sebagai penerapan beragam aspek semiotik atau peristiwa semiotik yang berlangsung bersamaan, serta bagaimana aspek-aspek tersebut disatukan untuk meningkatkan, melengkapi, atau disusun dengan urutan khusus (Kress, G. & Van Leeuwen, 2001).

Multimodal mengacu pada lebih dari satu mode atau aspek komunikasi yang digunakan secara bersamaan dalam memaknai informasi yang dikonstruksikan tidak hanya secara verbal, tetapi juga melalui gambar visual dan terkadang melalui suara. Saat ini, dengan perkembangan teknologi yang maju terus menerus dan berkesinambungan, ditemukan cara-cara baru dalam memaknai suatu kalimat menggunakan konsep multimodal, dan untuk menyikapi hal tersebut para ahli bahasa membuat makna dengan bentuk literasi baru yang melibatkan bahasa dan gambaran visual yaitu multiliterasi. Multimodal menekankan bahwa pentingnya setiap media komunikasi, baik itu verbal maupun non verbal dalam menghasilkan makna (Kress, G. & Van Leeuwen, 2001).

Struktur multimodal yang berfokus pada data audio merujuk pada poster penelitian oleh Bustamin, A. (2023) yang ditampilkan pada Gambar 1.



Gambar 1 Struktur multimodal pada data audio (Bustamin, 2023)

ltimodal merujuk pada kombinasi data audio dengan jenis data lain untuk tujuan tertentu. Dalam konteks data audio, khususnya dalam bidang

pengenalan emosi suara, aspek intonasi dapat menjadi parameter untuk mendapatkan informasi sentimen yang diharapkan.

2.2 Analisis Sentimen

Analisis sentimen merupakan suatu cara untuk menilai opini dengan mengekstrak, mengenali, dan mengevaluasi informasi tulisan atau lisan dalam menentukan suatu opini bersifat positif, negatif, atau netral. Biasanya, analisis sentimen berupaya untuk menentukan pendapat pembicara, penulis, atau subjek lain dalam kaitannya dengan tema melalui respons emosional terhadap suatu informasi atau peristiwa (Alsaeedi & Khan, 2019).

2.3 Emosi

Kondisi emosional seseorang merupakan faktor penting dalam interaksi antar manusia. Menurut penelitian oleh Picard (1997) menyatakan bahwa emosi memiliki peran penting dalam proses komunikasi antara manusia. Beberapa aspek komunikasi seperti ekspresi wajah, karakteristik suara, dan konten informasi linguistik secara signifikan dipengaruhi oleh kondisi emosional seseorang (Elfenbein, 2007). Emosi didefinisikan sebagai pengalaman sadar yang relatif singkat yang ditandai oleh aktivitas mental yang intens dan tingkat kesenangan atau ketidaknyamanan yang tinggi (Cabanac, 2002) . Emosi ini memiliki pengaruh besar pada aspek lain dalam kehidupan seperti sikap, perilaku, dan penyesuaian pribadi (Hurlock, 1999).

Emosi dapat mempengaruhi cara seseorang berinteraksi dengan orang lain, membuat keputusan, dan merespon situasi yang berbeda karena mengalami emosi bisa sangat subjektif. Emosi yang dirasakan oleh setiap individu sangatlah beragam. Sebagai contoh, tingkat keberagaman emosi marah yang dirasakan oleh masing-masing orang bisa sangat berbeda, mulai dari ketidaknyamanan yang ringan hingga ke rasa marah yang besar. Dengan demikian, penting untuk mempertimbangkan perbedaan individual seseorang dalam pengalaman emosi.

dalam bidang psikologi, emosi sering menyebabkan kontroversi karena
n dalam mendefinisikan dimensi dan kelas emosi yang terlibat (Barrett &
1999). Dalam penelitian ini, penulis mengadopsi deskripsi emosi

menggunakan dua dimensi yaitu dimensi *valence* dan *arousal*. Gambar 2 mengilustrasikan dimensi *valence* dan *arousal* sehubungan dengan kelas emosi secara umum.

Gambar 2 Dimensi valence-arousal emotion (Barrett & Russell, 1999)

Pada Gambar 2, *valence* diartikan sebagai peristiwa emosional *pleasant* (positif) atau *unpleasant* (negatif). Misalnya, emosi positif ditunjukkan dengan arah kanan yang diartikan sebagai hal yang menyenangkan sementara emosi negatif ditunjukkan dengan arah sebaliknya yang diartikan sebagai hal yang tidak menyenangkan. Di sisi lain, *arousal* merujuk pada intensitas emosi yang dirasakan (aktif atau pasif), yaitu kekuatan keadaan emosional yang terkait. Misalnya, ketakutan berada di tempat yang tinggi atau aktif dalam *arousal* sementara kesedihan berada di tempat yang cenderung rendah atau pasif. Dengan pendekatan ini, emosi dapat ditempatkan dalam bidang dua dimensi yang menggambarkan *valence* dan *arousal*.

Seiring dengan perkembangan teknologi, kebutuhan untuk interaksi cerdas antara manusia dan komputer juga meningkat. Dalam hal ini, pengenalan, penafsiran, dan respons terhadap emosi manusia menjadi penting. Picard (1997) dalam bukunya mengenai "*Affective Computing*" menggambarkan pentingnya komputer dalam mengenali dan menanggapi emosi manusia untuk menciptakan interaksi yang lebih alami dan efektif. Teknologi ini telah banyak ıentasikan dalam berbagai bidang. Misalnya, dalam survei kepuasan n, teknologi pengenalan emosi dapat digunakan untuk mendapatkan ı yang lebih baik tentang pengalaman pelanggan. Di sisi lain, dalam bidang

computer vision dan kecerdasan buatan, teknologi ini digunakan untuk membuat sistem yang lebih responsif dan mampu berinteraksi dengan pengguna secara lebih alami (Zeng et al., 2009).

2.4 *Convolutional Neural Network*

Convolutional Neural Network (CNN) adalah jenis jaringan saraf khusus yang digunakan untuk mengolah data dengan topologi mirip grid. Jaringan saraf konvolusional menunjukkan bahwa jaringan tersebut menggunakan operasi matematika yang disebut konvolusi. Konvolusi sendiri adalah operasi linier. Oleh karena itu, jaringan konvolusional adalah jaringan saraf yang menggunakan konvolusi minimal di setiap lapisannya (Zhao et al., 2019). CNN adalah kasus khusus Jaringan Syaraf Tiruan dan saat ini dianggap sebagai model terbaik untuk memecahkan masalah pengenalan dan deteksi objek (Lecun et al., 1998).

Struktur dasar dari *Convolutional Neural Network* (CNN) atau Jaringan Saraf Konvolusional biasanya meliputi tiga jenis lapisan yaitu lapisan konvolusi, lapisan *pooling*, dan lapisan *fully connected*.

2.4.1 *Convolution layer*

Convolution layer atau lapisan konvolusi adalah komponen kunci dalam arsitektur CNN. Konvolusi dalam konteks ini merujuk pada operasi matematika yang menerapkan filter atau *kernel* pada input. Filter ini biasanya berukuran lebih kecil daripada *input* dan akan digeser atau diaplikasikan pada seluruh *input* untuk menghasilkan yang dikenal sebagai *feature map*. *Feature map* ini merepresentasikan informasi spesifik yang telah ditemukan filter dalam *input* seperti garis, tepi, atau tekstur (Lecun et al., 1998). Secara umum operasi konvolusi dapat dilihat pada persamaan 1.

$$s(t) = (x * w)(t) \quad (1)$$

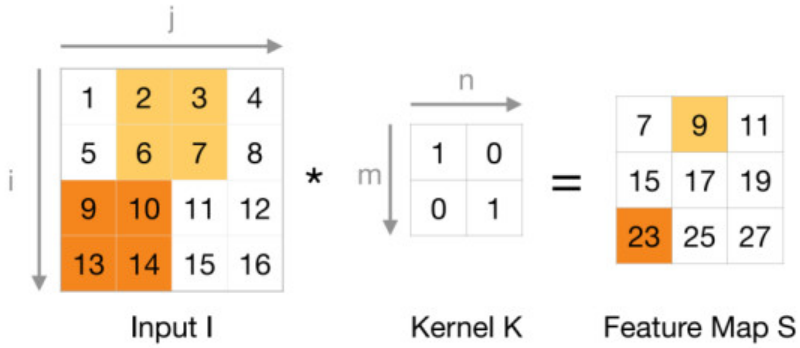
dimana,

$s(t)$ = Fungsi hasil operasi konvolusi (*feature map)*

x = Masukan (*input*)

w = Bobot atau *kernel*

Tujuan utama lapisan konvolusi adalah untuk mengekstrak fitur lokal dari *input* yang membuat CNN ideal untuk data yang memiliki struktur spasial (I. J. Goodfellow et al., 2012). Dalam praktiknya, banyak filter digunakan dalam setiap lapisan konvolusi dengan mencoba mendeteksi fitur yang berbeda. Setiap lapisan konvolusi menghasilkan beberapa *feature map* seperti pada Gambar 3.



Gambar 3 Ilustrasi operasi konvolusi (Neumann, 2021)

Gambar 3 mengilustrasikan konvolusi dengan *input* 4x4, *kernel* 2x2, dan menghasilkan *feature map* 3x3. Area berwarna pada input sesuai dengan *output* yang digabungkan dalam *feature map* yang dihasilkan dengan warna yang sama. Jarak perpindahan kernel disebut *stride* yang diatur dalam 1 *stride*. Menggunakan beberapa filter memungkinkan CNN untuk mendeteksi berbagai jenis fitur dalam *input* yang memungkinkan untuk merepresentasikan *input* secara lebih akurat dan lengkap (Krizhevsky et al., 2012).

2.4.2 *Pooling*

Pooling adalah operasi penting lainnya dalam arsitektur *Convolutional Neural Network* (CNN). Fungsi utamanya adalah untuk mengurangi dimensi spasial (panjang dan lebar) dari *input* yang diterima dari lapisan sebelumnya dengan cara mengurangi jumlah parameter, komputasi dalam jaringan, dan membantu mencegah *overfitting* (Lecun et al., 1998).

Lapisan *pooling* biasanya ditempatkan setelah satu atau beberapa lapisan konvolusi dalam arsitektur CNN. Lapisan *pooling* berfungsi untuk secara progresif mengurangi dimensi spasial dari *input* melalui *depth* dari jaringan, sehingga dapat fokus pada fitur-fitur lokal yang sangat spesifik di awal hingga ke

fitur yang lebih abstrak dan global di bagian akhir jaringan (Krizhevsky et al., 2012).

2.4.3 *Fully Connected Layer*

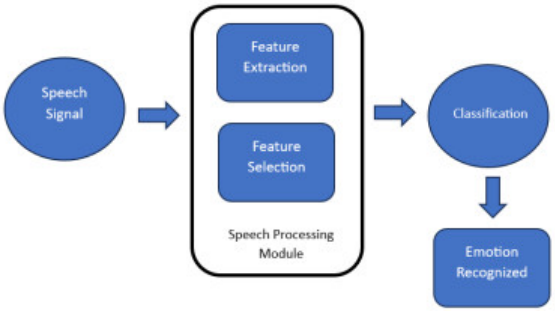
Fully Connected Layer yang juga dikenal sebagai *Dense Layer* adalah komponen integral dalam struktur *Convolutional Neural Network* (CNN). Pada lapisan ini, setiap neuron terhubung ke setiap neuron pada lapisan sebelumnya (I. Goodfellow et al., 2016). Fungsi utama dari *Fully Connected Layer* adalah untuk belajar fitur non-lokal dari representasi sebelumnya yang dipelajari oleh lapisan konvolusional dan *pooling*. Setelah beberapa tahap konvolusi dan *pooling*, representasi dari fitur asli telah diubah menjadi level yang sangat tinggi dan abstrak. Lapisan *Fully Connected* kemudian berfungsi untuk mempelajari kombinasi dari fitur-fitur abstrak ini, memetakan ke dalam ruang fitur yang lebih rendah dan akhirnya melakukan klasifikasi (Lecun et al., 2015).

2.5 *Speech Emotion Recognition*

Pengenalan emosi suara atau *Speech Emotion Recognition* (SER) adalah cabang penelitian dalam bidang pengolahan sinyal dan pembelajaran mesin yang berfokus pada identifikasi dan klasifikasi emosi dalam sinyal suara manusia (El Ayadi et al., 2011). Mengingat pentingnya emosional dalam interaksi atau komunikasi manusia, peningkatan SER memiliki aplikasi yang luas dalam berbagai bidang seperti perangkat interaktif yang responsif terhadap emosi, analisis suasana hati dalam layanan kesehatan mental, dan peningkatan interaksi manusia-komputer (Schuller et al., 2009).

Metode umum dalam pengenalan emosi suara melibatkan ekstraksi fitur, pemilihan fitur, dan klasifikasi. Ekstraksi fitur adalah proses mengubah input suara mentah menjadi set fitur yang mewakili informasi yang relevan untuk tugas pengenalan emosi. Pemilihan fitur melibatkan pengurangan dimensi fitur untuk menghilangkan fitur yang kurang informatif dan mempercepat proses klasifikasi. si dilakukan menggunakan teknik pembelajaran mesin untuk ıtifikasi emosi yang sesuai dari fitur-fitur tersebut (Eyben et al., 2013).

Gambar 4 menggambarkan sistem sederhana yang digunakan untuk pengenalan emosi suara atau ucapan.



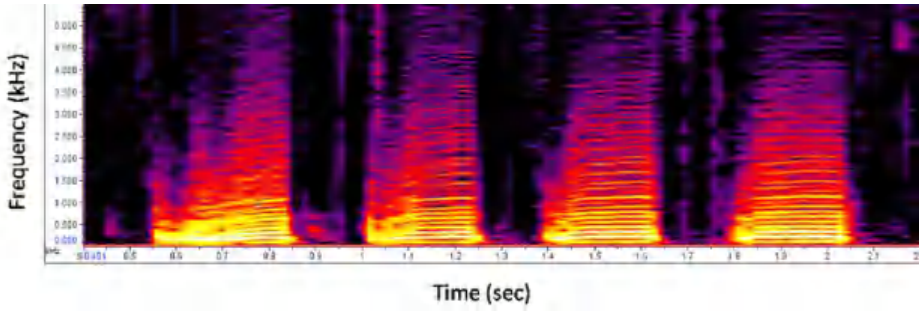
Gambar 4 Sistem pengenalan emosi suara sederhana (Rendi Nurcahyo & Mohammad Iqbal, 2022)

Pada Gambar 4, tahap pertama yaitu pemrosesan sinyal berbasis ucapan dan peningkatan suara dilakukan di mana gangguan bising dihilangkan. Tahap kedua melibatkan dua bagian yaitu ekstraksi fitur dan seleksi fitur. Fitur yang diperlukan diekstraksi dari sinyal ucapan yang telah diproses sebelumnya dan pemilihan dibuat dari fitur yang diekstraksi. Ekstraksi dan pemilihan fitur tersebut biasanya didasarkan pada analisis sinyal suara dalam domain waktu dan frekuensi. Selama tahap ketiga, berbagai pengklasifikasi digunakan untuk klasifikasi fitur. Terakhir, berdasarkan klasifikasi fitur, emosi yang berbeda dapat dikenali.

2.6 *Spectrogram*

Spectrogram adalah representasi visual dari spektrum frekuensi suara atau sinyal lainnya dalam waktu. Ini biasa digunakan dalam bidang pengolahan sinyal, neurologi, dan linguistik yang di mana *spectrogram* dapat membantu dalam analisis dan interpretasi karakteristik frekuensi suara (Quatieri, 2006).

Spectrogram dapat digunakan untuk analisis berbagai sinyal akustik, termasuk pidato, musik, suara lingkungan, dan lainnya yang berhubungan dengan sinyal suara (Fulop & Fitz, 2006). *Spectrogram* telah banyak digunakan dalam pelatihan model *deep learning* seperti *Convolutional Neural Networks* (CNN) untuk pengenalan suara dan pengenalan emosi (Amiriparian et al., 2018). Model ini ajar fitur penting dari *spectrogram* yang mencakup pola frekuensi dan erta digunakan untuk mengklasifikasikan suara atau emosi. Visual am ditampilkan pada Gambar 5.



Gambar 5 Visual spektrogram (Kovitvongsa & Lobel, 2009)

Pada Gambar 5, terlihat plot frekuensi suara dari waktu ke waktu, sumbu (x) adalah Waktu (sec) dan sumbu Y adalah Frekuensi (kHz). Jumlah energi yang ada di setiap frekuensi diwakili oleh intensitas warna. Semakin cerah warnanya, semakin banyak energi yang hadir dalam suara pada frekuensi tersebut dan semakin gelap warnanya semakin pelan dan sedikit energi yang hadir (Kovitvongsa & Lobel, 2009).

2.7 *Data Augmentation*

Data augmentation adalah teknik yang digunakan untuk meningkatkan satu atau lebih variasi pada kumpulan data pelatihan beranotasi yang menghasilkan data pelatihan tambahan baru dikarenakan *deep learning* sangat bergantung pada ketersediaan data pelatihan dalam jumlah besar untuk mempelajari fungsi non-linier dari *input* ke *output* yang dapat digeneralisasi dengan baik dan menghasilkan akurasi klasifikasi yang tinggi. (Salamon & Bello, 2017).

Hasil penelitian yang dilakukan oleh Wei, S. (2020) telah membuktikan secara eksperimental bahwa metode augmentasi data dengan menambahkan *noise*, *stretch*, *pitch shift* sangat membantu untuk meningkatkan kinerja klasifikasi pada data audio.

2.7.1 *Noise*

Noise merupakan cara augmentasi dengan menambahkan *gaussian noise* ke sampel audio. Amplitudo *noise* (σ) adalah faktor penting pada noise dengan *hyper-*
r yang dapat dikonfigurasi. Jika nilai amplitudo *noise* diatur terlalu kecil, t terjadi gangguan dan ketika diatur terlalu besar, pengklasifikasinya sulit elajari. Kita dapat mengatur dengan rentang $\sigma \in [0{,}001, 0{,}035]$ dan

mengikuti distribusi *uniform*, x(t) adalah sinyal mentah dari data audio (Wei et al., 2020). Sampel yang baru dihasilkan setelah augmentasi menggunakan *noise* dapat dinyatakan dengan persamaan 2.

$$x(t) = x(t) + \sigma \times n(t) \quad (2)$$

dimana,

x(t) = Sinyal mentah data audio

σ = *Amplitudo noise*

2.7.2 *Stretch*

Stretch atau *time stretch* merupakan teknik augmentasi dengan cara mengubah tempo dan panjang klip audio tanpa mengubah tingkat nadanya (Wei et al., 2020). γ merupakan faktor *stretch.* Kita dapat mengatur dengan rentang $\gamma \in [0{,}8, 1{,}25]$, mengikuti distribusi uniform. Jika $\gamma > 1$ maka sinyal data bertambah cepat dan jika $\gamma < 1$ maka sinyal data menjadi lambat. Agar sinyal yang telah diaugmentasi sesuai dengan ukuran *input*, kita perlu memastikan bahwa panjang sinyal yang telah dilakukan augmentasi *stretch* tetap sama. Kita bisa menerapkan *zero padding* untuk memenuhi panjang sinyal atau menentukan durasi panjang audio jika terlalu panjang (Wei et al., 2020).

2.7.3 *Pitch*

Pitch merupakan teknik augmentasi dengan mengubah *pitch* atau nada suatu gelombang dengan *n_steps semitones* tanpa mengubah temponya. Ini adalah proses timbal balik pada *time stretch*. Parameter *n_steps* dapat diatur dengan rentang *n_steps* $\in$ [-4, 4], mengikuti distribusi uniform (Wei et al., 2020).

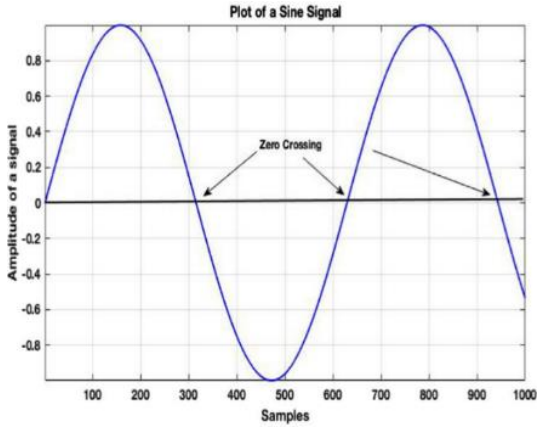
2.8 *Feature Extraction*

Ekstraksi fitur merupakan proses menandai karakteristik sinyal yang paling mendominasi dan membedakan. Evolusi fitur data audi dibagi menjadi beberapa kategori yaitu berdasarkan domain waktu, domain frekuensi, gabungan domain -waktu, dan *deep features*. Penggunaan ektraksi fitur seperti ZCR, MFCC, STFT, RMS, dan *Mel Spectrogram* masih memiliki peran penting untuk

melakukan ekstraksi fitur dalam analisis sinyal audio pada era saat ini (Sharma et al., 2020).

2.8.1 ZCR (*Zero Crossing Rate*)

ZCR atau *Zero Crossing Rate* merupakan perubahan tanda sinyal pada suatu *frame* tertentu. Secara matematis, yaitu berapa kali sinyal berpindah dari positif ke negatif dan sebaliknya dibagi dengan panjang *frame.* Sederhananya, ZCR adalah berapa kali sinyal melintasi level nol dalam interval satu detik (Sharma et al., 2020). Contoh visual konsep ZCR ditampilkan pada Gambar 6.



Gambar 6 Sinyal *zero-crossing*

Pada Gambar 6, visual sinyal *Zero-Crossing* yaitu perpindahan gelombang amplitudo yang melewati garis nol. ZCR untuk frame ke-i dengan panjang *N* dinyatakan pada persamaan 3.

$$Z(i) = \frac{1}{2N}\sum_{n=1}^{N} |sgn[Xi(n)] - sgn[Xi(n-1)]| \quad (3)$$

dimana, nilai *sgn* (*signum*) dapat dilihat pada persamaan 4.

$$sgn[Xi(N)] = \begin{cases} 1, x_i(n) \geq 0 \\ 0, x_i(n) < 0 \end{cases} \quad (4)$$

2.8.2 C*hroma* STFT

ır *chroma* adalah representasi audio yang pada dasarnya mewakili 12 kelas nada berbeda dan dapat digunakan untuk memberikan perbedaan

profil kelas nada pada sinyal audio. Hal ini dapat dihitung dari logaritmik *short-time fourier transform* dari sinyal audio. Ini juga disebut sebagai kromogram (Sharma et al., 2020).

2.8.3 MFCC (*Mel Frequency Cepstral Coefficient*)

MFCC merupakan salah satu metode yang sering digunakan dalam bidang teknologi suara, baik *speech recognition* maupun *speech emotion recognition*. Metode ini digunakan untuk melakukan *feature extraction*, sebuah proses yang mengkonversikan sinyal suara menjadi beberapa parameter (Solehudin et al., 2018). Proses pengolahan sinyal menggunakan ekstraksi fitur MFCC yaitu:

1. ***Pre-emphasis***

Pre-emphasis merupakan salah satu jenis filter yang biasa digunakan sebelum sinyal diproses. Filter tersebut mempertahankan frekuensi-frekuensi tinggi pada sebuah spektrum yang umumnya tereliminasi pada saat proses produksi suara. Tujuan dari *pre-emphasis* yaitu mengurangi *noise ratio* pada sinyal sehingga dapat meningkatkan kualitas sinyal selain itu untuk mendapatkan bentuk spektral frekuensi sinyal ucapan yang lebih halus dan menyeimbangkan spektrum dari sinyal suara. Filter *pre-emphasis* didasari oleh hubungan masukan dan keluaran dalam domain waktu yang dinyatakan dalam persamaan 5.

$$y(n) = \mathrm{s(n)} - \alpha s(n-1) \tag{5}$$

dimana,

y(n) = sinyal hasil *pre-emphasis*,
s(n) = sinyal sebelum *pre-emphasis*,
α = konstanta filter pre-emphasis, biasanya bernilai 0,97.

2. ***Frame Blocking***

Frame Blocking adalah proses membagi suara menjadi beberapa *frame*

dapat memudahkan dalam perhitungan dan analisa sinyal suara, satu

e terdiri dari beberapa sampel tergantung tiap berapa detik suara akan

npel dan berapa besar frekuensi *sampling*-nya. Panjang *frame* yang

digunakan sangat mempengaruhi keberhasilan dalam analisa spektral. Pada satu sisi, ukuran dari *frame* harus sepanjang mungkin untuk dapat menunjukkan resolusi frekuensi yang baik. Tetapi di sisi lain, ukuran *frame* juga harus cukup pendek untuk dapat menunjukkan resolusi waktu yang baik. Representasi fungsi *frame blocking* dapat dilihat pada persamaan 6.

$$x(n) = y\ (M + n) \tag{6}$$

dimana,

$x(n)$ = sinyal sesudah di-*frame blocking*,

y = sinyal hasil *pre-emphasis*,

M = *overlapping frame*.

3. *Windowing*

Suara yang dibagi menjadi beberapa potongan *frame*, membuat data atau sinyal suara menjadi *discontinue*. Hal tersebut mengakibatkan kesalahan data selama proses *fourier transform*, sehingga untuk menghindari terjadi kesalahan data pada proses *fourier transform* maka sampel suara yang telah dibagi menjadi beberapa *frame* perlu dijadikan suara kontinu dengan cara mengalikan tiap *frame* dengan *window* tertentu seperti ditampilkan pada persamaan 7.

$$(n) = x(n) \times w(n) \tag{7}$$

dimana,

n = *frame* ke-1, 2, 3, dan seterusnya,

$x(n)$ = sinyal setelah di-*frame blocking*,

$w(n)$ = sinyal setelah *windowing*.

2.8.4 RMS (*Root Mean Square*)

RMS adalah nilai akar kuadrat dari rata-rata jumlah kuadrat sinyal. Umumnya digunakan untuk mengevaluasi amplitudo sinyal dan energi sinyal dalam domain waktu (Patil et al., 2022). Untuk sinyal x(t), RMS dapat dihitung dengan persamaan 8.

$$MS = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(x(t)\right)^2} \tag{8}$$

dimana,

$x(t)$ = nilai pada set data,

N = jumlah elemen pada set data.

2.8.5 *Mel Spectogram*

Spektogram mel merupakan kolaborasi skala mel dan spektogram di mana skala mel mewakili transformasi non linier dari skala frekuensi. Dalam hal ini, sinyal audio dibagi menjadi beberapa *frame* yang lebih kecil dan *hamming window* diterapkan pada setiap frame (Das et al., 2020).

2.9 *Confusion Matrix*

Confusion Matrix digunakan untuk mengukur kinerja model dalam melakukan klasifikasi menggunakan data validasi untuk setiap jenis model pelatihan data. *Confusion Matrix* ini sangat berguna pada model *deep learning* terutama dalam masalah klasifikasi.

True Positives (TP) dan *True Negatives* (TN) adalah kasus di mana prediksi sesuai dengan kelas sebenarnya sedangkan *False Positives* (FP) *dan False Negatives* (FN) adalah kasus di mana prediksi tidak sesuai dengan kelas sebenarnya (Syaputra, 2023)

2.9.1 *Multiclass Confusion Matrix*

Multiclass confusion matrix adalah alat evaluasi yang digunakan dalam pemodelan dan pembelajaran mesin ketika memiliki lebih dari dua kelas pada klasifikasi. Contoh *multiclass confusion matrix* oleh Fokkema (2022) tentang cara menafsirkan *confusion matrix* untuk klasifikasi *multiclass* ditampilkan pada Gambar 7.

		PREDICTED classification			
	Classes	a	b	c	d
ACTUAL classification	a	TN	FP	TN	TN
	b	FN	TP	FN	FN
	c	TN	FP	TN	TN
	d	TN	FP	TN	TN

Gambar 7 *Multiclass confusion matrix* (Fokkema, 2022)

Pada Gambar 7, dimisalkan pada data kelas (b) diprediksi benar terhadap data aktual maka termasuk sebagai *True Positive* (TP) yang ditandai dengan warna merah muda kemudian jika terdapat data yang diprediksi sebagai kelas (b) terhadap data aktual selain kelas (b) maka termasuk sebagai *False Positive* (FP) yang ditandai dengan warna kuning sedangkan data yang diprediksi bukan kelas (b) terhadap data aktual kelas (b) maka termasuk sebagai *False Negative* (FN) yang ditandai dengan warna hijau dan *True Negative* (TN) adalah jika terjadi kondisi selain dari kondisi TP, FP, dan FN yang telah dijelaskan sebelumnya atau ditandai dengan warna abu-abu.

2.9.2 Metrik Evaluasi

Metrik evaluasi adalah ukuran kuantitatif yang digunakan untuk menilai kinerja model. Beberapa metrik evaluasi yang digunakan yaitu *accuracy, precision, recall,* dan *F1-Score* (Yao & Shepperd, 2020).

1. ***Accuracy***

   Akurasi yaitu persentase prediksi yang benar dari total sampel seperti pada persamaan 9.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

2. ***Precision***

   Presisi yaitu persentase prediksi positif yang benar dari total prediksi positif seperti pada persamaan 10.

$$\frac{TP}{TP + FP} \tag{10}$$

3. ***Recall***

   Recall yaitu persentase prediksi positif yang benar dari total aktual positif seperti pada persamaan 11.

$$\frac{TP}{TP + FN} \tag{11}$$

4. *F1-Score*

F1-score yaitu *harmonic mean* dari presisi dan *recall* seperti pada persamaan 12.

$$2\ x\left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (12)$$