Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 5 (2020) 799 - 804 ISSN Media Elektronik: 2580-0760

Deteksi Emosi Wicara pada Media On-Demand menggunakan SVM dan LSTM

Ainurrochman[1], Derry Pramono Adi[2], Agustinus Bimo Gumelar[3]
[1,2,3]Fakultas Ilmu Komputer, Universitas Narotama, Surabaya
[1]a1308.id@gmail.com, [2]derryalbertus@ieee.org, [3]bimogumelar@ieee.org

Abstract

To date, there are many speech data sets with emotional classes, but with impromptu or intentional actors. The native speakers are given a stimulus in each emotion expression. Because natural conversation from secretly recorded daily communication still raises ethical issues, then using voice data that takes samples from movies and podcasts is the most appropriate step to take the best insights from speech. Professional actors are trained to induce the most real emotions close to natural, through the Stanislavski acting method. The speech dataset that meets this qualification is the Human voice Natural Language from On-demand media (HENLO). Within HENLO, there are basic per-emotion audio clips of films and podcasts originating from Media On-Demand, a motion video entertainment media platform with the freedom to play and download at any time. In this paper, we describe the use of sound clips from HENLO, then conduct learning using Support Vector Machine (SVM) and Long Short-Term Memory (LSTM). In these two methods, we found the best strategy by training LSTMs first, then then feeding the model to SVM, with a data split strategy at 80:20 scale. The results of the five training phases show that the last accuracy results increased by more than 17% compared to the first training. These results mean both complement and methods are important for improving classification accuracy.

Keywords: Speech Emotion Detection, Media On-Demand, SVM, LSTM, Deep Learning

Abstrak

Hingga saat ini, terdapat banyak dataset wicara dengan kelas emosi, namun dengan aktor dadakan atau disengaja. Penutur asli diberi stimulus serta petunjuk emosi apa yang harus ditampakkan selanjutnya. Karena percakapan natural dari komunikasi sehari-hari masih menimbulkan isu etik ketika direkam secara diam-diam, maka, menggunakan data suara yang mengambil sampel dari film dan *podcast* adalah langkah paling tepat untuk mengambil *insight* terbaik dari wicara. Aktor profesional terlatih untuk menginduksikan emosi paling nyata mendekati natural, melalui metode akting Stanislavski. Dataset suara yang memenuhi kualifikasi ini adalah *Human voice Natural Language from On-demand media* (HENLO). Di dalam HENLO, terdapat klip audio per emosi dasar dari film dan *podcast* yang berasal dari *Media On-Demand*, yaitu *platform* media hiburan motion video dengan kebebasan *play* dan *download* kapanpun. Dalam makalah ini, kami menjabarkan penggunaan klip suara dari HENLO, lalu melakukan *learning* menggunakan *Support Vector Machine* (SVM) dan *Long Short-Term Memory* (LSTM). Pada dua metode tersebut, kami menemukan strategi terbaik dengan melatih LSTM terlebih dahulu, lalu selanjutnya memberi *feed model* ke SVM, dengan strategi *split data* di skala 80:20. Hasil dari lima kali fase *training* menunjukkan hasil akurasi terakhir meningkat lebih dari 17% dibandingkan dengan *training* pertama. Hasil ini berarti kedua metode komplemen dan penting untuk peningkatan akurasi klasifikasi.

Kata kunci: Deteksi Emosi Wicara, *Media On-Demand*, SVM, LSTM, *Deep Learning*

1. Pendahuluan

Emosi merupakan suatu faktor penting dalam komunikasi [1], [2]. Dalam kehidupan sehari-hari, hasil teks yang didikte secara sederhana tidak cukup mampu untuk mengungkapkan emosi. Sistem pengenalan emosi wicara dapat digunakan oleh penyandang disablitas untuk komunikasi [3], [4], oleh aktor untuk konsistensi wicara emosi serta untuk acara pada media televisi yang interaktif [5], untuk membangun model guru dalam bentuk virtual [6], juga digunakan dalam studi yang mendeteksi kerusakan otak pada manusia [7], dan desain canggih dari *speech embedding* [8], [9]. Salah satu contoh mengenai emosi lainnya adalah bahwa emosi mengatur kehidupan kita sehari-hari; emosi merupakan bagian besar dari pengalaman manusia dan mempengaruhi pengambilan keputusan kita [10]–[12]. Untuk mencapai tujuan ambisius seperti itu,

Diterima Redaksi : 21-06-2020 | Selesai Revisi : 02-09-2020 | Diterbitkan Online : 30-10-2020

pengumpulan *database* wicara emosional adalah prasyarat. Dalam penelitian ini, *database* wicara yang dimaksud adalah HENLO. Namun, *cost* untuk mendapatkan satu film terlalu besar, bertabrakan dengan misi untuk mendapatkan data sebanyak-banyaknya. Sehingga, diperlukan *platform* penyedia media hiburan yang lain.

Dataset suara berbasis emosi HENLO mengambil klip suara dari film dan *podcast* dari Media On-Demand (MOD), sebuah platform media hiburan yang memungkinkan pengguna untuk *play* dan *download* kapanpun. Keunggulan yang juga dimiliki MOD adalah satu kali berlangganan berarti akses penuh terhadap *database* film dan *podcast* milik MOD tersebut. Selain itu, akting oleh aktor profesional dalam film dalam menginduksikan emosi melalui metode Stanivlaski, dirasa sangat menguntungkan. Produksi film pada perusahaan besar juga memerlukan *mastering* dan *scoring* yang secara otomatis meminimalkan *noise* dari lingkungan.

Di penelitian terdahulu, *Support Vector Machine* (SVM) dan *Long Short-Term Memory* (LSTM) banyak digunakan untuk data suara, yang mana memperoleh hasil klasifikasi yang tinggi dan dalam fase *training* yang singkat. Dalam dunia *Deep Learning*, kebutuhan utama adalah: (1) banyaknya data, (2) kecepatan *training*, (3) tenaga pemrosesan yang tergolong sederhana sehingga mampu menangani data yang bahkan lebih besar [13], [14]. Hingga kini, HENLO memenuhi syarat pertama, dan SVM memenuhi syarat kedua, serta LSTM memenuhi syarat ketiga. SVM dan LSTM adalah dua metode terkenal untuk data berbasis *time-series*. Memproses sinyal audio pada skala tinggi memerlukan proses *windowing* yang memberikan informasi temporal, sehingga sangat tepat untuk menggunakan kedua metode tersebut.

Selain memahami persis apa itu emosi, banyak teori telah mengusulkan untuk mengklasifikasikannya ke dalam berbagai jenis. Di antara studi yang paling signifikan, diantara Plutchik dan Ekman:

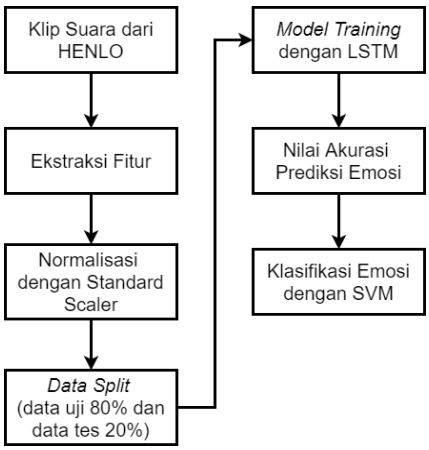
1. Paul Ekman adalah pelopor dalam studi emosi dan hubungannya dengan ekspresi wajah. Ekman mendefinisikannya sebagai emosi dasar: ketakutan, jijik, marah, terkejut, bahagia dan sedih [15], [16].
2. Robert Plutchik mengusulkan pendekatan klasifikasi psiko-evolusi untuk respons emosi umum. Plutchik menggambar "roda emosi" yang terkenal untuk menjelaskan usulannya menggunakan informasi yang dirupakan secara grafis (infografis), yang terdiri dari delapan emosi bipolar dasar: sukacita vs kesedihan, kepercayaan vs jijik, kemarahan vs ketakutan, dan kejutan vs antisipasi [17].

Makalah ini menjawab pertanyaan seputar keperluan klasifikasi emosi berdasarkan data suara yang memiliki keuntungan tipe komunikasi *natural*, namun tidak dengan *noise*, serta tipe komunikasi *acted* berdasarkan *script* yang mencontoh kehidupan sehari-hari. Simulasi emosi dengan aktor profesional terasa lebih *real*, dibandingkan simulasi emosi dengan penutur asli yang tidak memiliki talenta dalam induksi emosi dalam tekanan eksperimen.

Makalah ini disusun sebagai berikut: pada Bagian 1, dituliskan pendahuluan dan latar belakang penelitian. Bagian 2 menjabarkan metode yang digunakan dalam penelitian ini, sedangkan Bagian 3 menjelaskan penggunaan metode dalam kerangka kerja eksperimen yang runut. Akhirnya, Bagian 4 menyebutkan kesimpulan atas eksperimen ini.

2. Metode Penelitian

Pada eksperimen ini, kami menggunakan klip suara dari dataset publik HENLO yang menggunakan kelas emosi dari Plutchik [17]. Data suara tersebut kami lakukan ekstraksi fitur, normalisasi, lalu proses pembelajaran secara serial dari LSTM lalu diteruskan ke SVM.



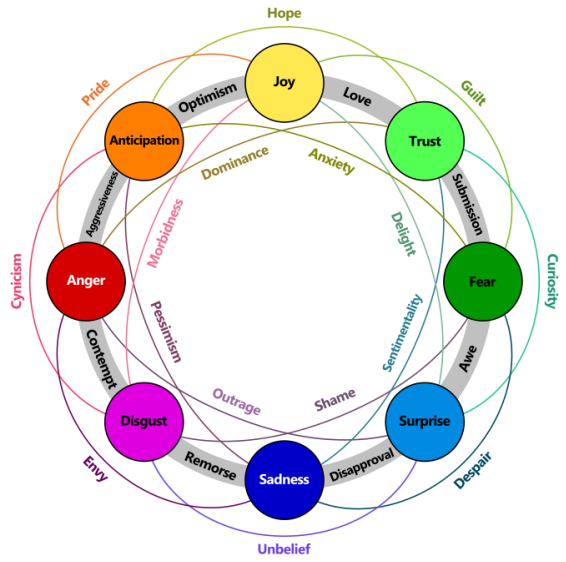
Gambar 1. Blok Diagram Alir Penelitian

Gambar 1 menjelaskan alir penelitian, dan sub bagian ini menguraikan setiap tahapan dari Gambar 1.

2.1. Klip Suara

Setiap klip suara diambil dari dataset publik HENLO. Dalam dataset ini, digunakan empat kelas emosi dasar, berdasarkan teori Plutchik, yaitu *Wheel of Emotion* [17]. Kelas emosi yang digunakan yaitu Takut, Senang, Sedih, dan Marah. HENLO memiliki 10.000 data klip suara untuk masing-masing kelas emosi.

Gambar 2 menunjukkan visualisasi *Wheel of Emotion* yang diinisiasi oleh Robert Plutchik [17]. Lingkaran dengan aneka ragam warna menunjukkan emosi dasar. Dalam HENLO, dipakai emosi dasar Marah, Senang, Sedih, dan Takut [2], [11].



Gambar 2. *Wheel of Emotion* milik Plutchik [2], [17]

Tabel 1 menunjukkan beberapa klip suara dari HENLO dengan kelas emosi Marah. Klip suara tersebut kemudian dipersiapkan untuk dilanjut ke proses ekstraksi fitur suara.

Tabel 1. Pratinjau data klip suara Marah dari HENLO

ID	*Path*	*Label*	*Source*
0	data/HENLO/Marah/Mar_275.mp3	angry	HENLO
1	data/HENLO/Marah/Mar_261.mp3	angry	HENLO
2	data/HENLO/Marah/Mar_249.mp3	angry	HENLO
3	data/HENLO/Marah/Mar_301.mp3	angry	HENLO
4	data/HENLO/Marah/Mar_315.mp3	angry	HENLO
5	data/HENLO/Marah/Mar_329.mp3	angry	HENLO
6	data/HENLO/Marah/Mar_88.mp3	angry	HENLO
7	data/HENLO/Marah/Mar_103.mp3	angry	HENLO
8	data/HENLO/Marah/Mar_117.mp3	angry	HENLO
9	data/HENLO/Marah/Mar_77.mp3	angry	HENLO
10	data/HENLO/Marah/Mar_63.mp3	angry	HENLO
11	data/HENLO/Marah/Mar_62.mp3	angry	HENLO
12	data/HENLO/Marah/Mar_76.mp3	angry	HENLO

2.2. Ektraksi Fitur Suara

Sundawa dkk berhasil mempergunakan fitur *Mel Frequency Cepstral Coeffecient* (MFCC) pada penelitiannya [18], dan berhasil mengenali empat dari tujuh kelas emosi dalam wicara. Anggraini dkk memakai frekuensi dan intensitas untuk alat pengenalan emosi yang dikembangkan, dan berhasil mengenali emosi berdasarkan tinggi/rendahnya frekuensi suara [19]. Walau keduanya tidak memberi nilai akurasi klasifikasi, namun penelitian tersebut membuktikan kemampuan fitur dasar MFCC dan intensitas dalam pengenalan emosi.

Setelah pengumpulan klip suara, proses selanjutnya adalah mengekstrak fitur suara dari klip HENLO. Diantara ribuan fitur suara [2], [20], kami menggunakan fitur dasar *Energy*, MFCC, *Pitch*, *Intensity*, *Jitter*, *Shimmer*, *dan Harmonics-to-Noise* (HNR). Tabel 2 menunjukkan fitur-fitur suara yang diambil dalam eksperimen ini. Total 38 fitur suara tersebut berbentuk nilai numerik, yang kemudian dimasukkan ke dalam *file* .csv.

2.3. Normalisasi dan *Data Split*

Setiap *Standard Scaler* menormalisasikan fitur dengan mengurangi rata-rata dan kemudian menskala ke varian unit [21], [22]. Varians unit berarti membagi semua nilai dengan standar deviasi. *Standard Scaler* menghasilkan distribusi dengan deviasi standar sama dengan satu varian unit sama dengan 1 juga, karena varians sama dengan standar deviasi kuadrat. *Standard Scaler* membuat rata-rata distribusi 0. Sekitar 68% nilai akan berada di antara -1 dan 1. Angka ini tetap bergantung pada jumlah dan variabilitas data. Algoritma *Machine Learning* sering menyerukan *zero mean* (rata-rata) dan varian unit.

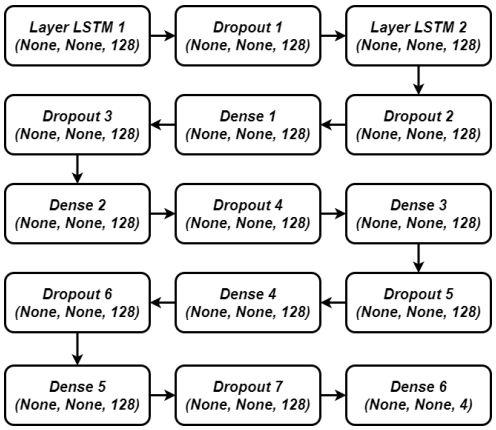
Tabel 2. Fitur Suara dalam Eksperimen

No	Fitur	Sub-Fitur	Fitur Statistik Global	*n*
1.	MFCC	-	-	20
2.	*Pitch*	*Strength*	▪ Rata-rata	5
			▪ Maks	
			▪ Min	
			▪ *Range*	
			▪ *Std. Dev*	
		Freq.	▪ Rata-rata	5
			▪ Maks	
			▪ Min	
			▪ *Range*	
			▪ *Std. Dev*	
3.	*Intensity*	-	▪ Rata-rata	5
			▪ Maks	
			▪ Min	
			▪ *Range*	
			▪ *Std. Dev*	
4.	*Jitter*	-	-	1
5.	*Shimmer*	-	-	1
6.	HNR	-	-	1
7.	*Energy*	-	-	1
			Jumlah Fitur Suara	39

Algoritma bertipe regresi juga mendapat manfaat dari data yang didistribusikan secara normal dengan ukuran sampel data yang lebih kecil. *Standard Scaler* mendistorsi jarak relatif antara nilai-nilai fitur, sehingga mengecilkan data sampel yang *imbalance*. Pada data HENLO, data sampel dirasa *imbalance* karena variabilitas yang tinggi, yaitu jumlah aktor yang setiap klip suara dapat berbeda-beda.

Kemudian, proses *data split* dilakukan dengan menggunakan *library* sklearn. Dalam proses ini, kami memecah data menjadi dua bagian, yaitu data untuk melatih mesin *(training data)* dan data untuk menguji mesin *(testing data)*, dengan skala masing-masing 80% dan 20%.

2.5. *Model* dalam *Long Short-Term Memory*



Gambar 3. Blok Diagram dari Arsitektur LSTM

Schmidhuber, pada penelitiannya berhasil mengembangkan modifikasi dari *Recurrent Neural Network* (RNN), yang diberi nama *Long Short-Term Memory* (LSTM). Hingga saat ini, LSTM telah dibuktikan mampu memecahkan masalah bervariasi pada data berupa *sequential* [23]–[25].

Tabel 3. *Log Summary* dari Arsitektur LSTM

Layer	***Output Shape***	***Param#***
lstm_1	(None, None, 128)	88.576
dropout_1	(None, None, 128)	0
lstm_2	(None, None, 128)	131.584
dropout_2	(None, None, 128)	0
dense_1	(None, None, 128)	16.512
dropout_3	(None, None, 128)	0
dense_2	(None, None, 128)	16.512
dropout_4	(None, None, 128)	0
dense_3	(None, None, 128)	16.512
dropout_5	(None, None, 128)	0
dense_4	(None, None, 128)	16.512
dropout_6	(None, None, 128)	0
dense_5	(None, None, 128)	16.512
dropout_7	(None, None, 128)	0
dense_6	(None, None, 4)	516
	Total Params	303.236
	Trainable Params	303.236
	Non-trainable Params	0

Model LSTM dalam eksperimen ini dibuat menggunakan *library* terkenal untuk pembelajaran mesin dan kecerdasan buatan, yaitu Keras. LSTM dalam penelitian ini memiliki dua *hidden layer*, lima *dense layer*, serta nilai 0,3 sebagai nilai *dropout*. Setiap *hidden layer*, termasuk *dense*, memiliki besaran jumlah unit 128. Gambar 3 dan Tabel 3 masing-masing menunjukkan arsitektur LSTM dalam bentuk blok diagram dan *log summary* pada eksperimen ini. Model LSTM kemudian "dilatih" menggunakan data latih HENLO berdasarkan empat kelas emosi. Data uji kemudian digunakan sebagai validasi untuk menghitung nilai akurasi model. Proses ini juga dikenal sebagai proses *fitting*.

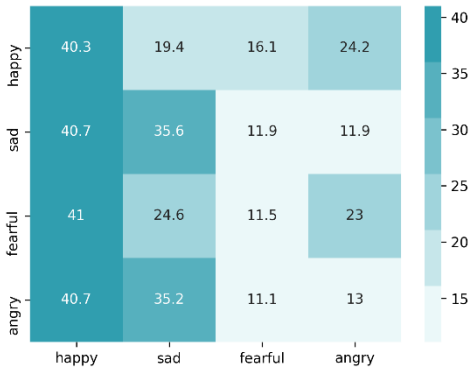
3. Hasil dan Pembahasan

Bagian ini menjelaskan hasil dari rancangan eksperimen, yaitu berupa nilai akurasi prediksi emosi pada LSTM dan SVM.

3.1. Klasifikasi Emosi dengan LSTM

Pada proses *fitting*, kami juga merencanakan proses learning bagi LSTM, sehingga memunculkan nilai akurasi prediksi per kelas emosi. Fungsi *"predict"* yang tertanam dalam LSTM, secara default mengambil data uji. Fungsi *"predict"* mengembalikan nilai prediksi berbasis probabilitas jenis emosi tiap suara dengan rentang nilai 0 sampai dengan 1. Proses ini dilakukan sebagai validasi dari model yang sebelumnya sudah dilatih.

Jenis emosi ditentukan berdasarkan nilai probabilitas tertinggi diasumsikan sebagai variabel y_pred. Setelah mendapatkan nilai y_pred, nilai y_pred dibandingkan dengan jenis emosi suara yang asli (dalam variabel y_test).



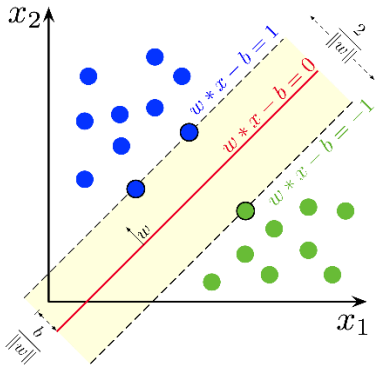
Gambar 4. Validasi LSTM dalam bentuk *Confusion Matrix*

Hasil perbandingan divisualisasikan menggunakan *Confusion Matrix* seperti pada Gambar 4. Perhitungan nilai akurasi model kemudian diperoleh dengan menghitung nilai rata-rata tingkat akurasi pada setiap jenis emosi.

3.2. Klasifikasi Emosi dengan SVM

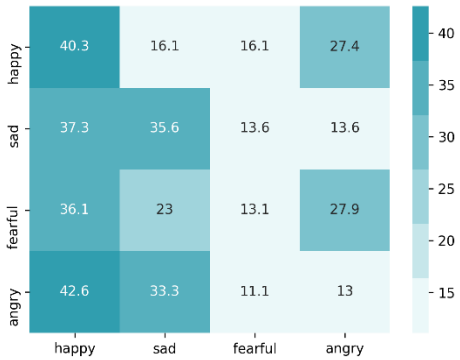
Vapnik mengembangkan *Support Vector Machine* (SVM) pada tahun 1992 [26], [27]. SVM adalah model algoritma *learning* analisis data yang menunjukkan performa tinggi untuk klasifikasi dan analisis regresi. Gambar 5 menunjukkan visualisasi SVM.

Setelah mendapatkan nilai probabilitas data uji y_pred, diperlukan dataset dengan fitur baru untuk mentraining fungsi klasifikasi SVM. Dataset baru ini adalah data yang berisikan fitur *utterance-level* (nilai rata-rata, nilai maksimal, nilai minimal dan nilai rata-rata di atas 0.2) dari nilai probabilitas prediksi emosi data train dan data uji (nilai probabilitas jenis emosi data train dan data uji diumpakan masing-masing sebagai X_DNN_train dan X_DNN_test).



Gambar 5. Visualisasi SVM [28]

Nilai X_DNN_train dapat diperoleh dengan memprediksi jenis emosi suara menggunakan model LSTM dengan X_train sebagai parameter, sedangkan nilai X_DNN_test dapat diperoleh melalui nilai y_pred.



Gambar 6. Visualisasi Akurasi Klasifikasi pada SVM

Selanjutnya dilakukan proses ekstraksi *fitur utterance-level* dengan menggunakan fungsi iterasi, kemudian dilanjutkan proses data latih pada metode klasifikasi SVM. Apabila proses training data sudah selesai, lalu dilanjutkan dengan membandingkan hasil prediksi jenis emosi suara metode klasifikasi SVM dengan jenis emosi suara yang asli (y_test). Mendekati dengan model LSTM, kami melakukan proses validasi akurasi dengan SVM melalui perbandingan hasil prediksi jenis emosi suara metode klasifikasi SVM dengan jenis emosi suara yang asli (y_test). Perbandingan ini dapat divisualisasikan dengan *Confusion Matrix*, seperti ditunjukkan pada Gambar 6.

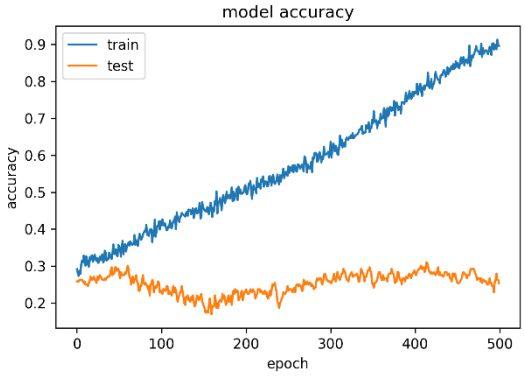
3.3. Nilai Akurasi Klasifikasi Emosi

Pada bagian ini, kami memberi tabel berisi angka akurasi pada setiap iterasi *learning*. Tabel 4 memberi angka akurasi pada masing-masing metode SVM dan LSTM.

Pada SVM dan LSTM, iterasi *learning* dengan akurasi tertinggi yaitu pada 47,4%. Gambar 7 menunjukkan model akurasi terbaik pada iterasi *learning* kelima, masing-masing untuk SVM dan LSTM. Sedangkan Gambar 8 menunjukkan model *loss* pada iterasi *learning* kelima.

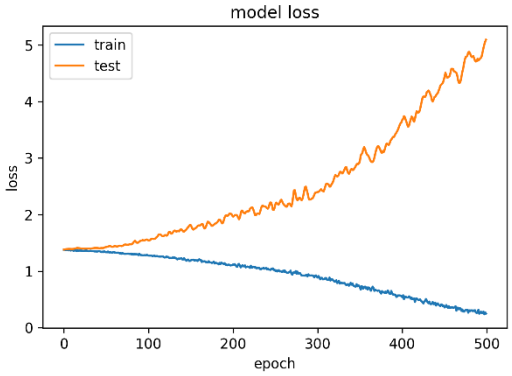
Tabel 4. Iterasi *Training* dan Nilai Akurasi

Training* ke-*n	**LSTM**	**SVM**
1	30%	30%
2	35,6%	35,6%
3	45,1%	45,1%
4	43,7%	43,7%
5	47,4%	47,4%

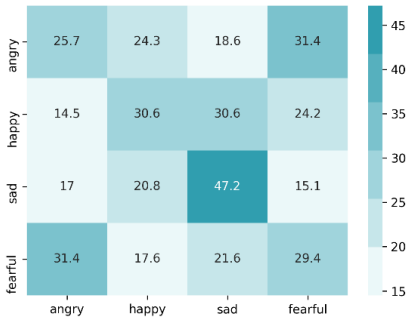


Gambar 7. Model Akurasi pada iterasi *learning* kelima

Gambar 9 adalah hasil validasi akurasi pada iterasi *learning* kelima, yang kami lakukan dengan menggunakan *Confusion Matrix*.



Gambar 8. Model *Loss* pada iterasi *learning* kelima



Gambar 9. Hasil Validasi dengan *Confusion Matrix*

4. Kesimpulan

Dalam eksperimen ini, kami melakukan proses identifikasi emosi wicara manusia menggunakan dataset HENLO. Dataset HENLO mempergunakan film dan podcast dari MOD, yang menguntungkan karena mengurangi *cost* dalam akuisisi data film. MOD memungkinkan data film diakuisisi dalam jumlah yang sangat banyak; berpengaruh langsung kepada jumlah klip suara yang akan di-*learning*.

SVM dan LSTM memberikan hasil yang cukup dalam pendeteksian emosi. SVM yang diberi *trained model* dalam LSTM ternyata memberi nilai akurasi yang lebih tinggi. Hasil *training* pertama pada SVM dan LSTM memberi angka 30%, sedangkan hasil *training* kelima pada SVM dan LSTM memberi angka 47,4%. Peningkatan dalam angka akurasi ini berarti performa SVM dan LSTM cukup besar serta berpengaruh klasifikasi emosi multi kelas.

Sebagai rekomendasi untuk penelitian selanjutnya, kami memilih metode *Data Augmentation* sebagai langkah yang harus dijalankan, untuk menghindari *overfitting* atau *underfitting*. Metode *Data Augmentation* juga dipilih untuk mengatasi variabilitas data (aktor) yang tinggi.

Daftar Rujukan

[1] R. M. Nesse, "Evolutionary Explanations of Emotions," *Hum. Nat.*, vol. 1, no. 3, pp. 261–289, Sep. 1990.

[2] A. B. Gumelar, Eko Mulyanto Yuniarno, Wiwik Anggraeni, Indar Sugiarto, A. A. Kristanto, and M. H. Purnomo, "Kombinasi Fitur Multispektrum Hilbert dan Cochleagram untuk Identifikasi Emosi Wicara," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 9, no. 2, pp. 180–189, May 2020.

[3] M. A. W. Martens, M. J. Janssen, W. A. J. J. M. Ruijssenaars, M. Huisman, and J. M. Riksen-Walraven, "Fostering Emotion Expression and Affective Involvement with Communication Partners in People with Congenital Deafblindness and Intellectual Disabilities," *J. Appl. Res. Intellect. Disabil.*, vol. 30, no. 5, pp. 872–884, Sep. 2017.

[4] N. Adibsereshki, M. Shaydaei, and G. Movallali, "The Effectiveness of Emotional Intelligence Training on the Adaptive Behaviors of Students with Intellectual Disability," *Int. J. Dev. Disabil.*, vol. 62, no. 4, pp. 245–252, Oct. 2016.

[5] K. An and M. Chung, "Cognitive Face Analysis System for Future Interactive TV," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 2271–2279, Nov. 2009.

[6] M. Uitto, K. Jokikokko, and E. Estola, "Virtual Special Issue on Teachers and Emotions in Teaching and Teacher Education (TATE) in 1985–2014," *Teach. Teach. Educ.*, vol. 50, pp. 124–135, Aug. 2015.

[7] K. Kucharska-Pietura, M. L. Philips, W. Gernand, and A. S. David, "Perception of Emotions from Faces and Voices Following Unilateral Brain Damage," *Neuropsychologia*, vol. 41, no. 8, pp. 1082–1090, Jan. 2003.

[8] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech Emotion Recognition Using Spectrogram and Phoneme Embedding," in *Interspeech 2018*, 2018, pp. 3688–3692.

[9] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," no. 7491, pp. 85–90, 2017.

[10] D. P. Adi, A. B. Gumelar, and R. P. Arta Meisa, "Interlanguage of Automatic Speech Recognition," in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2019, pp. 88–93.

[11] A. B. Gumelar *et al.*, "Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks," *IEEE 7th Int. Conf. Serious Games Appl. Heal.*, pp. 1–8, Aug. 2019.

[12] J. Ahmad, M. Sajjad, S. Rho, S. il Kwon, M. Y. Lee, and S. W. Baik, "Determining Speaker Attributes from Stress-affected Speech in Emergency Situations with Hybrid SVM-DNN Architecture," *Multimed. Tools Appl.*, vol. 77, no. 4, pp. 4883–4907, 2018.

[13] A. Huang and P. Bao, "Human Vocal Sentiment Analysis," pp. 1–16, 2019.

[14] X. Zhengqiao and Z. Dewei, "Research on Clustering Algorithm for Massive Data Based on Hadoop Platform," in *2012 International Conference on Computer Science and Service System*, 2012, pp. 43–45.

[15] K. R. Scherer, *Approaches To Emotion*. Psychology Press, 2014.

[16] P. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions*. Oxford University Press USA, 1994.

[17] R. Plutchik, "The Nature of Emotions: Human Emotions Have Deep Evolutionary Roots, a Fact That May Explain Their Complexity and Provide Tools for Clinical Practice," *Am. Sci.*, vol. 89, no. 4, pp. 344–350, 2001.

[18] A. A. Sundawa, A. G. Putrada, and N. A. Suwastika, "Implementasi Dan Analisis Simulasi Deteksi Emosi Melalui Pengenalan Suara Menggunakan Mel-frequency Cepstrum Coefficient Dan Hidden Markov Model Berbasis IoT," *eProceedings Eng.*, vol. 6, no. 1, 2019.

[19] N. A. Anggraini and N. Fadillah, "Analisis Deteksi Emosi Manusia dari Suara Percakapan Menggunakan Matlab dengan Metode KNN," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 3, no. 2, pp. 176–179, Mar. 2019.

[20] A. B. Gumelar, M. H. Purnomo, E. M. Yuniarno, and I. Sugiarto, "Spectral Analysis of Familiar Human Voice Based On Hilbert-Huang Transform," in *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, 2018, pp. 311–316.

[21] P. Ferreira, D. C. Le, and N. Zincir-Heywood, "Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection," in *2019 15th International Conference on Network and Service Management (CNSM)*, 2019, pp. 1–7.

[22] B. Lamichhane, U. Großekathöfer, G. Schiavone, and P. Casale, "Towards Stress Detection in Real-Life Scenarios Using Wearable Sensors: Normalization Factor to Reduce Variability in Stress Physiology," 2017, pp. 259–270.

[23] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, Nov. 2019.

[24] F. A. Gers, "Learning to Forget: Continual Prediction with LSTM," in *9th International Conference on Artificial Neural Networks: ICANN '99*, 1999, vol. 1999, pp. 850–855.

[25] S. An, Z. Ling, and L. Dai, "Emotional Statistical Parametric Speech Synthesis using LSTM-RNNs," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1613–1616.

[26] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting Time Series with Support Vector Machines," 1997, pp. 999–1004.

[27] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems*, 1997, pp. 155–161.

[28] Y. Chavhan, M. Dhore, and Y. Pallavi, "Speech Emotion Recognition Using Support Vector Machines," *Int. J. Comput. Appl.*, vol. 1, 2010.