Jurnal Mahajana Informasi, Vo.4, No.2, 2019
e-ISSN:2527-8290

DETEKSI UJARAN KEBENCIAN BERBASIS VIDEO DENGAN METODE MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC) - HIDDEN MARKOV MODEL (HMM) DAN CONVOLUTIONAL NEURAL NETWORK (CNN)

Sunaryo Winardi[1)], **Ng Poi Wong**[2)], **Tiartha Triagustinus Sitanggang**[3)], **Juangsyah Putra Nasution**[4)]
[1,2,3,4)]Program StudiTeknik Informatika,STMIK Mikroskil
Jl. Thamrin No.112, 124, 140, Medan 20212,Telp (061)-4573767
e-mail : [1)]sunaryo.winardi@mikroskil.ac.id, [2)]poiwong@mikroskil.ac.id, [3)]151112416@students.mikroskil.ac.id, [4)]151112637@students.mikroskil.ac.id

Abstrak

Peningkatan penyebaran konten kebencian di media sosial membutuhkan tindakan penanggulangan yang serius. Sejumlah metode telah dikembangkan untuk mendeteksi konten kebencian secara otomatis dengan tujuan mengklasifikasikan konten tekstual sebagai ujaran kebencian atau bukan. Dalam penelitian ini, pendeteksian ujaran kebencian akan dilakukan pada konten video dengan model isolated word recognition. Model ini hanya dapat mendeteksi kata, bukan kalimat sehingga harus dilakukan pemotongan kalimat menjadi kata menggunakan metode silence split. Metode Mel Frequency Cepstral Coefficients (MFCC) - Hidden Markov Model (HMM), dan Convolutional Neural Network (CNN) digunakan untuk mengklasifikasikan konten video sebagai ujaran kebencian. Pengujian pada penelitian ini terdiri dari 2 bagian, yakni pengujian speech to text menggunakan metode word error rate (WER) dan menghasilkan WER sebesar 9.433% untuk data testing perempuan dan WER sebesar 7.54% untuk data testing laki-laki, serta pengujian text classification menggunakan metode confusion matrix dan mendapatkan nilai akurasi sebesar 88%.

Kata Kunci: Mel Frequency Cepstral Coefficients, Hidden Markov Model, Convolutional Neural Network, Ujaran Kebencian

1. PENDAHULUAN

Kejahatan ujaran kebencian (*Hate Speech*) adalah tindakan komunikasi yang dilakukan oleh suatu individu atau kelompok dalam bentuk provokasi, hasutan, ataupun hinaan kepada individu atau kelompok yang lain dalam hal berbagai aspek seperti ras, warna kulit, gender, cacat, orientasi seksual, kewarganegaraan, agama, dan lain-lain (Febriyani, et al, 2018). Perkembangan kasus ujaran kebencian di Indonesia telah meningkat pesat dari tahun ke tahun, tercatat 3.325 kasus pada tahun 2017, angka tersebut naik 44,99% dari tahun sebelumnya yang berjumlah 1.829 kasus(Medistiara Y, 2017). Dari angka tersebut, perlu adanya tindakan untuk mengurangi penyebaran ujaran kebencian, salah satunya adalah melalui media video.

Hal tersebut sudah dilakukan dengan adanya kebijakan yang dibuat penyedia video dari anak perusahaan google yaitu *youtube* tentang ujaran kebencian. Oleh karena itu, dibutuhkan sistem pendeteksi ujaran kebencian untuk memfilter video yang diunggah untuk mengurangi penyebaran ujaran kebencian. (Chen C Y, 2014). Pada penelitian ini digunakan metode *Silence Split* untuk pemotongan kalimat menjadi kata (Montacie M & Caraty M J, 1998) dan model *Isolated Word Recognition* untuk mendeteksi kata (Huang, et al, 2001). Proses *speech to text* akan menggunakan metode *Hidden Markov Model* (HMM), kemudian ciri-ciri suara diambil dengan metode ekstraksi ciri *Mel Frequency Cepstral Coefficients* (MFCC).

Selanjutnya adalah proses klasifikasi teks dengan menggunakan *Convolution Neural Network* (CNN).

2. TINJAUAN PUSTAKA

Beberapa penelitian deteksi ujaran kebencian yang pernah dilakukan diantaranya penelitian deteksi ujaran kebencian berbasis teks pada media *twitter* dengan algoritma *Random Forest Decision Tree* diperoleh *F-measure*sebesar 93.5% (Alfina, et al, 2017). Penelitian deteksi ujaran kebencian pernah ditingkatkan menjadi berbasis gambardengan metode *Optical Character Recognition* dan diperolehtingkat akurasi sebesar 96%(Putra, et al., 2018).

Mel Frequency Cepstral Coefficients(MFCC) merupakan suatu metode yang digunakan dalam pemrosesan sinyal suara sehingga sinyal suara yang diolah memiliki ciri-ciri tertentu yang dapat dibedakan oleh sistem, dimana ciri-ciri yang dihasilkan adalah berupa koefisien (Hasan, et al, 2004). MFCC mengadopsi cara kerja dari organ pendengaran manusia sehingga mampu untuk menangkap karakteristik suara yang sangat penting (Tychtl Z.& Psutka J., 1999).

Hidden Markov Model (HMM) merupakan model statistik untuk memodelkan parameter tersembunyi dari suatu sistem. Pemodelan sinyal ucapan dilakukan melalui proses estimasi berulang pada

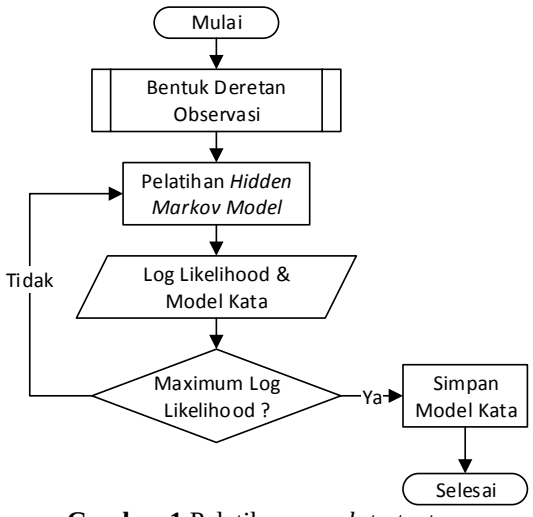
parameter HMM sehingga akhirnya didapat satu set parameter HMM yang memodelkan suatu kata. Proses pengenalannya dilakukan dengan mencari kata yang memberikan peluang terbesar terhadap kemunculan urutan pengamatan tersebut pada model HMM (Jalan, et al, 2013).MFCC dinilai sangat cocok dengan HMM karena berhasil mendapatkan tingkat akurasi paling tinggi yaitu 92% saat dilakukan perbandingan dengan metode sejenis lainnya (Abushariah, et al., 2010).Kombinasi metode HMM dan MFCC terbukti dapat menurunkan waktu komputasi dan meningkatkan akurasi (Patel I.& Rao Y. S., 2010). Algoritma CNN mampu mendapatkan akurasi paling tinggi yaitu sebesar 81,5% dari algoritma klasifikasi teks lainnya seperti Multinomial Naif Bayes (MNB) sebesar 79,0% dan Naive Bayes Support Vector Machine (NBSVM) sebesar 79,4% (Kim Y., 2014).

3. METODE PENELITIAN

Deteksi ujaran kebencian pada penelitian ini dilakukan melalui 2 tahapan, dimulai dari *Speech to Text* dan *Text Classification*.

Pada *speech to text*(Gambar 1) berfungsi untuk mengambil ciri suara dan melatih model HMM untuk setiap kata. Proses dimulai dari membentuk deretan observasi yang didapatkan dengan dilakukan perekaman suara sebagai data training. Sample suara akan dinormalisasi dan melewati proses *Silent Removal* agar bentuk sinyal suara lebih kecil yang dapat mempercepat proses komputasi. Proses normalisasi dilakukan agar pada pemrosesan sinyal selanjutnya tidak dipengaruhi oleh amplitudo sinyal yang terlalu besar atau terlalu kecil dengan rumussebagai berikut.

$$s(n)_{norm} \frac{s[n]}{|max\,(s(n))|} \quad (1)$$



Gambar 1. Pelatihan*speech to text*

Proses *silence removal* dilakukan untuk menghilangkan daerah *silence* dari sinyal suara untuk meningkatkan akurasi sistem. Proses ini dilakukan dengan mencari nilai standar deviasi dari sinyal suara menggunakan rumus sebagai berikut.

$$std = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_{i-\bar{x}})^2\right)^{1/2} \quad (2)$$

Berikutnya dilakukan ekstraksi ciri menggunakan MFCC yang terdiri dari proses*Pre-emphesis, Framing, Windowing, Fast Fourier Transform, Filter Bank, Discreate Cosine Transform,* dan *Cepstral Lifering.*

*Pre-emphasis*untuk menyaring sinyal suara agar sinyal menjadi lebih jelas ketika masuk ke tahap selanjutnya dengan rumus sebagai berikut.

$$y[n] = x[n] - a * x[n-1] \quad (3)$$

Berikutnya *Framing* dilakukandengan menggunakan rumus*frame blocking* sebagai berikut.

$$isi\ frame = \frac{fs}{0.025} \quad (4)$$

$$Overlap = \frac{fs}{0.010} \quad (5)$$

$$banyak\ frame = \frac{panjang\ data - isi\ frame}{isi\ frame - overlap} \quad (6)$$

Windowing untuk mengurangi terjadinya kebocoran spektral atau aliasing yang merupakan efek dari timbulnya sinyal baru yang memiliki frekuensi yang berbeda dengan sinyal aslinya, dengan rumus sebagai berikut.

$$X_n = S_{i[n]}W_n \quad (7)$$

Window yang digunakan pada penelitian ini adalah *Hamming window* dengan rumus matematis sebagai berikut.

$$w[n] = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N-1}\right) \quad (8)$$

dimana n = 0,1,2,…,n-1

Fast Fourier Transform (FFT)untuk mendapatkan sinyal dalam domain frekuensi dari sebuah sinyal diskritdengan rumus sebagai berikut.

$$X(k) = \frac{1}{n}\sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi kn}{N}} \quad (9)$$

Proses perhitungan untuk membentuk *filter bank* diawali dengan menentukan *lower* dan *upperfrequency*. Jika sinyal suara memiliki frekuensi sampling 16000, maka nilai *lower* dan *upperfrequency* yang mungkin adalah 300 Hz dan 8000 Hz. Jika sinyal suara memilki frekuensi sampling sebesar 16000, maka nilai *upper frequency* terbatas hanya sampai 8000 Hz.

Skala *Mel-Frequency* adalah frekuensi yang linier di bawah 1 kHz dan logaritmik di atas 1 kHz. Skala Mel dapat diperoleh dengan rumus sebagai berikut.

$$m = 2595\ log_{10}\left(\qquad (10)$$

Rumus untuk kembali ke frekuensi dari skala Mel adalah sebagai berikut.

$$f = 700\left(10^{m/2595} - 1\right) \qquad (11)$$

Kemudian digunakan rumus untuk membuat penampung sementara *filter bank* sebagai berikut.

$$f(i) = floor\left(\frac{(NFFT+1)*h(i)}{fs}\right) \qquad (12)$$

Kemudian masuk pada tahap membuat *filter bank* dengan menggunakan rumus sebagai berikut.

$$H_m(k) \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \qquad (13)$$

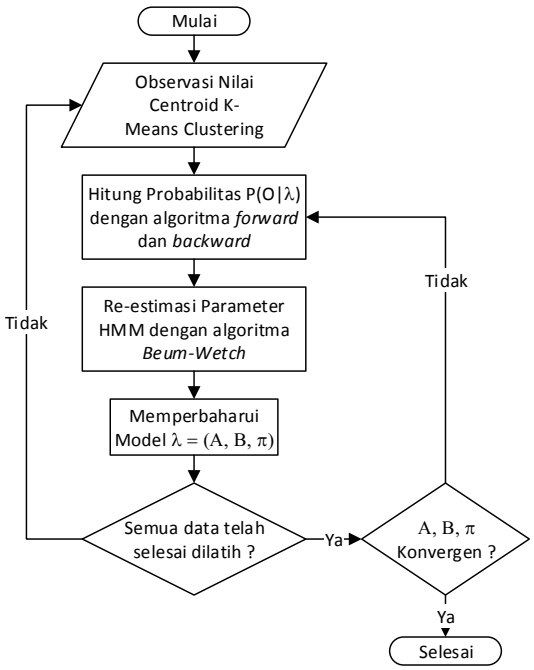
*Mel-Frequency Cepstral*kemudian didapatkan dari invers *Discrete CosineTransform* (DCT) untuk mendapatkan kembali sinyal dalam domain waktu. Hasilnya disebut sebagai *Mel-Frequency Cepstral Coefficient*(MFCC). MFCC bisa didapat dari pendekatan rumus sebagai berikut.

$$c_k = 2\sum_{n=0}^{N-1} x[n]cos\left[\qquad (14)$$

Dengan *x[n]* adalah hasil akumulasi dari kuadratik magnitude FFT yang dikalikan dengan *Mel-Filter Bank*. Setelah itu didapatkanlah MFCC. Pada sistem pengenalan suara, umumnya hanya 13 koefisien*cepstrum* pertama yang digunakan.
Hasil dari DCT merupakan *cepstrum* namun masih terdapat beberapa kelemahan,maka dilakukan proses *cepstral liftering* untuk memperhalus *cepstrums* dan meningkatkan akurasi sistem dengan rumus sebagai berikut.

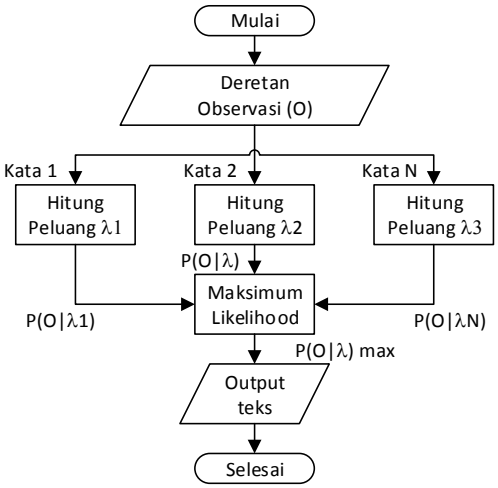
$$C = 1 + (L/2) \times \sin(\pi \qquad (15)$$

Langkahterakhiruntuk mendapatkanderetobservasi yaitu dengan Algoritma *K-Means Clustering*untuk memperkecil data ekstraksi ciri menjadi klaster sesuai dengan nilai yang berdekatan. Data diperkecil agar dapat mengurangi proses komputasi dan sebagai masukan nilai observasi pada HMM. Nilai koefisien MFCC yang sudah didapat akan dikelompokkan menjadi 5 klaster (K = 5) untuk setiap dimensi dan maksimal iterasi akan ditentukan sebanyak 300 iterasi agar mendapat nilai *centroid* terbaik.



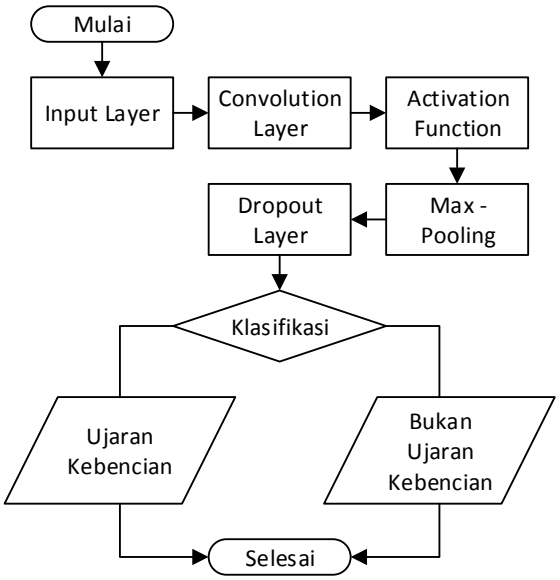
Gambar 2. Pelatihan HMM

Tahapan*training* HMM (Gambar 2) dilakukan dengan nilai probabilitas (π,A,B) yang dimodelkan dengan menggunakan notasi λ=(A,B, π). Probabilitas transisi (A), probabilitas simbol observasi (B), dan probabilitas initial state (π). Memiliki elemen pembentuk yaitu jumlah state tersembunyi (N), dan jumlah simbol observasi (M). Tahap training pada HMM dilakukan dengan 2 proses yakni tahap *Evaluation*dan *Learning*.Pada tahap *Evaluation*, dilakukan dengan menggunakan algoritma*forward-backward*, dimana merupakan penggabungan dari algoritma*forward* dan algoritma*backward*. Pada algoritma *forward*, dihitung variabel sepanjang deretan observasi, dan pada algoritma *backward*,dihitung variabel mundur secara berulang ke belakang sepanjang deretan observasi.Pada tahap *Learning* dilakukan estimasi parameter modelλ=(A,B, π)dengan menggunakan metode *Baum-welch.*
Gambar 3 menggambarkan pencarian nilai maksimum *likelihood*dari beberapa log yang telah didapatkan.Proses pencocokan dilakukan dengan mencari nilai *likelihood* terhadap semua model λ dengan menggunakan algoritma *forward* yang sama saat waktu proses pelatihan. Nilai yang diambil adalah nilai *likelihood* yang paling besar.



Gambar 3. Pencarian *maximum likelihood*

Pada tahapan *text classification*untuk membersihkan data latih dan mengubahnya menjadi vektor, yang dimulai dari mengambil dataset yang bersumber dari *twitter* sebagai data training.Data *tweets* yang didapatkan berbentuk format *txt* yang sudah difilter, dimana data *tweets* tersebut berisikan kalimat ujaran kebencian, kemudian diberi label pada setiap kalimat.Label tersebut digunakan untuk scoring pada data, dimana untuk data yang mengandung ujaran kebencian diberikan nilai (1) dan untuk data yang tidak mengandung ujaran kebencian diberikan nilai (0).Data yang didapatkan dari *twitter* dan telah ditentukan adalah data yang belum siap diolah sehingga perlu dilakukan *pre-processing*dengan dilakukan *Case folding, Remove Punctuation, Tokenizing, Stopword Removal,* dan *Stemming.Case folding* merupakan proses untuk mengubah semua huruf pada teks menjadi huruf kecil. *Remove Punctuation* untuk penghilangan tanda baca dan simbol pada kalimat. *Tokenizing* merupakan proses pemisah kalimat dengan karakter spasi menjadi token-token berupa kata. *Stopword removal* untuk menghapus kata yang tidak penting dan tidak memiliki arti berdasarkan kamus *Stopword*. *Stemming*untukmengubah token-token yang dihasilkan menjadi kata dasar.Berikutnya dilakukan proses vektorisasi dengan menggunakan *word2vec*untuk pembobotan dari data yang diproses menjadi bentuk vektor sehingga dihasilkan dataset untuk proses *Convolutional Neural Network* (CNN). Input yang digunakan ke dalam CNN adalah berupa kalimat atau dokumen yang direpresentasikan sebagai matriks (*matrix filter*). Setiap baris dari matriks sesuai dengan 1 token (kata), tetapi bisa juga karakter yang setiap barisnya adalah vektor yang merepresentasikan sebuah kata.Pengklasifikasian data dilakukan dengan metode *Convolutional Neural Network* (CNN) seperti gambar 4.



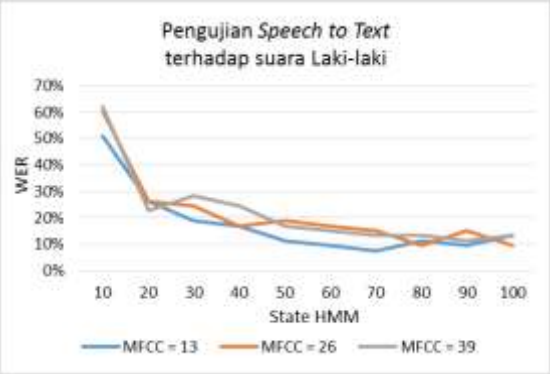
Gambar 4. Proses*text classification*

4. HASIL DAN PEMBAHASAN

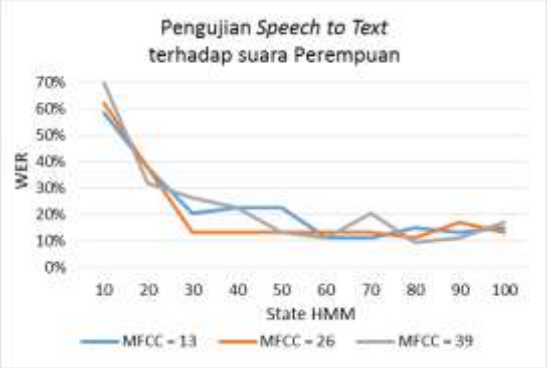
Pengujian dilakukan terhadap *speech to text* dengan metode *word error rate (WER)*, dan *text classification* dengan metode *confusion matrix* untuk menghitung akurasi.

Pengujian *speech to text* dilakukan terhadap 20 video, yang terdiri dari 10 video dengan suara laki-laki, dan 10 video dengan suara perempuan, dengan setiap video mengandung 1 kalimat, dandari 10 video dari masing-masing gender mengandung total 53 kata. Video tersebut adalah berformat mp4 yang direkam manual di dalam ruangan tertutup dengan aturan 50dB noise untuk video bersuara laki-laki dan 40dB noise untuk video bersuara perempuan, dimana perbedaan noise disebabkan oleh pengambilan video yang diambil di waktu dan tempat yang berbeda. Pengujian *speech to text* dilakukan sebanyak 60 kali percobaan sesuai kombinasi parameter berupa gender terdiri dari laki-laki dan perempuan, Koefisien MFCC terdiri dari 13, 26, dan 39, serta State HMM terdiri dari 10, 20, 30, 40, 50, 60, 70, 80, 90, dan 100.Gambar 5 dan gambar 6 menunjukkan bahwa semakin banyak state HMM yang digunakan maka WER yang didapatkan cenderung akan semakin rendah, hal ini terjadi karena semakin banyak state HMM maka model yang dibentuk untuk setiap kata akan lebih beragam sehingga kata akan lebih mudah dikenali.Dari perbedaan gender antara suara laki-laki dan perempuan menunjukkan hasil yang hampir sama. Hasil pengujian terhadap koefisien MFCC menunjukkan bahwa banyaknya koefisien MFCC tidak mempengaruhi WER yang dihasilkan. Dari hasil pengujian dengan koefisien MFCC = 13 diperoleh bahwa nilai WER terendah berada di antara 60 s/d 70 state HMM, pada koefisien MFCC = 26 diperoleh bahwa nilai WER terendah berada di antara 80 s/d 100 state HMM, dan pada koefisien

MFCC = 39 diperoleh bahwa nilai WER terendah berada di antara 80 s/d 90 state HMM.



Gambar 5. Pengujian *speech to text* terhadap suara Laki-laki



Gambar 6. Pengujian *speech to text* terhadap suara Perempuan

Pengujian *text classification* dilakukan dengan mengambil 25 video bersuara gender laki-laki, koefisien MFCC = 13, dan 70 state HMM, dengan setiap video mengandung 1 kalimat. Pengujian dilakukan sebanyak 5 kali dengan parameter jumlah matrix filter terdiri dari 2, 3, 4, 5, dan 6. Tabel 1 menunjukkan bahwa semakin kecil ukunran matrix filter maka akan menghasilkan akurasi yang lebih tinggi.

Tabel 1. Pengujian *text classification*

Matrix Filter	TP	TN	FP	FN	Akurasi
2	7	3	0	15	88%
3	8	2	4	11	76%
4	6	4	5	10	64%
5	10	5	7	3	52%
6	6	9	7	3	36%

5. KESIMPULAN

Dari hasil pengujian speech to text dan text classification dalam deteksi ujaran kebencian berbasis video dengan metode MFCC-HMM dan CNN, dapat disimpulkan bahwa :

1. Semakin banyak state HMM yang digunakan maka nilai WER yang diperoleh akan cenderung semakin rendah, dimana WER terendah dapat diperoleh dengan kisaran 60 s/d 100 state HMM.
2. Semakin besar ukuran matrix filter maka semakin rendah tingkat akurasi yang diperoleh, dimana tingkat akurasi tertinggi dapat diperoleh dengan ukuran matrix filter adalah sebesar 2.

DAFTAR PUSTAKA

Abushariah A. A. M., GunawanT. S., KhalifaO. O., AbushariahM. A. M., 2010, *English Digits Speech Recognition Based on Hidden Markov Models*, International Conference on Computer and Communication Engineering

Alfina I., Mulia R., Fanany M. I., EkanataY., 2017, *Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study*, 9th InternationalConference on Advanced ComputerScience and Information Systems 2017(ICACSIS)

Chen C. Y., 2014, *Fighting Online Hate Speech*, https://publicpolicy.googleblog.com/2014/09/fighting-online-hate-speech.html

Febriyani M., Sunarto, Husin B. R., 2018, *Analisis Faktor Penyebab Pelaku Melakukan Ujaran Kebencian (Hate Speech) dalam Media Sosial*, Jurnal Poenale, Universitas Lampung, Vol. 6, No. 3

Hasan M. R., Jamil M., Rabbani M. G., Rahman M. S., 2004, *Speaker Identification using Mel Frequency Cepstral Coefficients*, 3rd International Conference on Electrical & Computer Engineering (ICECE).

Huang X., Acero A., Hon H. W., 2001, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, USA

Jalan L., Masrani R., Jadhav R., Palav T., 2013, *Speech Recognitiion Based Learning System*, International Journal of Engineering Trends and Technology, Vol. 4, No. 2.

Kim Y., 2014, *Convolutional Neural Networks for Sentence Classification*, Conference on Empirical Methods in Natural Language Processing (EMNLP)

Medistiara Y., 2017, *Selama 2017 Polri Tangani 3.325 Kasus Ujaran Kebencian*, https://news.detik.com/berita/d-3790973/selama-2017-polri-tangani-3325-kasus-ujaran-kebencian

Montacie C., Caraty M. J., 1998, *A Silence / Noise / Music / Speech Splitting Algorithm*, 5th International Conference on SpokenLanguage Processing (ICSLP 98)

Patel I., Rao Y. S., 2010, *Speech Recognition using HMM with MFCC – An Analysis using Frequency Spectral Decomposion Technique*, Signal & Image Processing : An International Journal (SIPIJ), Vol. 1, No. 2.

Putra B. P., Irawan B., Setianingsih C., 2018, *Deteksi Ujaran Kebencian dengan*

Menggunakan Algoritma Convolutional Neural Network pada Gambar, e-Proceeding of Engineering, Universitas Telkom, Vol. 5, No. 2

Tychtl Z., Psutka J., 1999, *Speech Production Based on the Mel-Frequency Cepstral Coefficients*, Speech Communication and Technology