Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

Identifikasi Ujaran Kebencian Multilabel Pada Teks Twitter Berbahasa Indonesia Menggunakan Convolution Neural Network

Aditya Perwira Joan Dwitama, Syarif Hidayat*

Fakultas Teknologi Industri, Program Studi Magister Informatika, Universitas Islam Indonesia, Yogyakarta, Indonesia
Email: [1]aditya.dwitama@students.uii.ac.id, [2,*]syarif@uii.ac.id
Email Penulis Korespondensi: syarif@uii.ac.id
Submitted: **09/12/2021**; Accepted: **15/12/2021**; Published: **31/12/2021**

Abstrak−Besarnya jumlah pengguna media sosial mengakibatkan peningkatan aktivitas komunikasi antar warganet dalam media daring. Sebagai contoh, pengguna Twitter dapat berkomunikasi melalui twit. Twit dapat mengandung makna negatif. Hal ini perlu mendapatkan perhatian khusus karena berpotensi mengandung ujaran kebencian. Bahkan pemerintah merasa perlu menerbitkan peraturan khusus untuk menangani kasus ujaran kebencian seperti Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) yang dikeluarkan pada tahun 2018 pasal 28 ayat 2 tentang ujaran kebencian. Machine Learning (ML) adalah salah satu teknik yang bisa digunakan dalam mengidentifikasi suatu pola. ML dapat digunakan pada berbagai jenis data yang salah satunya adalah teks (dikenal sebagai text analytic). Penelitian sebelumnya telah mengidentifikasi ujaran kebencian pada teks Twitter dengan label yang lebih dari satu (multilabel) menggunakan metode Support Vector Machine (SVM). Penelitian ini dilakukan untuk mengidentifikasi ujaran kebencian pada teks Twitter dengan label yang lebih dari satu (multilabel) menggunakan metode Convolutional Neural Network (CNN) . Penelitian ini berhasil mendapatkan model CNN terbaik dengan angka akurasi sebesar 98,76% dari dataset multilabel ujaran kebencian pada teks bahasa Indonesia.

Kata Kunci: Bahasa Indonesia; Cnn; Machine Learning; Twitter; Ujaran Kebencian

Abstract−There has been a significant increase in communication activities between internet users in online media due to the increase in social media users. For instance, Twitter users may send messages via their tweets. However, tweets can also contain negative meanings. Therefore, it deserves special attention as it has the potential to contain hate speech. Even the government deems it necessary to publish regulations to deal with hate speech cases such as the Information and Electronic Transactions Law (ITE Law) issued in 2018 Article 28 paragraph 2 of the Hate Speech. Machine Learning (ML) is one of the techniques that can be used in identifying patterns. There are various types of data that ML can be applied to, including text (known as Text Analytic). Previous research has used the Support Vector Machine (SVM) method to identify hate speech on Twitter text with more than one label (multilabel). The purpose of this study was to identify hate speech on Twitter with a label of more than one (multilabel) via Convolutional Neural Network (CNN). The study obtained the best CNN model with an accuracy of 98.76% from the multi-label dataset on hate speech in Indonesian texts

Keywords: Indonesian; CNN; Machine Learning; Twitter; Hate Speech

1. PENDAHULUAN

Pengguna internet di Indonesia sudah mencapai jumlah 202,6 juta pengguna atau sekitar 73,7% dari total jumlah penduduk yang ada di Indonesia [1]. Dimana sebagian besar dari penggunaan internet tersebut ditujukan untuk beraktivitas di media sosial [1]. Hal ini ditunjukkan oleh jumlah pengguna media sosial yang mencapai 170 juta pengguna atau sekitar 83,9% dari total pengguna internet di Indonesia [1]. Adapun lima media sosial yang paling sering digunakan oleh rentang pengguna dengan umur 16-64 tahun adalah YouTube, WhatsApp, Instagram, Facebook, dan Twitter [1].

Media sosial menjadi media yang menyediakan sarana bagi warga internet atau biasa disebut dengan sebutan warganet untuk berkomunikasi secara daring. Misalkan Twitter, warganet bisa berkomunikasi antar sesamanya melalui cuitan-cuitan yang dilontarkan di Twitter. Cuitan-cuitan ini bisa memiliki dua tipe; ada yang bersifat positif dan ada pula yang bersifat negatif. Komentar yang negatif menjadi masalah karena biasanya mengandung unsur dari ujaran kebencian dan bisa berakibat sanksi hukum bagi penulisnya [2].

Direktorat Siber Bareskrim Polri telah mencatat bahwa sebanyak 125 akun media sosial sudah mendapat teguran oleh *virtual police* terkait dengan konten yang terindikasi mengandung unsur ujaran kebencian. Dari akun-akun yang mendapat teguran tersebut, Twitter memiliki angka terbesar yakni sebanyak 79 akun. Angka ini tercatat dalam periode 23 Februari sampai dengan 11 Maret 2021 [3].

Pemerintah dalam mencegah dan mengatasi permasalahan terkait dengan ujaran kebencian telah menerbitkan peraturan perundang-undangan dalam wujud UU ITE. Pada UU ITE Pasal 28 ayat 2 disebutkan bahwa warganet dilarang untuk menyebarkan informasi untuk menimbulkan rasa kebencian [2]. Selain itu, penjelasan mengenai kasus-kasus ujaran kebencian ini pernah diperjelas dalam Surat Edaran (SE) Kapolri No. SE/06/X/2015 mengenai cakupan-cakupan dari bentuk ujaran kebencian yang dapat diberikan pada konten media sosial [4]. Namun aturan ini dirasa masih perlu pembenahan terutama pada UU ITE. UU ini masih memiliki frase yang bersifat multitafsir yaitu pada frase "menyebarkan informasi" dan "rasa kebencian". Untuk itu perlu dilakukan penyusunan ulang terkait dengan kualifikasi dan ruang lingkup dari ujaran kebencian yang ada pada suatu konten media sosial [5].

Machine Learning dalam sub bagiannya yaitu *text analytic* memiliki algoritma-algoritma yang dapat melakukan pengenalan atau pengelompokan terhadap suatu objek text. *Text analytic* ini dapat dimanfaatkan di

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

dalam mengatasi kasus ujaran kebencian dalam media sosial melalui kemampuannya dalam mendeteksi *cyberbullying*, bahasa kasar, maupun *cyberhate* [6].

Penelitian mengenai ujaran kebencian dalam bahasa Indonesia pernah dibahas dalam penelitian yang berjudul "Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines". Penelitian ini menggunakan objek penelitian berupa *text* yang diambil dari Twitter. Akurasi yang dihasilkan pada penelitian ini adalah sebesar 74,88% [7]. Selain itu, penelitian mengenai ujaran kebencian juga pernah dilakukan dengan judul "Hate Speech Detection and Racial Bias Mitigation in Social Media Based On BERT Model". Penelitian ini menggabungkan metode BERT dan Convolution Neural Network (CNN) untuk mengklasifikasi teks berbahasa Inggris dengan hasil akhir F1 sebesar 92% [8].

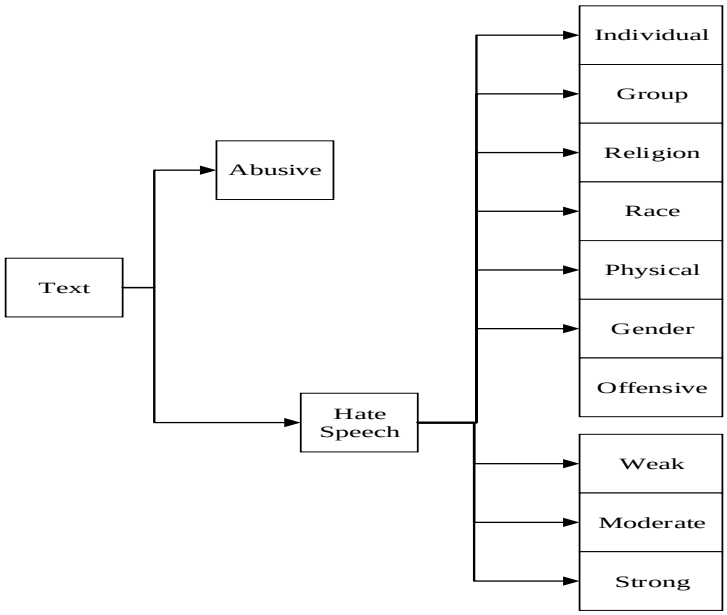
Berdasarkan pembahasan tersebut, penelitian kali ini akan dilakukan untuk membangun model *Machine Learning* menggunakan metode CNN untuk melakukan klasifikasi terhadap ujaran kebencian pada teks Twitter. Penelitian ini diharapkan mampu memberikan model terbaik dengan performa yang baik dalam melakukan klasifikasi sehingga dapat menjadi acuan untuk pengembangan di penelitian-penelitian selanjutnya.

2. METODOLOGI PENELITIAN

2.1 Literatur Review

Berdasarkan pembahasan tersebut, penelitian kali ini akan dilakukan untuk membangun model *Machine Learning* menggunakan metode CNN untuk melakukan klasifikasi terhadap ujaran kebencian pada teks Twitter. Penelitian ini diharapkan mampu memberikan model terbaik dengan performa yang baik dalam melakukan klasifikasi sehingga dapat menjadi acuan untuk pengembangan di penelitian-penelitian selanjutnya.

Penelitian sebelumnya yang membahas mengenai multilabel ujaran kebencian pada teks Twitter telah dilakukan menggunakan algoritma Support Vector Machine (SVM) [9]. Penelitian ini menerapkan desain Hierarchical Multi-Label Classification (HMC) dalam melakukan pengujiannya. HMC dilakukan dengan membentuk beberapa skenario berdasarkan hirarki yang dihasilkan dari label-label dengan karakteristik yang sama. Sebanyak lima skenario labeling dilakukan untuk mengukur kombinasi model mana yang memiliki akurasi terbaik. Lebih detail mengenai gambaran hirarki label pada *dataset* yang digunakan dapat dilihat pada Gambar 1.



Gambar 1. Hirarki label pada dataset multilabel ujaran kebencian bahasa Indonesia.

Penelitian pada literatur [9] berhasil mendapatkan akurasi terbaik sebesar 68,43%. Akurasi ini didapatkan dengan mereduksi jumlah label dari 12 menjadi 9. Label-label yang digunakan antara lain "Abusive", "Hate Speech", "Individual", "Group", "Religion", "Race", "Physical", "Gender", dan "Offensive". Sedangkan label yang direduksi adalah label yang dapat dikelompokkan sebagai tingkatan ujaran kebencian yaitu "Weak", "Moderate", dan "Strong".

Penelitian lain yang membahas klasifikasi terhadap multi-label ujaran kebencian pada teks Twitter menggunakan SVM adalah literatur [7]. *Dataset* yang digunakan juga sama dengan *dataset* pada literatur [9] yaitu *dataset* dari penelitian [10]. SVM dikombinasikan dengan Classifier Chains (CC). CC memiliki konsep untuk melakukan klasifikasi secara serial atau berurutan untuk tiap label yang ada. Secara konsep, teknis dari CC adalah membangun model dengan *single label* untuk kelas satu, kemudian hasilnya akan menjadi tambahan fitur untuk model klasifikasi selanjutnya untuk output berupa label berikutnya. Kombinasi SVM dan CC pada literatur [7]

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

berhasil mendapatkan akurasi terbaik sebesar 74,88%. Angka ini menunjukkan hasil akurasi 68,45% lebih baik dari literatur [9].

Selain menggunakan SVM+CC, literatur [7] juga menggunakan algoritma Convolution Neural Network (CNN) sebagai pembanding dari model SVM. Pada penelitian ini, model CNN memberikan nilai akurasi yang lebih kecil dibandingkan dengan model SVM+CC yaitu 65.07%. hasil ini diperoleh dengan skenario *pre-processing* tanpa menggunakan *stopword removal*, *stemming*, dan *translation*. Selain itu tidak imbangnya jumlah data *(imbalance)* menjadi salah satu faktor yang menyebabkan akurasi dari model CNN tidak bisa melebihi 70%.

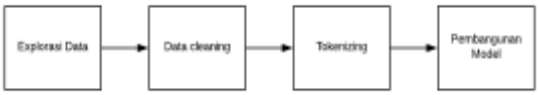
Hasil yang berbeda ditunjukkan pada model CNN yang dilakukan pada literatur [8]. Namun berbeda dengan literatur [7], literatur ini melakukan pengujian *multilabel* ujaran kebencian pada teks berbahasa Inggris. Penelitian ini melakukan penggabungan metode antara CNN dengan BERT *pretrained model*. Selain *pretrained model*, BERT juga menyediakan *tokenizer* yang berfungsi untuk mengkonversi teks menjadi *vector*. Fungsi *tokenizer* ini setara dengan TFIDF yang biasa digunakan pada *text analytic*. Adapun hasil dari gabungan model CNN dan BERT berhasil mencapai *metric* pengukuran F1 yang sangat baik, yaitu sebesar 92%.

Penelitian mengenai ujaran kebencian juga pernah dilakukan untuk teks Twitter pada bahasa Arab. Penelitian pada literatur melakukan perbandingan performa antara CNN dengan kombinasi antara CNN dan LSTM serta kombinasi antara CNN dan BiLSTM. Selain melakukan perbandingan performa pada arsitektur pemodelan. penelitian ini juga menguji performa model untuk melakukan klasifikasi terhadap *dataset binary class* (satu label) dan *multilabel*. Adapun hasil yang diperoleh adalah model terbaik dari kedua skenario dataset diperoleh model model CNN dimana model mampu memberikan performa 81% untuk klasifikasi *binary class*. Sedangkan untuk klasifikasi dengan skenario *dataset multilabel*, performa model menurun menjadi 73%. Penurunan performa ini bisa disebabkan karena bertambahnya jumlah target yang harus dikenali oleh model dalam melakukan klasifikasi [11].

Berdasarkan reviu yang sudah dilakukan pada beberapa literatur tersebut, dapat diketahui bahwa CNN dapat digunakan untuk menangani kasus pada *text analytic* yang dalam hal ini adalah kasus mengenai ujaran kebencian baik dalam bahasa Indonesia, bahasa Inggris, maupun bahasa Arab. Oleh karena itu, penelitian kali ini akan difokuskan untuk melakukan optimasi model CNN untuk menangani kasus ujaran kebencian pada bahasa Indonesia. Model yang dibangun akan didukung oleh tokenisasi menggunakan BERT *tokenizer*. BERT sebagai *tokenizer* digunakan mengingat hasil yang baik diperoleh pada penelitian ujaran kebencian berbahasa Inggris [8].

2.2. Metodologi Penelitian

Penelitian kali ini membangun model *Machine Learning* untuk mendeteksi teks yang mengandung ujaran kebencian menggunakan metode CNN. Sebelum membangun model CNN tentunya perlu dilakukan terlebih dahulu pengolahan terhadap *dataset text* yang dimiliki. Lebih jelas mengenai langkah-langkah dari penelitian yang dilakukan dapat dilihat pada Gambar 2.



Gambar 2. Alur penelitian.

2.3. Data Set

Penelitian kali ini akan menggunakan *dataset* teks ujaran kebencian berbahasa Indonesia yang berasal dari Twitter. *Dataset* ini merupakan *dataset* yang dihasilkan pada penelitian [10] dan sudah dapat dapat diakses oleh publik melalui github (https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection). Penggunaan *dataset* ini sudah digunakan oleh banyak penelitian ujaran kebencian berbahasa Indonesia dan diantaranya adalah literatur yang sudah disebutkan pada reviu literatur yaitu [9] [7]. *Dataset* ini memiliki jumlah kelas sebanyak 12 dimana kelas-kelas tersebut merupakan hasil diskusi antara peneliti dengan pihak Bareskrim Polri. Kelas-kelas yang dimaksud dapat dilihat pada Gambar 1. Jumlah twit yang terkandung dalam *dataset* ini adalah sebanyak 131.169 baris twit yang sudah ter-anotasi ke dalam 12 label [10].

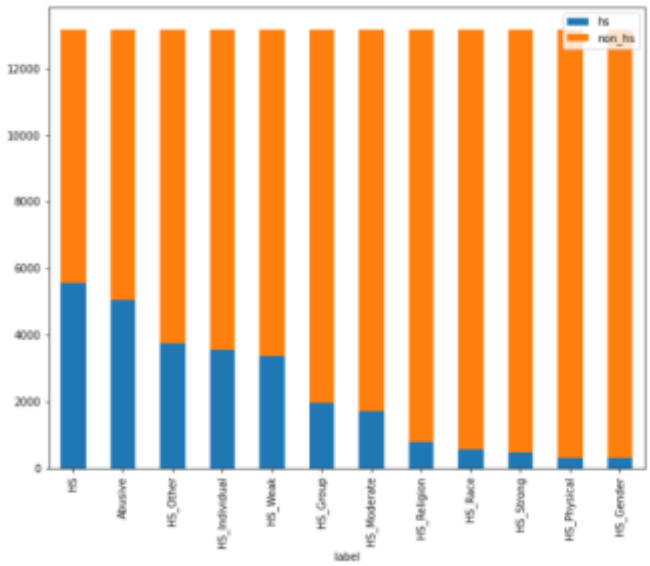
2.4. Eksplorasi Data

Sebelum masuk ke dalam pemodelan, data terlebih dahulu akan dieksplorasi. Eksplorasi yang dilakukan pada *dataset* juga dilakukan untuk mengetahui sebaran data untuk tiap label, histogram dari panjang kata tiap teks, dan tingkat kemunculan dari kata pada *dataset*. Hasil Eksplorasi ini nantinya akan digunakan untuk melihat bagaimana data akan ditangani untuk proses selanjutnya.

2.4.1. Sebaran Data tiap label

Eksplorasi terhadap sebaran data bertujuan untuk melihat keseimbangan dari data tiap kelas dan labelnya. Eksplorasi ini bermanfaat pada penelitian untuk mengetahui sifat dari data apakah termasuk data *imbalance* atau tidak. Adapun hasil eksplorasi dapat dilihat pada Gambar 3.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

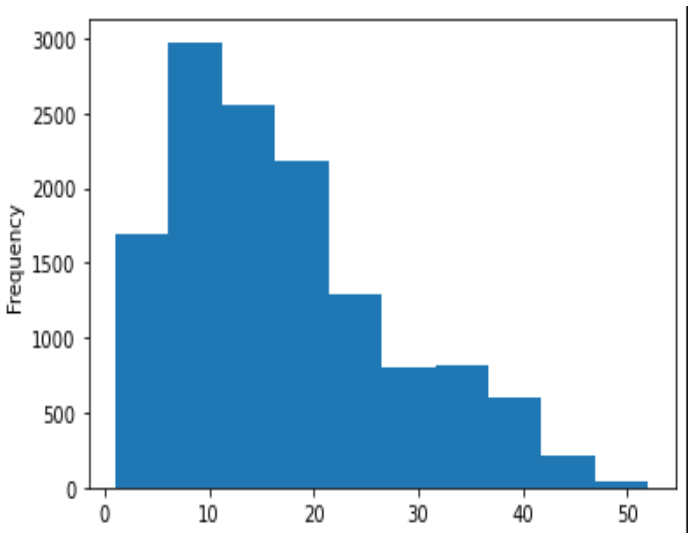


Gambar 3. Perbandingan jumlah kelas ujaran kebencian pada tiap label.

Pada Gambar 3 dapat dilihat bahwa jumlah kelas *hate speech* (ujaran kebencian) pada tiap label memiliki jumlah yang tidak seimbang. Label "HS_Religion", "HS_Race", "HS_Strong", "HS_Physical", dan "HS_Gender" menjadi 3 label dengan jumlah kelas positif ujaran kebencian terendah yaitu dibawah 1000 twit dari total 131.169 twit atau dengan kata lain berada pada angka di bawah 0,76% dari total jumlah twit.

2.4.2 Jumlah kata tiap teks

Eksplorasi ini bertujuan untuk melihat bagaimana sebaran panjang kata yang dimiliki oleh tiap teks pada dataset. Hasil eksplorasi bermanfaat ketika masuk ke dalam fase tokenisasi. Adapun hasil dari eksplorasi disajikan pada Gambar 4.



Gambar 4. Histogram panjang kata pada tiap teks.

Gambar 4 menunjukkan bahwa sebagian besar twit yang terdapat pada dataset memiliki panjang kata sekitar 10-20. Jika ditotalkan, jumlah twit dengan panjang kata kurang dari 20 adalah sebanyak 68,73% dari total twit dimana 61% dari twit tersebut merupakan twit dengan panjang kata sebanyak 10-20 kata.

2.4.3 Tingkat kemunculan kata

Eksplorasi ini bertujuan untuk melihat seberapa sering suatu kata muncul di tiap teks dalam *dataset*. Eksplorasi bermanfaat ketika penelitian masuk ke dalam fase *data cleaning*. Hal ini dikarenakan data yang paling sering muncul bisa jadi tidak diperlukan dalam proses pembangunan model. Hasil dari eksplorasi ini disajikan pada Gambar 5.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

Gambar 5. Wordcloud kata pada dataset.

Pada Gambar 5 dapat dilihat bahwa kata "USER" memiliki tingkat kemunculan yang paling tinggi pada *dataset*. Selanjutnya terdapat kata "xf0", "x9f", "x9f", dan beberapa kata serupa lain dengan intensitas yang cukup tinggi jika dilihat dari Gambar 5. Kata-kata tersebut merupakan emoji yang ditulis dalam format Unicode.

2.5. Data Processing

Data processing merupakan proses yang dilakukan pada data sebelum masuk ke dalam tahap pemodelan. *Data processing* menjadi penting untuk meningkatkan efisiensi dari pembangunan model yang akan dibangun [10]. Selain itu, penerapan *data processing* juga akan sangat berpengaruh pada performa dari model yang dihasilkan nantinya [7].

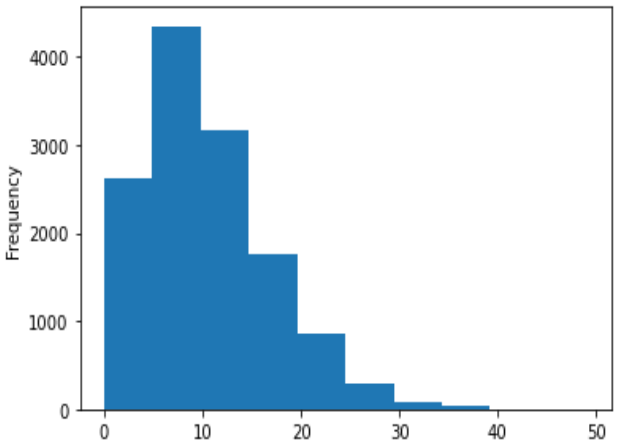
2.5.1 Case Folding

Case folding dilakukan dengan cara mengkonversi keseluruhan karakter menjadi huruf kecil. *Case folding* bertujuan untuk men-general-kan keseluruhan kata pada teks untuk memudahkan proses selanjutnya. Selain itu penerapan proses ini dapat meningkatkan performa dari model dalam melakukan klasifikasi [12]. Hal ini dikarenakan dalam ASCII code, huruf besar dan kecil memiliki nilai yang berbeda sehingga perbedaan case pada karakter bisa membuat kata yang sama memiliki nilai token yang berbeda.

2.5.2 Text Cleaning

Text cleaning diperlukan untuk menghilangkan kata-kata atau karakter yang dirasa tidak perlu. *Text cleaning* pada penelitian kali ini dilakukan dengan menghilangkan kata-kata seperti *link*, *username*, *hashtag*, angka, dan *punctuation*. Selain itu hasil eksplorasi pada Gambar 5 juga menjadi pertimbangan dalam melakukan *cleaning*. Berdasarkan Gambar 5, *text cleaning* ditambahkan dengan menghilangkan kata-kata "rt", "user", dan emoji Unicode.

Hasil dari *text cleaning* ini tentunya akan merubah jumlah sebaran kata pada *dataset*. Selain itu, sebaran mengenai panjang kata pada teks juga akan berefek karena beberapa kata sudah dihilangkan dari teks. Lebih jelas mengenai efek tersebut disajikan pada Gambar 6 dan gambar 7.



Gambar 6. Histogram panjang kata pada teks setelah cleaning.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

Pada Gambar 6 dapat dilihat bahwa sebaran panjang kata pada tiap teks dalam dataset mengalami perubahan jika dibandingkan dengan Gambar 4. Pada Gambar 4 sebaran data memiliki nilai memuncak ada panjang kata 10-20. Sedangkan pada gambar 6, teks dengan panjang kata sebanyak 10 mendominasi *dataset*. Kemudian teks dengan panjang kata lebih dari 40 bisa dikatakan menjadi tidak ada.

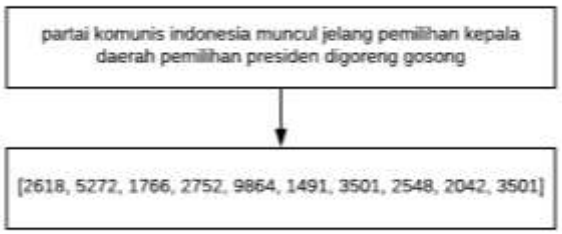
Selanjutnya pada Gambar 7, terlihat bahwa kata *"user"* dan emoji Unicode sudah tidak terlihat lagi. Kata-kata lain menjadi bermunculan dengan intensitas yang hampir merata. Dengan ini, data hasil *cleaning* sudah siap untuk digunakan pada tahap tokenisasi untuk menjadi masukan bagi model yang akan dibangun.

Gambar 7. Wordcloud kata pada dataset setelah cleaning.

2.5.3 Tokenisasi

Tokenisasi bertujuan untuk menghasilkan vektor dari tiap teks twit pada dataset. Transformasi dari teks menjadi vektor diperlukan karena pada dasarnya pemodelan pada algoritma Machine Learning bekerja dengan melakukan perhitungan terhadap angka-angka. Selanjutnya vektor hasil tokenisasi ini akan menjadi fitur dari teks dan menjadi input untuk pemodelan.

Tokenisasi pada penelitian kali ini akan memanfaatkan tokenizer dari model BERT yang sudah ada. Tokenizer dari pretrained model yang akan digunakan pada penelitian kali ini adalah model "cahya/bert-base-indonesian-522M" (https://huggingface.co/cahya/bert-base-indonesian-522M).



Gambar 8. Ilustrasi hasil tokenisasi oleh BERT Tokenizer.

BERT Tokenizer memiliki dua buah spesial token yang ditambahkan pada teks yaitu "[CLS]" dan "[SEP]". Kedua token ini berguna ketika melakukan tokenisasi terhadap 2 kalimat atau lebih. Token "[CLS]" ditambahkan pada awal teks sedangkan token "[SEP]" ditambahkan di setiap akhir kalimat pada teks [8]. Namun penambahan spesial token ini bersifat *optional* dalam implementasinya. Artinya, untuk kasus pendeteksian ujaran kebencian kali ini spesial token ini tidak perlu untuk ditambahkan ke dalam teks. Adapun ilustrasi dari tokenisasi yang dihasilkan oleh BERT Tokenizer ditunjukkan pada Gambar 8.

2.6. Model Klasifikasi

Model klasifikasi multilabel ujaran kebencian menggunakan metode CNN. Output dari model diharapkan mampu melakukan klasifikasi pada 12 label data ada pada *dataset*. Adapun model terbaik akan dicari dengan menggunakan beberapa parameter pengujian seperti *learning rate*, jumlah *convolution layer*, ukuran kernel konvolusi, jumlah layer nn, serta jumlah node pada layer nn.

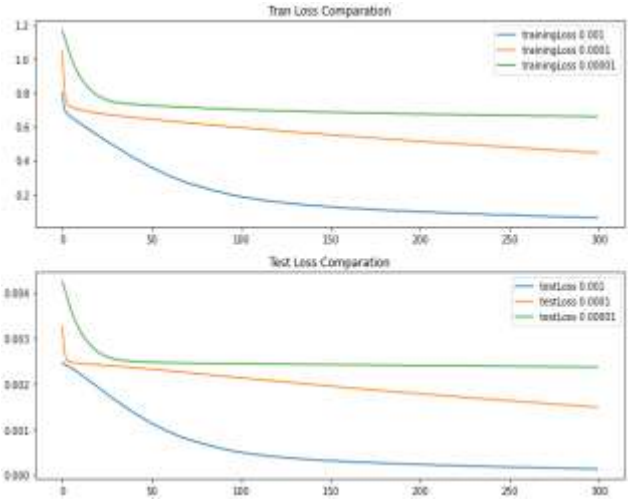
Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

3. HASIL DAN PEMBAHASAN

Pengujian dilakukan dengan beberapa skenario perbandingan parameter CNN seperti yang disebutkan pada bagian 3.4. Performa dari masing-masing skenario akan diukur menggunakan nilai rata-rata dari *loss* yang dihasilkan model dengan teknik pengujian menggunakan *cross validation k-fold.*

3.1 Hasil pengujian pada *Learning Rate*

Pengujian dilakukan untuk mendapatkan nilai *learning rate* terbaik untuk model. *Learning rate* berfungsi untuk mengatur seberapa besar perubahan bobot dari model yang bangun di tiap iterasinya. Nilai dari parameter *learning rate* yang diujikan pada penelitian ini ada 3 yaitu 0.001, 0.0001, dan 0.00001. nilai 0.001 merupakan nilai *default* dari *learning rate* yang disediakan pada *library pytorch*. Adapun hasil dari pengujian disajikan dalam Gambar 9.



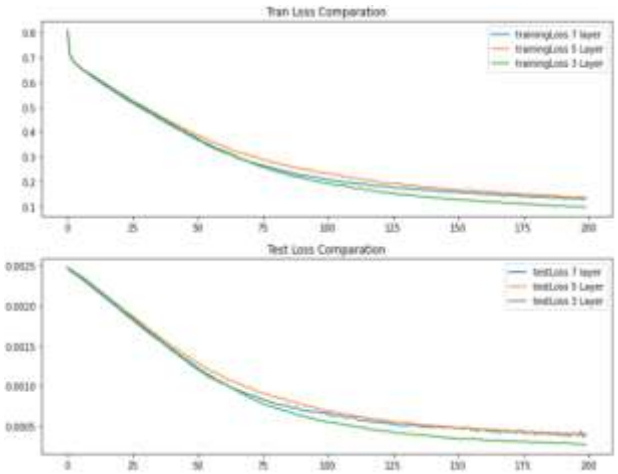
Gambar 9. Perbandingan nilai loss parameter pengujian learning rate.

Pengujian dilakukan dengan beberapa skenario perbandingan parameter CNN seperti yang disebutkan pada bagian 3.4. Performa dari masing-masing skenario akan diukur menggunakan nilai rata-rata dari *loss* yang dihasilkan model dengan teknik pengujian menggunakan *cross validation k-fold.*

Pada Gambar 9, dapat dilihat bahwa nilai *learning rate* terbaik didapatkan pada nilai 0.001. *Loss* yang didapat dengan *learning rate* 0.001 berada dibawah 0.001 pada *epoch* ke 300. Sedangkan 2 *learning rate* lainnya berada di atas 0.001. selanjutnya dpaat dilihat bahwa penurunan *loss* tidak terlalu signifikan pada *epoch* setelah 200. Oleh karena itu pada pengujian bagian ini didapatkan hasil bahwa nilai terbaik untuk penggunaan learning rate pada model adalah 0.001. Selanjutnya akan digunakan jumlah epoch sebanyak 200 untuk efisiensi waktu komputasi.

3.2 Hasil Pengujian pada Layer Konvolusi

CNN memiliki layer konvolusi sebagai layer awal dari model. Pengujian dilakukan untuk mendapatkan jumlah layer konvolusi terbaik untuk melakukan klasifikasi terhadap ujaran kebencian pada teks. Adapun nilai parameter yang akan diujikan pada pengujian kali ini adalah jumlah layer konvolusi sebanyak 3 layer, 5 layer, dan 7 layer. Hasil pengujian pada bagian ini disajikan dalam Gambar 10.



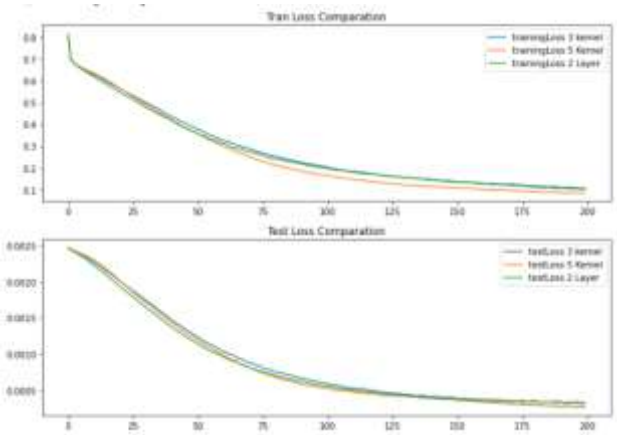
Gambar 10. Perbandingan nilai *loss* parameter pengujian jumlah layer konvolusi.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

Sebanyak 3 pengujian dilakukan untuk mencari nilai terbaik dari parameter jumlah layer konvolusi. Gambar 10 menunjukkan bahwa jumlah layer terbaik untuk melakukan klasifikasi terhadap ujaran kebencian pada teks didapatkan pada angka 3 layer. Nilai *loss* dari *training* dan *testing* terakhir pada *epoch* ke 200 dari layer berjumlah 3 memiliki nilai yang paling kecil dibandingkan dengan *loss* ketika menggunakan layer kovolusi sebanyak 5 dan 7. Sehingga dapat ditarik kesimpulan bahwa pada pengujian bagian ini parameter dengan jumlah layer konvolusi terbaik didapatkan pada layer konvolusi berjumlah 3.

3.3 Hasil pengujian pada Jumlah Kernel Konvolusi

Layer konvolusi memiliki kernel berupa matriks atau vektor sesuai dengan dimensi yang dimilikinya. Pada kali ini dimensi yang digunakan untuk penelitian adalah dimensi 1 sehingga kernel akan berbentuk vektor. Pengujian pada bagian ini dilakukan untuk mendapatkan ukuran vektor terbaik dari kernel untuk melakukan klasifikasi terhadap ujaran kebencian pada teks. Adapun mengenai nilai dari jumlah kernel akan menggunakan 3 parameter nilai yaitu 3, 5, dan 7. Hasil dari pengujian disajikan dalam Gambar 11.

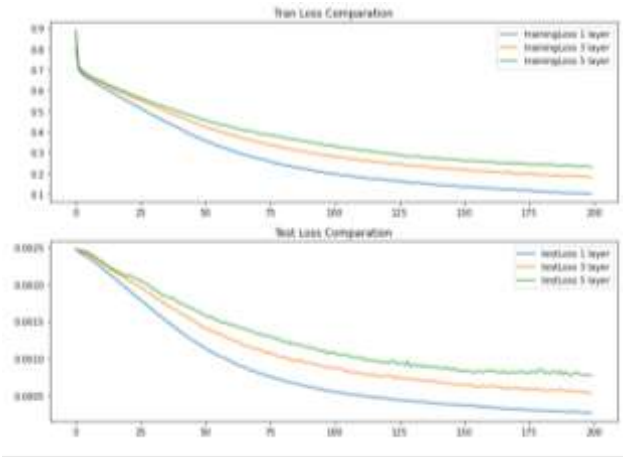


Gambar 11. Perbandingan nilai loss parameter pengujian jumlah layer konvolusi.

Pengujian terhadap ukuran kernel dari layer konvolusi terlihat memiliki perbedaan nilai loss di setiap iterasi yang tidak terlalu signifikan. Line chart pada test loss untuk 3 parameter pengujian menunjukkan garis yang saling berhimpitan satu sama lain. Namun pada epoch terakhir, parameter dengan kernel berukuran 2 berhasil memisahkan diri dan memiliki nilai loss yang paling kecil dibandingkan dengan 2 parameter lain yaitu 3 dan 5 kernel. Sehingga parameter terbaik dalam pengujian kali ini adalah parameter dengan ukuran kernel sebesar 2.

3.4 Hasil pengujian pada Jumlah Layer NN

Layer NN merupakan layer terakhir dari CC. layer ini menjadi penentu dari hasil klasifikasi dari model. Jumlah layer ini sendiri untuk tiap model dapat diinisiasi dengan jumlah yang berbeda-beda untuk tiap permasalahan. Oleh karena itu, pangujian ni dilakukan untuk mendapatkan jumlah layer nn terbaik untuk melakukan klasifikasi ujaran kebencian pada teks. Adapun parameter yang digunakan pada pengujian bagian ini adalah dengan menggunakan jumlah layer nn sebanyak 1, 3, dan 5 layer.



Gambar 12. Perbandingan nilai loss parameter pengujian jumlah layer nn.

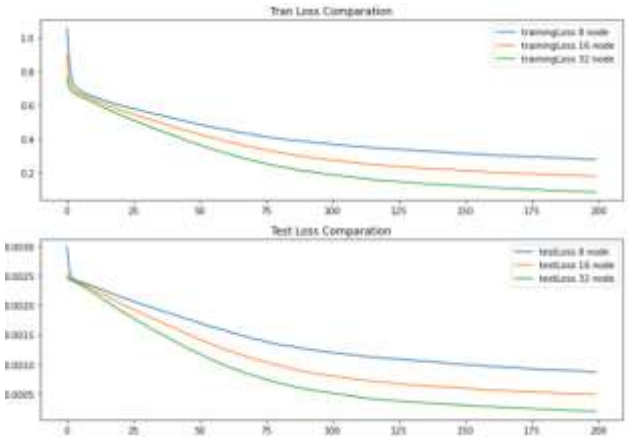
Pada Gambar 12, dapat dilihat bahwa perbedaan loss yang signifikan terlihat antara ketiga parameter yang diujikan. Layer nn dengan jumlah 1 layer memiliki nilai loss yang paling kecil dibandingkan denga nn dengan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

jumlah layer 3 dan 5. Penurunan loss layer nn berjumlah 3 sudah mengungguli parameter lain sejak *epoch-epoch* awal. Oleh karena itu, didapatkan bahwa jumlah layer nn terbaik untuk klasifikasi ujaran kebencian pada teks adalah sebesar 1 layer.

3.5 Hasil Pengujian pada Jumlah Node layer NN

Pengujian terakhir dilakukan untuk mencari jumlah node yang digunakan pada layer NN. Layer NN sejatinya adalah sebuah vektor yang berisi node-node atau elemen dari vektor itu sendiri. Jumlah node ini tidak memiliki nilai yang pasti karena dapat diberikan dengan nilai yang berbeda-beda. Pada bagian ini, akan dilakukan pengujian untuk mendapatkan jumlah node terbaik pada layer NN untuk dapat melakukan klasifikasi ujaran kebencian pada teks.



Gambar 13. Perbandingan nilai loss parameter pengujian jumlah node layer nn.

Pada pengujian terkahir, hasil untuk pengujian layer nn pada Gambar x menunjukkan perbedaan yang signifikan juga seperti pada Gambar y. Dapat dilihat pada Gambar x bahwa penurunan nilai *loss* pada pengujian menggunakan parameter jumlah node layer nn sebanyak 32 lebih baik daripada 2 parameter lain. Dengan demikian pada pengujian bagian ini didapatkan bawa jumlah node layer nn terbaik untuk melakukan klasifikasi ujaran kebencian pada teks adalah sebesar 32 node.

3.6. Hasil pengujian Model Terbaik

Berdasarkan pengujian yang sudah dilakukan pada bagian-bagian sebelumnya, maka didapatkan model CNN terbaik untuk melakukan klasifikasi multilabel ujaran kebencian pada teks berbahasa Indonesia dengan komposisi *learning rate* 0.001, epoch 200, layer konvolusi sebanyak 3 dengan kernel 2, dan layer nn sebanyak 1 dengan jumlah node sebanyak 32. Model terbaik ini berasal dari loss paling kecil dari masing-masing parameter pengujian. Selanjutnya akan dilihat bagaimana performa klasifikasi model terhadap 12 label ujaran kebencian pada dataset.

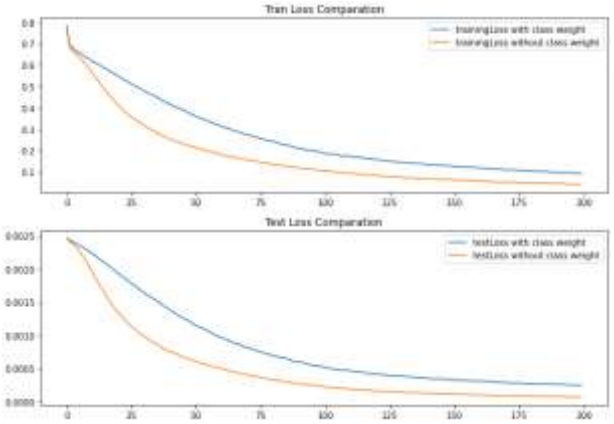
Kondisi *imbalance* pada *dataset* menjadi permasalahan dalam melakukan pemodelan. Oleh karena itu, pada kali ini akan dilakukan penerapan teknik *class weighting* ketika membangun model. *Class weight* ini berperan ketika melakukan *update weight* saat proses *backpropagation* [13]. Adapun hasil dari klasifikasi lebih detail yang dihasilkan oleh model dengan penerapan *class weighting* disajikan dalam *confusion matriks* pada Gambar 14.

Gambar 14. *Multilabel confusion matrix* model dengan class weighting.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 3, Nomor 2, Desember 2021
e-ISSN 2685-998X
DOI 10.30865/json.v3i2.3610

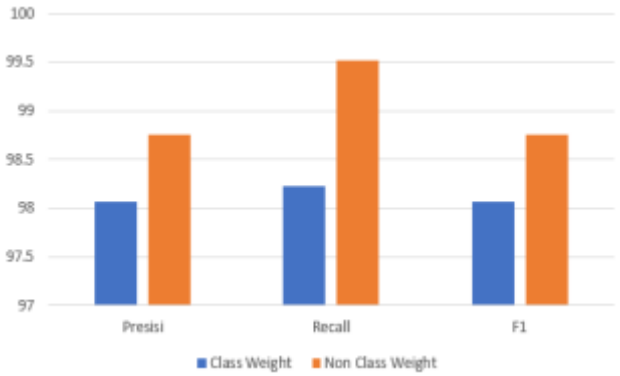
Confusion matriks pada Gambar 14 merupakan confusion matriks hasil akumulasi dari tiap *fold* pada cross Gambar 3 menunjujkkan bahwa nilai *True Positif* (TP) tiap kelas memiliki nilai yang jauh lebih besar daripada nilai *False Negatif* (FN). Hal ini dapat dilihat dari *heatmap* TP memiliki nilai yang lebih terang dibandingkan dengan nilai FN yang memiliki posisi vertikal dengan TP. Penelitian ini juga melakukan perbandingan hasil model dengan model CNN tanpa penerapan *class weighting* di akhir pengujian. Hal ini dilakukan sebagai upaya untuk melihat efektifitas dari penerapan *class weighting* pada model.

Pada Gambar 15, dapat dilihat bahwa *loss* dari model tanpa penerapan *class weighting* memiliki nilai yang lebih kecil dibandingkan dengan model yang menerapkan *class weighting*.



Gambar 15. Perbandingan *loss* penerapan *class weighting*.

Hasil ini berbanding lurus juga dengan hasil dari perhitungan rata-rata *metric* pengukuran antara kedua skema pembangunan model yang disajikan pada Gambar 16. Nilai presisi, recall, dan f1 dari model yang menerapkan *class weighting* memiliki nilai yang lebih rendah dibandingkan dengan model tanpa penerapan *class weight*. Dari nilai rata-rata presisi, dapat dinyatakan bahwa model tanpa penerapan *class weighting* lebih akurat dengan nilai akurasi 98,76% daripada model dengan penerapan *class weighting* yaitu 98,07%.



Gambar 16. Perbandingan rata-rata presisi, recall, dan f1 tiap label pada penerapan class weighting.

4. KESIMPULAN

Model CNN pada penelitian ini didapatkan pada jumlah *epoch* 200, *learning rate* 0.001, 3 layer konvolusi dengan kernel 2, dan 1 layer nn dengan node sebanyak 32. Model CNN yang didapat berhasil memberikan akurasi yang sangat baik dalam melakukan klasifikasi multilabel ujaran kebencian pada teks twitter pada dataset dengan akurasi 98,07%. Pada bagian akhir dari pengujian, penulis melakukan perbandingan antara pembangunan model dengan menerapkan *class weighting* dan tanpa penerapan *class weighting*. Secara teori, model yang digunakan bersifat *imbalance* sehingga sudah tepat untuk ditangani dengan *class weighting*. Akan tetapi, akurasi model dengan penerapan *class weighting* memiliki nilai akurasi d bawah model yang tidak menerapkannya. Oleh karena itu, penulis merekomendasikan untuk menganalisa lebih jauh lagi mengenai penerapan *model weighting* dalam menangani kasus ujaran kebencian pada teks twitter terlebih khusus lagi dengan dataset yang sama dengan yang digunakan pada penelitian ini. Selain itu, baiknya performa dari model CNN ini bisa diuji coba pada teks yang lebih panjang lagi. Hal ini dikarenakan pada penelitian ini panjang vektor hasil tokenisasi adalah 10. Mengingat sebaran data yang lebih banyak pada panjang teks sekitar 10 kata.

REFERENCES

[1] S. Kemp, "Digital in Indonesia: All the Statistics You Need in 2021 — DataReportal – Global Digital Insights," 2021. https://datareportal.com/reports/digital-2021-indonesia (accessed Jun. 27, 2021).
[2] A. Briliani, B. Irawan, and C. Setianingsih, "Hate speech detection in indonesian language on instagram comment section using K-nearest neighbor classification method," *Proc. - 2019 IEEE Int. Conf. Internet Things Intell. Syst. IoTaIS 2019*, pp. 98–104, 2019, doi: 10.1109/IoTaIS47347.2019.8980398.
[3] S. Y. Hukmana, "125 Akun Medsos Terjaring Virtual Police - Medcom.id," 2020. https://www.medcom.id/nasional/hukum/gNQ5RnwN-125-akun-medsos-terjaring-virtual-police (accessed Jun. 27, 2021).
[4] L. P. A. S. Tjahyanti, "Pendeteksian Bahasa Kasar (Abusive Language) Dan Ujaran Kebencian (Hate Speech) Dari Komentar Di Jejaring Sosial," *J. Chem. Inf. Model.*, vol. 07, no. 9, pp. 1689–1699, 2020.
[5] P. Devita, "Apakah semua ujaran kebencian perlu dipidana? Catatan untuk revisi UU ITE." https://theconversation.com/apakah-semua-ujaran-kebencian-perlu-dipidana-catatan-untuk-revisi-uu-ite-156132 (accessed Jun. 27, 2021).
[6] F. Alzami, N. P. P, R. A. P, R. A. Megantara, and D. P. Prabowo, "SENTIMENT ANALYSIS UNTUK DETEKSI UJARAN KEBENCIAN PADA DOMAIN POLITIK," vol. 5, no. Sens 5, pp. 213–218, 2020.
[7] K. M. Hana, Adiwijaya, S. Al Faraby, and A. Bramantoro, "Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines," *2020 Int. Conf. Data Sci. Its Appl. ICoDSA 2020*, 2020, doi: 10.1109/ICoDSA50139.2020.9212992.
[8] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," *PLoS One*, vol. 15, no. 8 August, 2020, doi: 10.1371/journal.pone.0237861.
[9] F. A. Prabowo, M. O. Ibrohim, and I. Budi, "Hierarchical multi-label classification to identify hate speech and abusive language on Indonesian twitter," *2019 6th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2019*, pp. 1–5, 2019, doi: 10.1109/ICITACEE.2019.8904425.
[10] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.
[11] R. Duwairi, A. Hayajneh, and M. Quwaider, "A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets," *Arabian Journal for Science and Engineering*, vol. 46, no. 4. pp. 4001–4014, 2021, doi: 10.1007/s13369-021-05383-3.
[12] N. A. Setyadi, M. Nasrun, and C. Setianingsih, "Text Analysis for Hate Speech Detection Using Backpropagation Neural Network," *Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018*, pp. 159–165, 2018, doi: 10.1109/ICCEREC.2018.8712109.
[13] K. Sozykin, S. Protasov, A. Khan, R. Hussain, and J. Lee, "Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks," *Proc. - 2018 IEEE/ACIS 19th Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distributed Comput. SNPD 2018*, pp. 146–151, 2018, doi: 10.1109/SNPD.2018.8441034.