DETEKSI UJARAN KEBENCIAN PADA TWITTER BAHASA INDONESIA MENGGUNAKAN MACHINE LEARNING: REVIU LITERATUR

Aditya Perwira Joan Dwitama
Program Studi Informatika
Universitas Islam Indonesia
Jl. Kaliurang 14,5 Sleman,
Yogyakarta, Indonesia 55584
20917035@students.uii.ac.id

ABSTRAKSI

Meningkatnya pengguna media sosial mengakibatkan peningkatan aktifitas komunikasi antar warganet dalam media daring. Misalnya media twitter, warganet dapat berkomunikasi melalui twit. Twit pada twitter dapat memiliki sifat negatif dan sifat ini perlu memiliki perhatian khusus karena kemungkinan besar akan mengandung ujaran kebencian. Kasus ujaran kebencian ini kemudian oleh pemerintah diatasi atau dicegah dengan salah satunya Undang-undang Informasi dan Transaksi Elektronik (UU ITE) yang dikeluarkan pada tahun 2018 pasal 28 ayat 2 tentang ujaran kebencian. *Machine learning* dalam penerapannya mampu mengolah data dalam bentuk teks atau (*text analytic).* Artikel ini menyajikan reviu terhadap literatur-literatur yang memanfaatkan *machine learning* untuk membantu mendeteksi teks yang mengandung ujaran kebencian. Hasil reviu menunjukkan bahwa penelitian terhadap kebencian menggunakan *machine learning* dapat dilakukan untuk mendeteksi kategori teks; apakah tergolong sebagai ujaran kebencian, serta memprediksi kategori dari ujaran kebencian yang terkandung dalam suatu teks. Algoritma RNN ternyata memiliki akurasi yang terbaik jika dibandingkan dengan algoritma lain yaitu 91%.

Kata Kunci

Ujaran kebencian; Bahasa Indonesia; Twit; *Machine learning*.

1. PENDAHULUAN

Pengguna internet di Indonesia sudah mencapai jumlah 202,6 juta pengguna atau sekitar 73,7% dari total jumlah penduduk Indonesia [1]. Dimana sebagian besar dari penggunaan internet tersebut ditujukan untuk beraktifitas 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 5 aplikasi media sosial yang paling sering digunakan oleh rentang pengguna dengan umur 16-64 tahun adalah YouTube, WhatsApp, Instagram, Facebook, dan Twitter [1].

Media sosial menyediakan sarana bagi warga internet atau biasa disebut dengan sebutan warganet untuk berkomunikasi secara daring. Misalnya Twitter, warganet dapat berkomunikasi melalui twit yang dilontarkan pada aplikasi. Twit ini dapat bersifat positif dan ada juga yang bersifat negatif. Komentar yang negatif menjadi masalah karena biasanya mengandung unsur ujaran kebencian dan dapat berakibat sanksi hukum bagi penulisnya [2].

Direktorat Siber Bareskrim Polri telah mencatat bahwa sebanyak 125 akun media sosial sudah mendapat teguran oleh polisi virtual terkait dengan konten yang terindikasi mengandung unsur ujaran kebencian. Dari akun-akun yang mendapat teguran tersebut, akun 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. Dalam 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 multi tafsir 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].

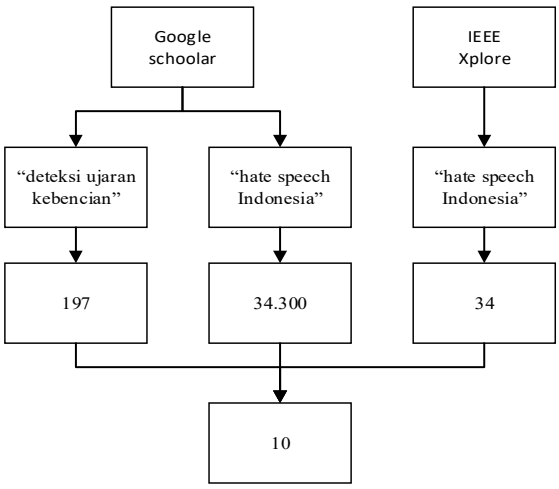
Machine learning dalam sub-bagiannya yaitu *text analytic* memiliki algoritma-algoritma yang dapat melakukan pengenalan atau pengelompokan terhadap suatu objek teks. *Text analytic* ini dapat dimanfaatkan dalam mengatasi kasus ujaran kebencian dalam media sosial melalui kemampuannya dalam mendeteksi *cyberbullying*, bahasa kasar, maupun *cyberhate* [6].

Oleh karena itu, dilakukan reviu literatur mengenai terkait dengan objek penelitian ujaran kebencian dilakukan untuk mengumpulkan metode atau algoritma apa saja yang sudah pernah digunakan untuk menganalisis ujaran kebencian dalam bahasa Indonesia. Bahasa Indonesia dipilih sebagai objek dari reviu literatur karena selain bahasa ini adalah bahasa utama dari penulis dan kasusnya dari Bahasa Indonesia. Selanjutnya, reviu literatur ini dilakukan untuk melihat ragam dataset yang digunakan dalam penelitian serta performa dari masing-masing algoritma yang digunakan dalam penelitian tersebut.

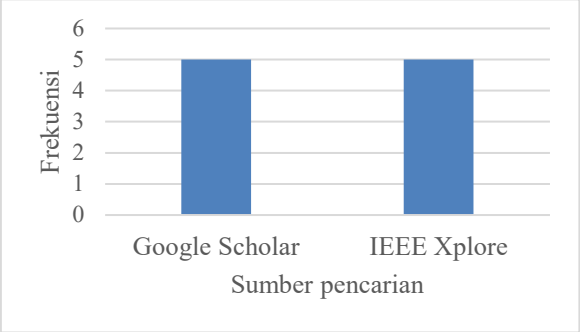
2. STRATEGI SELEKSI LITERATUR

Sebelum melakukan reviu terhadap literatur, dilakukan pengumpulan terhadap literatur-literatur terlebih dahulu. Pengumpulan literatur untuk direviu perlu dilakukan untuk melihat penelitian-penelitian yang sudah dilakukan sebelumnya. Selanjutnya dapat diketahui apa saja yang sudah dilakukan pada literatur tersebut dan bagaimana hasilnya sehingga kedepannya dapat diketahui posisi dari penelitian yang akan dilakukan terhadap ujaran kebencian menggunakan metode *machine learning.*

Adapun strategi utama yang dilakukan adalah dengan memanfaatkan mesin pencarian pada google scholar (https://scholar.google.co.id/) dan IEEE Xplore (https://ieeexplore.ieee.org/).



Gambar 1. Hasil pencarian.



Gambar 2. Perbandingan tahun publikasi pada literatur yang di-reviu.

Gambar 3. Perbandingan tahun publikasi pada literatur yang direviu.

Pada google scholar dicoba 2 kata kunci yaitu "deteksi ujaran kebencian" dan "hate speech Indonesia". Kedua kata kunci ini bertujuan agar *return* dari mesin pencarian spesifik mengarah kepada literatur berbahasa Indonesia atau berbahasa Inggris dengan objek penelitian berupa bahasa Indonesia.

Berbeda dengan google scholar, pada IEEE Xplore digunakan satu kata kunci yaitu "hate speech Indonesia". Hal ini dikarenakan dari IEEE Xplore yang menghimpun literatur berbahasa Inggris. Kata kunci ini ditujukan agar *return* dari mesin pencarian berupa objek penelitian berupa teks berbahasa Indonesia. Adapun hasil dari pencarian dapat dilihat pada Gambar 1.

Tabel 1. Literatur-literatur yang direviu.

Penulis	Judul
Farrikh A., dkk [6]	Sentimen Analysis untuk Deteksi Ujaran Kebencian pada Domain Politik
Dayang Putri Nur L. [7]	Deteksi Ujaran Kebencian pada Twitter Menjelang Pilpres 2019 dengan Machine Learning
Guntur Budi H, dkk [8]	Hate Speech and Abusive Language Classification using fastText
Arum Sucia S., dkk [9]	Analysis Text of Hate Speech Detection Using Recurrent Neural Network
Nabiila Adani S., dkk [10]	Text Analysis for Hate Speech Detection Using Backpropagation Neural Network
Faizal Adhitama P., dkk [11]	Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter
Karimah Mutisari H., dkk [12]	Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines
M. Okky I., dkk [13]	Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter
Ika Alfina, dkk [14]	Hate Speech Detection in the Indonesian Language:A Dataset and Preliminary Study
Luh Putu Ary Sri T. [15]	Pendeteksian Bahasa Kasar (Abusive Language) Dan Ujaran Kebencian (Hate Speech) dari Komentar di Jejaring Sosial

Pada Gambar 1, dapat dilihat bahwa total makalah yang didapat dari kedua mesin pencarian adalah 34.531 literatur. Kemudian dilakukan filtering terhadap hasil *return* teratas. Filtering dilakukan dengan melihat objek dari penelitian pada literatur. Reviu dilakukan pada literatur dengan objek penelitian berupa teks pada twitter. Selain itu dilihat tahun publikasi dari literatur yang muncul pada mesin pencarian. Hasil dari filtering ini mendapatkan 10 literatur yang dirasa paling relevan untuk direviu. Adapun untuk perbandingan literatur yang diambil berdasarkan sumber mesin pencariannya dapat dilihat pada Gambar 2.

Grafik pada Gambar 2 memperlihatkan jumlah literatur yang diperoleh dari setiap mesin pencarian. Literatur dari hasil mesin pencarian pada IEEE Xplore dan Google Scholar memiliki jumlah yang sama yakni sebanyak 5 literatur.

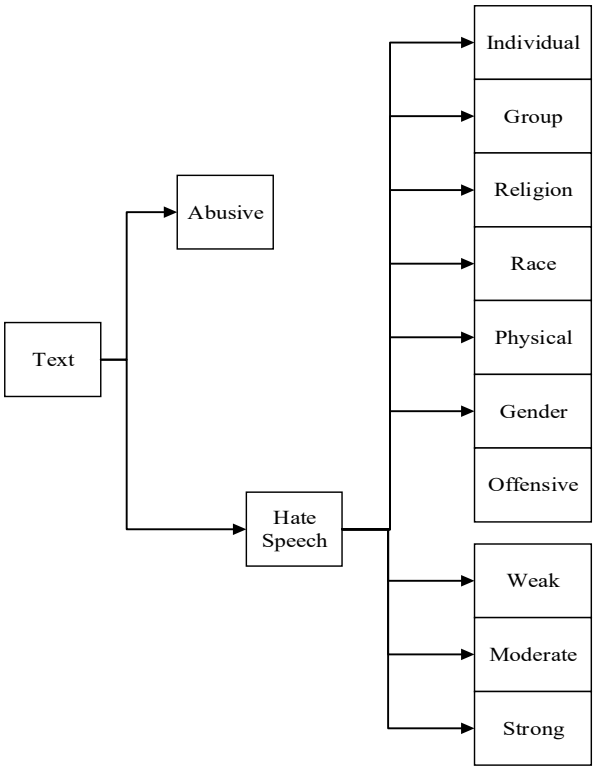
Banyaknya hasil *return* dari mesin pencarian pada Google Scholar membuat filterisasi hanya dilakukan pada 10 halaman awal saja. Kemudian penentuan mengenai literatur mana saja yang akan direviu pada kedua mesin pencarian mempertimbangkan tahun publikasi. Literatur yang diambil adalah yang memiliki tahun publikasi di atas 5 tahun sejak tanggal penulisan reviu ini. Perbandingan jumlah tahun publikasi dari literatur yang digunakan dapat dilihat pada Gambar 3.

Pada Gambar 3 dapat dilihat bahwa keseluruhan literatur yang digunakan memiliki tahun publikasi di atas 5 tahun ke belakang. Literatur yang paling tua didapatkan dengan tahun publikasi 2017. Sedangkan literatur yang dipublikasikan pada tahun 2019 adalah yang terbanyak dengan jumlah 4 literatur. Adapun mengenai detail judul dan penulis dari literatur yang akan digunakan sebagai bahan reviu dapat dilihat pada Tabel 1.

Selanjutnya literatur-literatur tersebut direviu untuk mendapatkan gambaran mengenai isi dari setiap literatur. Adapun fokus dari faktor yang akan diteliti kali ini ada 4 yaitu tujuan pembangunan

Tabel 2. Tujuan dari penelitian dalam literatur.

No	Jenis	Literatur
1	Membangun model untuk mengategorikan teks dalam satu label (ujaran kebencian dan bukan ujaran kebencian)	[6] [7] [8] [9] [10] [14]
2	Membangun model untuk mengategorikan teks dalam label lebih dari 1	[11] [12] [13] [15]



Gambar 4. Ilustrasi pelabelan teks pada *dataset* multiabel.

model, penggunaan *dataset*, penerapan teknik *pre-processing* dan performa dari metode yang digunakan.

Fokus pertama adalah untuk melihat bagaimana *output* dari model yang sudah dibangun. *Output* yang dimaksud adalah jumlah label yang akan diklasifikasi oleh model pengenalan pada teks. Hal ini didasarkan pada *case* dari teks ujaran kebencian yang dapat dilakukan untuk mendeteksi teks menjadi ujaran kebencian atau tidak ataupun melakukan klasifikasi terhadap kategori dari setiap ujaran kebencian dengan label yang akan berjumlah lebih dari satu sesuai dengan jumlah kategori yang diberikan.

Fokus kedua adalah mengenai penggunaan *dataset*. Penggunaan *dataset* ini menjadi fokus pada reviu karena untuk membangun suatu model *machine learning*, *dataset* sangat berpengaruh terhadap bagaimana model mencari bobot dari setiap perhitungannya ketika menjalankan proses training. Model yang sama dengan *dataset* yang berbeda akan memberikan hasil yang berbeda pula pada performa model.

Fokus ketiga adalah mengenai teknik *pre-processing* yang diterapkan pada setiap literatur. Fokus ini menjadi penting untuk direviu karena berkaitan dengan fitur yang akan didapat untuk membangun model *machine learning*. *Pre-processing* akan menentukan bagaimana cara model melakukan *update* terhadap tiap bobot-bobot pemodelannya sehingga menghasilkan model *machine learning* yang dapat bekerja dengan baik untuk mengenali objek yang sedang diuji.

Fokus terakhir adalah untuk melihat bagaimana performa dari model. Tiap literatur menggunakan metode yang berbeda untuk membangun model *machine learning* guna mengatasi masalah untuk pengenalan teks ujaran kebencian. Ada beberapa literatur yang menggunakan metode klasifikasi yang sama, namun akan memiliki perbedaan pada teknis pengerjaannya. Misalkan perbedaan pada tahap ekstraksi fitur atau pun perbedaan dari segi struktur modelnya.

3. HASIL

Melihat dari isi tiap literatur yang direviu, penulis dapat melakukan pengelompokan menjadi 4 bagian. Pertama adalah mengenai tujuan dari masing-masing literatur. Kedua adalah mengenai sumber *dataset* yang digunakan. Ketiga adalah mengenai teknik-teknik *pre*-processing pada teks yang digunakan. Kemudian yang terakhir adalah mengenai bagaimana performa yang dihasilkan oleh masing-masing penelitian.

3.1 Tujuan literatur

Berdasarkan penelitian yang sudah dilakukan, terdapat 2 tujuan umum dari penelitian yang dilakukan pada setiap literatur. Detail mengenai tujuan dari model yang dibangun pada masing literatur disajikan pada Tabel 2.

3.1.1 Model dengan output satu label

Pertama, penelitian dilakukan untuk membangun model *machine learning* yang dapat mengategorikan teks ke dalam satu label saja. model diharapkan mampu memberikan output apakah suatu objek yang dalam hal ini adalah teks bernilai iya atau tidak untuk satu jenis kelas kategori. Pada tiap literatur, kelas yang dimaksud adalah kelas ujaran kebencian. Jadi yang menjadi poin penelitiannya adalah mengenai kemampuan model yang dibangun dalam melakukan deteksi terhadap suatu teks apakah termasuk ke dalam teks yang bersifat ujaran kebencian atau tidak [6] [7] [8] [9] [10] [14].

3.1.2 Model dengan output multilabel

Tujuan lain dari literatur yang diperoleh adalah untuk membangun model yang dapat mengenali klasifikasi dengan jumlah label per teksnya lebih dari satu [11] [12] [13] [15]. Masing-masing dari literatur memiliki jenis output yang berbeda antara satu dengan yang lain.

Pada literatur pertama [11] dalam subbab ini dilakukan 5 jenis skenario pengujian terhadap model yang diberikan. Skenario-skenario tersebut berupa manipulasi terhadap label dari tiap teks. Adapun label asli dari *dataset* yang dimiliki adalah berjumlah 12 label dengan nama “Hate Speech (HS)”, “HS Individual”, “HS Group”, “HS Religion”, “HS Race”, “HS Physical”, “HS Gender”, “HS Other”, “HS Weak”, “HS Moderate”, “HS Strong”, dan “Abusive” [11]. Dari dasar label ini kemudian dimanipulasi menjadi 5 skenario yang berbeda. Ilustrasi dari label *dataset* dapat dilihat pada Gambar 4.

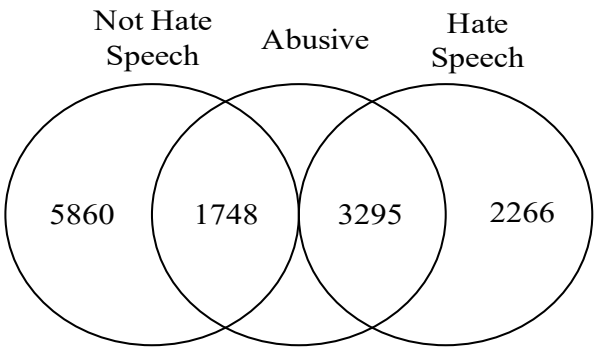
Pada literatur selanjutnya [12], pelabelan pada teks menggunakan variabel yang sama dengan literatur [11]. Skenario yang digunakan ada 2 yaitu untuk melakukan pengenalan terhadap keseluruhan label kemudian untuk melakukan pengenalan terhadap kategori dari label “Hate speech”. Kategori label yang dimaksud adalah “HS

Tabel 3. Sumber *dataset* untuk penelitian

No	Jenis	Literatur
1	The Dataset for Hate Speech Detection in Indonesian https://github.com/ialfina/id-hatespeech-detection	[6] [8] [14]
2	Multi-label hate speech https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection	[11] [12] [13]
3	Dataset dari peneliti	[7] [9] [10] [15]

Tabel 4. Sample data pada *dataset* satu label

Twit	Hate Speech
Rencana Bapak yang di surga itu lebih indah yang kita inginkan bapak ahok tetap semangat ya pak	0
Kami sebagai rakyat indonesia yang bukan berasal dari DKI turut mendoakan semoga pak Basuki dan pak Djarot tidak pernah lelah melayani kami dan tetap terus semangat	0
"FPI: Pendukung Ahok kalah jumlah, dengan izin Allah kita gusur Gubernur koruptor dan penista agama https:\/\/t.co\/Jp7pGLRGiE"	1
"Si penista agama, Jaga mulut aja gabisa apalagi jaga jkt?"	1



Gambar 5. Distribusi label pada *dataset*.

Individual", "HS Group", "HS Religion", "HS Race", "HS Physical", "HS Gender", dan "HS Other" [12].

Selajutnya adalah literatur [13] yang melakukan pengujian untuk membangun model dengan output multilabel. Literatur ini membangun *dataset*nya sendiri dan telah digunakan dalam literatur beberapa penelitian termasuk 2 penelitian sebelumnya yang tersitasi dalam sitasi [11] [12]. Ada 2 skenario yang digunakan dalam penelitiannya yaitu membangun model untuk 2 label yaitu label "Hate Speech" dan "Abusive". Kemudian skenario kedua adalah untuk membangun model untuk keseluruhan label seperti yang diilustrasikan pada Gambar 4.

Literatur terakhir yang membangun model *machine learning* dengan *output* multilabel adalah literatur [15]. Secara konsep, literatur ini hampir sama dengan literatur [13]. Hanya saja model yang dibangun hanya terbatas pada skenario untuk menghasilkan *output* sebanyak 3 yaitu kata kasar (*abusive*), ujaran kebencian (HS), dan bukan ujaran kebencian (Non HS).

3.2 Sumber *dataset*

Reviu dilakukan dengan cara difokuskan pada literatur yang melakukan pengujian untuk melakukan membangun model dengan masukan atau bahan *training* dan *testing* berupa teks dari twitter. Setiap literatur mempunyai cara tersendiri untuk membangun *dataset* yang digunakan. Penulis secara garis besar dapat mengelompokkan 3 jenis *dataset* yang digunakan pada 8 literatur yang diperoleh. Adapun sebaran mengenai penggunaan jenis *dataset* disajikan dalam Tabel 3.

3.2.1 Dataset pertama

Dataset pertama merupakan *dataset* hasil penelitian yang disebarkan secara publik melalui platform GitHub. *Dataset* ini berisikan sentimen-sentimen pada twitter yang berkaitan dengan peristiwa politik tentang pemilihan gubernur di DKI Jakarta.

Adapun mengenai teknik pengumpulan data (*crawling*), *dataset* ini memanfaatkan *library* "tweepy" pada bahasa pemrograman python untuk mendapatkan kumpulan twit dengan kata kunci "#DebatPilkadaDKI", "#SidangAhok", "Pilkada Jakarta 2017", dan lainnya. *Crawling* data kemudian mendapatkan jumlah teks sebanyak 1.100 [14].

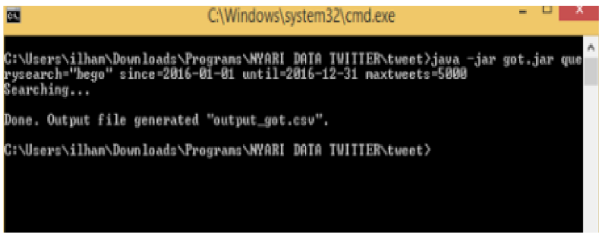
Tiap teks kemudian diproyeksikan untuk memiliki satu buah label "Hate Speech" atau label "ujaran kebencian". Tiap teks pada *dataset* dianotasi dengan memberikan nilai 0 dan 1 pada variabel label untuk menyatakan teks sebagai ujaran kebencian atau tidak. Dimana nilai 0 berarti bukan merupakan teks ujaran kebencian dan 1 untuk teks yang merupakan ujaran kebencian.

Dari hasil anotasi kemudian didapatkan 713 teks yang digunakan sebagai *dataset* dengan rincian: 250 teks ujaran kebencian dan 453 teks lainnya bukan termasuk ujaran kebencian [14]. Anotasi tersebut dilakukan oleh 30 orang volunter yang berasal dari mahasiswa di DKI Jakarta. Contoh data pada *dataset* pertama tersaji pada Tabel 4.

Dataset ini kemudian menjadi terpercaya dan dapat digunakan untuk penelitian terkait dengan deteksi ujaran kebencian pada teks. Hal ini dapat dilihat pada penggunaan *dataset* ini pada 3 literatur dalam reviu kali ini [6] [8] [14].

3.2.2 Dataset Kedua

Poin ke dua merupakan dataset yang secara garis besar memiliki 2 label dalam satu teks yaitu "Hate Speech" dan "Abusive". Dataset ini dibangun dengan menggabungkan hasil pembangunan dataset pada penelitian sebelumya ditambah dengan teks yang di crawling lagi pada twitter dengan menggunakan library "tweepy" pada bahasa pemrograman python. Dataset pada penelitian sebelumnya diambil dari dataset yang dihasilkan pada penelitian [14]. Jumlah teks yang dimuat dalam dataset ini adalah sebanyak 13.169 dengan rincian 5.561 teks yang termasuk ke dalam teks ujaran kebencian dan 7.608 lainnya tidak termasuk ujaran kebencian. Pada dataset, terdapat label "Abusive" yang bermakna teks yang mengandung kata-kata bersifat kasar. Teks dengan label "Abusive" belum tentu tergolong sebagai teks yang mengandung ujaran kebencian. Hal ini dapat dilihat dari hasil anotasi yang dilakukan oleh annotator terhadap dataset ini [13]. Distribusi dari teks berlabel "Abusive" dapat dilihat pada Gambar 5. Dataset ini digunakan oleh 3 literatur yang diperoleh [11] [12] [13].



Gambar 6. Contoh pengambilan data twitter pada *library* Jefferson Henrique.

Tabel 5. Teknik pre-processing pada teks

No	Jenis	Literatur
1	*Text Normalization*	[11] [13]
2	*Stemming*	[6] [7] [9] [10] [11] [12] [13]
3	*Stop word Removal*	[6] [8] [11] [12] [13]
4	*Word Filtering*	[7]
5	*n-gram*	[6] [11] [13] [14] [15]

3.2.3 Dataset ketiga

Terakhir, terdapat jenis *dataset* yang dibangun sendiri oleh para peneliti di masing-masing literatur [7] [9] [10]. Ada 2 cara yang dapat dilakukan untuk mendapatkan data teks pada twitter berdasarkan reviu yang sudah dilakukan. Cara pertama adalah dengan memanfaatkan API yang sudah disediakan oleh Twitter. Kemudian cara kedua adalah dengan memanfaatkan *opensource library* yang sudah ada.

Pemanfaatan twitter API dilakukan dengan cara memberikan kata kunci pada pengaksesannya. Literatur menggunakan kata kunci yang sesuai dengan kebutuhan peelitiannya dan kemudian melakukan filter teradap teks yang didapat menurut waktu posting dari tiap teks. Kata kunci yang digunakan adalah "#pilpres2019", "#2019gantipresiden", "#debatcapres", "debatpilpres" dan "#jokowi2periode". API berhasil memberikan 54.650 teks dengan rentang posting dari jam 11.00 hingga 19.00. Teks-teks tersebut kemudian disaring dengan hasil akhir sebanyak 597 teks dengan rincian 320 yang termasuk ke dalam ujaran kebencian dan 277 lainnya bukan termasuk ke dalam ujaran kebencian [7].

Cara kedua yaitu dengan memanfaatkan *opensource library* yang sudah ada. *Library* ini dibangun oleh Jefferson Henrique yang pada dasarnya akan mengakses API dari twitter. *Library* ini dapat didapatkan pada website dan akun github dari developernya. *Library* ini dapat langsung menyimpan teks hasil *crawling* langsung ke dalam format csv [9] [10]. Contoh penggunaan *library* ini disajikan pada Gambar 6.

3.3 Penerapan teknik *pre-processing*

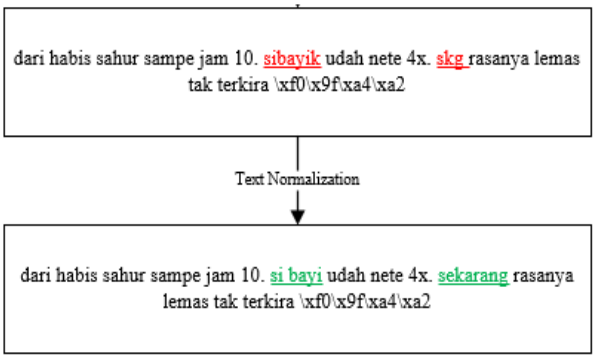
Teknik *pre-processing* merupakan hal yang menarik untuk direviu. *Pre-processing* dibutuhkan dalam pembangunan model *machine learning* karena akan sangat mempengaruhi performa model dalam melakukan pengenalan [13].

Reviu mengenai teknik *pre-processing* ini lebih kepada teknik yang digunakan setelah data melalui proses *cleaning*. *Cleaning* pada teks twitter pada tiap literatur akan menyesuaikan dengan kondisi *dataset* yang digunakan. Misalkan dengan cara menghilangkan kata-kata atau simbol yang tidak perlu seperti “RT”, username, link, dan karakter *non alpha numeric*.

Setelah melewati tahap *cleaning*, masing-masing literatur ada yang menerapkan teknik yang sama, namun ada juga yang berbeda. Secara garis besar, terdapat 3 teknik yang dapat diambil dari keseluruhan literatur pada tahap *pre-processing*. Teknik tersebut dapat dilihat pada Tabel 5.

Tabel 6. Jenis metode/algoritma machine learning yang digunakan

Kata Kunci	Kata Pengganti
sibayik	si bayi
skg	sekarang
nyariin	Mencarikan



Gambar 7. Ilustrasi penerapan normalisasi pada teks.

Tabel 5 memperlihatkan bahwa *Stemming* menjadi teknik yang paling sering digunakan yaitu pada 7 literatur. Kemudian untuk teknik yang dikombinasikan, *stemming* dan *Stop word removal* menjadi kombinasi teknik yang paling sering digunakan yaitu pada 4 literatur.

3.3.1 Text normalization

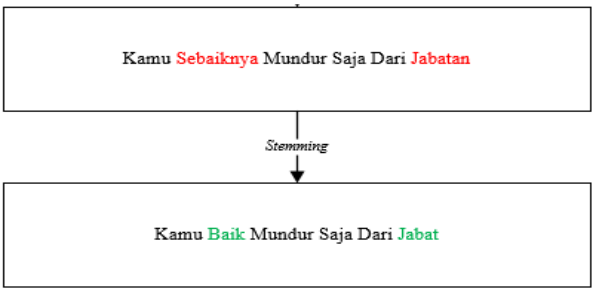
Text normalization atau normalisasi pada teks merupakan teknik yang digunakan untuk mengkonversi tiap kata pada teks menjadi bentuk yang lebih umum. Tiap kata pada teks akan dicek satu-persatu apakah terdapat kata dengan penulisan yang tidak umum. Jika kata tersebut ditemukan makan akan diganti dengan penulisan yang sebenarnya.

Normalisasi teks dapat dilakukan dengan membuat kamus khusus yang disesuaikan dengan *dataset* yang dimiliki. Kamus tersebut akan diacu ketika proses normalisasi berlangsung [13]. Contoh kosakata dari kamus untuk melakukan normalisasi pada teks dapat dilihat pada Tabel 6.

Tabel 6 menunjukkan perpaduan antara kata yang harus dinormalisasi (Kata Kunci) dengan kata yang digunakan untuk normalisasi (Kata Pengganti). Jika suatu teks mengandung memiliki kata yang terdaftar pada kolom Kata Kunci, maka kata tersebut akan diganti langsung dengan kata pada kolom Kata Pengganti yang sebaris dengan kata tersebut. Gambar 7 memperlihatkan bagaimana teknik normalisasi teks melakukan konversi terhadap kata pada teks. Pada teks yang di atas, terdapat 2 kata yang harus diganti (mengacu pada Tabel 7) yaitu “sibayik” dan “skg” yang ditandai dengan warna merah. Kedua kata tersebut kemudian diganti secara berurut dengan “si bayi” dan “sekarang” sehingga kalimat menjadi seperti kalimat dibawahnya.

3.3.2 Stemming

Teknik *Stemming* merupakan teknik yang diterapkan untuk mendapatkan kata dasar dari suatu kata. Dengan kata lain, teknik ini dilakukan untuk mendapatkan kata dasar dari teks yang berimbuhan. Tujuan dari dilakukannya teknik ini adalah agar ketika



Gambar 8. Ilustrasi penerapan normalisasi pada teks.

Gambar 9. Ilustrasi proses stop words removal.

pembobotan dilakukan, kata-kata yang memiliki kata dasar sama akan mendapat bobot yang sama [10].

Penerapan *stemming* pada teks berbahasa Indonesia dapat memanfaatkan *library* "sastrawi" pada *library* python [6] [11] [12] [13]. Ilustrasi dari penerapan stemming pada teks dapat dilihat pada Gambar 8. Pada Gambar 8, terdapat 2 kata yang memiliki imbuhan yaitu "Sebaiknya" dan "Jabatan". Selanjutnya, kata-kata tersebut dikonversi menjadi bentuk dasarnya yaitu secara berurut menjadi "Baik" dan "Jabat".

3.3.3 Stop word removal

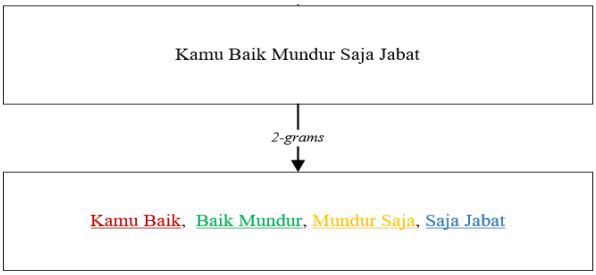
Stop word merupakan kata yang sangat umum dan keberadaannya tidak mengubah makna dari kalimat. Sesuai dengan namanya yang mengandung kata "removal", teknik ini akan menghilangkan kata-kata tersebut dalam kalimat [8].

Berdasarkan reviu terhadap literatur, terdapat 3 kamus yang dapat digunakan sebagai acuan menentukan kata pada teks tergolong ke dalam *stop word* atau tidak. Pertama adalah dengan menggunakan *library*. Python sudah menyediakan *library* untuk mendapatkan list dari *stop words* dengan bahasa Indonesia. *Library* tersebut adalah NLTK dan Sastrawi [6] [12]. Selain itu, *stop word* juga dapat dibuat secara mandiri dalam bentuk kamus. Kata-kata yang ada di dalam kamus ini disesuaikan *dataset* yang ada dan kebutuhan penelitian yang dilakukan [13]. Lebih jelas mengenai proses *stop words removal* dapat diihat pada Gambar 9. Gambar 9 merupakan ilustrasi penerapan *stop word removal* menggunakan *library* "sastrawi". Pada kalimat yang atas ditemukan satu kata yang tergolong sebagai *stop words*. Oleh karena itu, dalam prosesnya kalimat tersebut dihilangkan sehingga kalimat berubah menjadi "Kamu Baik Mundur Saja Jabat".

3.3.4 N-gram

Machine learning dalam melakukan pemodelan dengan cara melakukan perhitungan terhadap nilai numerik. Oleh karena itu, perlu dilakukan proses terlebih dahulu untuk melakukan pembobotan terhadap teks tersebut menjadi list bobot berupa angka.

Pembobotan pada *text analytic* didasarkan pada hasil tokenisasi. Umumnya, tokenisasi pada teks dilakukan dengan melakukan pemecahan teks menjadi kata perkata. Namun, terdapat metode n-gram yang memungkinkan untuk melakukan tokenisasi/pemecahan teks dengan tiap pecahan terdiri atas n kata. Tiap pecahan diambil dari urutan kata pada teks asli dengan perpindahan sebanyak satu kata [6]. Gambar 10 merupakan ilustrasi penerapan 2-grams pada teks. Dapat dilihat bahwa tiap pecahan teks terdiri atas 2 kata. Kata-kata tersebut masih berurutan dengan teks aslinya. Jumlah token yang dihasilkan pada Gambar 10 sebanyak 4 dari total 5 huruf pada teks sebelum diterapkan 2-grams.



Gambar 10. Ilustrasi penerapan 2-grams pada teks.

Tabel 7. Jenis metode/algoritma machine learning yang digunakan

No	Jenis	Literatur
1	Support Vector Machine (SVM)	[7] [11] [12] [13] [14] [15]
2	Random Forest Decision Tree (RFDT)	[11] [13] [14] [15]
3	Naïve Bayes (NB)	[6] [11] [13] [15]
4	Convolution Neural Network (CNN)	[12]
5	FastText	[8]
6	Recurrent Neural Network (RNN)	[9]
7	Backpropagation Neural Network	[10]

3.4 Metode yang digunakan

Dilihat dari perspektif metode atau algoritma yang digunakan pada setiap literatur, ada 2 hal yang dapat diambil dari literatur-literatur tersebut. Pertama adalah mengenai algoritma apa saja yang digunakan dalam penelitian di setiap literatur. Kedua adalah mengenai teknis yang dilakukan pada penelitian setiap literatur apakah melakukan perbandingan terhadap lebih dari satu algoritma atau melakukan optimalisasi hanya pada satu algoritma saja. Adapun sebaran dari metode yang digunakan oleh tiap literatur disajikan dalam Tabel 7.

Dari Tabel 7, dapat dilihat bahwa 3 algoritma yang paling sering digunakan pada literatur adalah algoritma SVM, RFDT, dan NB dengan masing-masing digunakan oleh 4, 2, dan 3 literatur. Satu literatur dapat memuat lebih dari satu metode/algoritma. Hal ini berkaitan dengan teknis dari penelitian yang dilakukan adalah berupa perbandingan metode yang digunakan.

3.4.1 Support Vector Machine (SVM)

Penelitian menggunakan SVM mendapatkan akurasi dengan rentang 61.67% - 74.88% [7] [11] [12] [13] [14] [15]. SVM diujikan dengan menggunakan 2 jenis data transformation yaitu Label Power Set (LP) dan Classifier Chains (CC). LP melakukan transformasi dari data dengan label yang banyak menjadi satu label dengan jumlah kelas yang banyak. Sedangkan CC melakukan

transformasi pada label dengan mengubah label dari tiap teks ke dalam bentuk binary. Model terbaik dari kedua kombinasi algoritma tersebut akan diperoleh dengan cara melakukan pengujian terhadap kernel, parameter regularisasi, dan nilai gamma. Hasil akhir dari pemodelan adalah didapatkannya performa model terbaik dari SVM yakni dengan mengkombinasikan CC sebagai data transformasinya. Model ini berhasil mendapatkan nilai akurasi maksimum sebesar 74.88% [12].

3.4.2 Random Forest Decision Tree (RFDT)

Penelitian menggunakan RFDT dilakukan pada 2 literatur. RFDT dikombinasikan dengan ekstraksi fitur menggunakan metode word uni-gram. Model yang dibangun berhasil melakukan klasifikasi terhadap teks yang berlabel lebih dari satu dengan performa akurasi yang cukup rendah yaitu 64.38%, 66.12%, dan 68.34% [11] [13] [15]. Masalah ini kemungkinan disebabkan oleh tidak imbangnya sebaran data untuk masing-masing label. Hal ini ditunjukkan oleh banyaknya nilai *false negatif* yang diperoleh pada saat pengujian terhadap model [13]. Selain itu, penerapan dari teknik pre-processing menjadi salah satu faktor penyebab dari kurang baiknya akurasi yang didapat [15]. Masih terdapat karakter tanda baca dan *stop word* masuk dalam pembobotan membuat mesin pemodelan kesulitan dalam mengenali fitur yang menjadi *inputnya* [15].

Di sisi lain, RFDT mampu dengan baik dalam melakukan klasifikasi terhadap data dengan satu label. Akurasi terbaik dari RFDT didapatkan sebesar 93.5%. Akurasi tersebut diperoleh untuk melakukan klasifikasi apakah teks tergolong sebagai teks yang mengandung ujaran kebencian atau tidak [14].

3.4.3 Naïve Bayes (NB)

Penelitian menggunakan naïve bayes dilakukan pada 2 jenis penelitian dengan tujuan yang berbeda. Maksudnya adalah metode ini digunakan melakukan klasifikasi terhadap data satu label dan data multi label [6] [11] [13]. Pada data satu label, model berhasil mendapatkan nilai akurasi yang baik yaitu 85% [6]. Sedangkan pada multilabel, model masih belum menemukan performa terbaiknya dimana akurasi yang diperoleh adalah sebesar 64.38% [11].

3.4.4 Neural Network

Selanjutnya terdapat penelitian menggunakan metode dengan dasar algoritma berupa Neural Network. Perkembangan dari metode neural network berhasil mendapatkan metode yang baru yang digunakan oleh literatur pada reviu literatur kali ini. Metode-metode tersebut adalah CNN, fastText, RNN, dan Backpropation Neural Networks.

3.4.4.1 Convolution Neural Network (CNN)

Metode CNN digunakan untuk melakukan klasifikasi teks dengan jumlah label lebih dari 1 [12]. Metode CNN pada kasus penelitian bekerja tidak cukup baik. CNN tidak cukup mampu untuk mengenali label dari teks prediksi yang diberikan. CNN diuji menggunakan label-label *low occurrence rate* yaitu "HS Race", "HS Physical", "HS Gender", dan "HS Strong". Dari 8 twit yang diuji, model hanya mampu mengklasifikasikan 6 twits saja. Hal ini terjadi kemungkinan karena jumlah data yang tidak seimbang untuk tiap labelnya. Kemudian pada *pre-processing* teks tidak dilakukan secara maksimal dengan tidak melakukan pembersihan terhadap *stop word*, tidak melakukan *stemming*. Analisis masalah juga muncul pada penggabungan *dataset* dari [13] dan twit yang diambil sendiri oleh peneliti. Permasalahan yang kemungkinan terjadi adalah karena ada twit dalam bahasa Inggris yang juga dimuat dalam *dataset*. Twit berbahasa Inggris tersebut kemudian mengalami masalah saat proses translasinya [12].

3.4.4.2 fastText

Metode fastText digunakan untuk mengklasifikasi teks ke dalam 2 jenis label yaitu label lebih dari satu dan satu label. FastText disusun dengan input berupa vektor. Komponen utama dari metode ini ada 3 yaitu *embedding layer*, *hidden layer*, dan *output layer*. Penelitian mengkombinasikan penggunaan sub-word dalam implementasinya. Selain itu dilakukan percobaan untuk memasukkan pre-train dari "wiki" pada *word embedding layer*. Model yang dihasilkan memiliki performa yang diukur menggunakan metrics F1 dengan nilai lebih baik ditunjukkan untuk klasifikasi pada teks satu label. Dengan komposisi yang sama, model memperoleh nilai F1 sebesar 87.3% untuk klasifikasi teks dengan satu label. Sedangkan untuk 2 label, model memberikan nilai F1 sebesar 85.6% [8].

3.4.4.3 Recurrent Neural Network (RNN)

Pendekatan lain dari neural network adalah dengan menggunakan metode RNN [9]. Penelitian ini menggunakan *dataset* yang berbeda dari penelitian pada literatur yang sudah disebutkan pada sub bagian ini. Penelitian pada literatur menitikberatkan pada pengujian terhadap akurasi yang diperoleh dengan perubahan pada nilai *epoch, learning rate*, dan ukuran *batch* pada data *training*. Hasil dari penelitian adalah berupa model RNN yang bekerja sangat baik dalam mengklasifikasikan teks apakah merupakan ujaran kebencian atau tidak. Model mampu mendapatkan nilai akurasi terbaik sebesar 93%. Adapun nilai terbaik dari parameter uji untuk *epoch, learning rate*, dan ukuran *batch* secara berturut-turut mendapatkan nilai 150, 0.007, dan 64 [9].

3.4.4.4 Backpropagation

Terakhir adalah yang menggunakan metode Backpropagation. Literatur ini melakukan pengujian terhadap model untuk melakukan klasifikasi terhadap teks twitter. klasifikasi yang dilakukan merupakan klasifikasi terhadap satu label data. Teks yang terlibat dalam penelitian sebelumnya melalui proses TF-IDF terlebih dahulu. Proses ini diperlukan untuk mengkonversi nilai dari setiap kata pada teks menjadi deretan angka yang siap untuk dikalkulasi pada saat pembangunan model. Performa dari model dalam melakukan klasifikasi sudah baik dengan nilai akurasi mencapai 89.47% [10]. Namun nilai ini masih lebih rendah jika dibandingkan dengan model dengan RNN [9].

Dari penjabaran di atas, tiap metode memiliki nilai performanya masing-masing. Perbandingan tingkat akurasi dari masing-maisng metode dengan lingkup penelitiannya disajikan dalam Tabel 8 dan Tabel 9.

Berdasarkan Tabel 8 dapat dilihat bahwa tingkat akurasi tertinggi diperoleh pada literatur dengan menggunakan algoritma RNN. Kemudian untuk akurasi terendah didapatkan sebesar 61.67% dengan metode yang digunakan adalah SVM.

Berdasarkan Tabel 9, dapat dilihat bahwa model yang dibangun untuk melakukan klasifikasi terhadp teks dengan jumlah label lebih dari 1 memberikan hasil akurasi di bawah 80%. Hanya satu model yang berhasil menembus angka 70% yaitu model SVM dengan kombinasi data transformation berupa CC.

4. PEMBAHASAN

Hal yang pertama kali menjadi catatan dalam melakukan reviu mengenai literatur-literatur yang sudah melakukan penelitian mengenai ujaran kebencian dalam bahasa Indonesia adalah keberadaan *dataset* yang dapat diakses secara publik dan telah

Tabel 8. Perbandingan tingkat akurasi pada model dengan klasifikasi pada satu label

No	Metode	Literatur	Akurasi
1	SVM	[7]	61.67%
2	NB	[6]	85.00%
3	fastText	[8]	87.30%
4	RNN	[9]	91.00%
5	Backpropagation	[10]	89.47%

Tabel 9. Perbandingan tingkat akurasi pada model dengan klasifikasi pada lebih dari 1 label

No	Metode	Literatur	Akurasi
1	SVM+CC	[12]	74.88%
2	SVM	[11]	68.43%
3	RFDT	[13]	66.12%
4	fastText	[8]	69.40%

melalui proses anotasi yang menjadikan *dataset* tersebut valid untuk digunakan sebagai bahan penelitian [13][14]. Bukti dari kelayakan *dataset* ini adalah dengan penggunaannya pada 5 literatur yang didapatkan pada reviu kali ini.

Selanjutnya mengenai reviu pada model yang dihasilkan pada tiap literatur. Secara performa, literatur yang melakukan klasifikasi terhadap teks dengan label berjumlah hanya 1 memiliki tingkat akurasi yang baik. Hal ini dapat dilihat dari sebagian besar literatur berhasil mendapatkan akurasi di atas 80% [6] [8] [9] [10]. Kecuali model yang dibangun dengan model SVM [7]. Akan tetapi, pada model yang ditujukan untuk melakukan klasifikasi dengan jumlah label lebih dari 1, model di tiap literatur belum mampu memberikan nilai performa akurasi di atas 80%. Model terbaik dihasilkan menggunakan metode fastText yaitu 69.4% untuk jumlah label sebanyak 3 [8]. Sedangkan untuk jumlah label yang lebih banyak lagi, model terbaik didapatkan pada penggunaan SVM dengan tingkat akurasi sebesar 68.43% [11].

Alasan mengenai performa yang lemah dari pemodelan terhadap jumlah label data yang lebih dari 1 adalah pada masalah data. Ketidaksemimbangan jumlah data pada beberapa label mengakibatkan model kurang mampu dalam mengenali dengan baik teks pada label tertentu. Hal ini tentunya akan berakibat pada nilai akurasi model secara kumulatif untuk seluruh label. Selain itu, tahapan pada *pre-processing* juga mesti diperhatikan. Teknik apa saja yang diterapkan pada saat *pre-processing* akan sangat berpengaruh kepada bagaimana cara model untuk melakukan pembelajaran terhadap data yang diberikan [12].

Berdasarkan hasil reviu yang sudah dilakukan, terdapat beberapa saran yang dapat diajukan untuk penelitian selanjutnya antara lain:

a. Membangun model *machine learning* untuk mengoptimalkan performa model dalam melakukan klasifikasi terhadap data multilabel.
b. Menentukan model terbaik dengan parameter uji berupa teknik *pre-processing* yang diterapkan.
c. Algoritma *Neural Network* (RNN dan CNN) memiliki performa yang sangat baik dalam melakukan klasifikasi terhadap data satu label. Algoritma tersebut sangat menarik jika diuji untuk melakukan klasifikasi terhadap data multilabel.
d. Melakukan pengujian menggunakan data yang sudah dipublikasikan untuk publik. Hal ini bertujuan untuk memudahkan dalam melakukan komparasi terhadap penelitian yang sudah ada. Misalkan pada *dataset* nomor 1 dan 2 pada Tabel 3.

5. KESIMPULAN

Berdasarkan reviu yang sudah dilakukan terhadap literatur-literatur yang diperoleh, didapatkan bahwa kasus ujaran kebencian pada teks twitter sudah dilakukan dengan menggunakan beberapa metode klasifikasi dengan masing-masing *dataset* sebagai bahan penelitiannya. Penggunaan *dataset* sangat berpengaruh ke dalam performa yang akan diberikan oleh model *machine learning*. Penggunaan *dataset* ini maksudnya adalah mengenai bagaimana keseimbangan data antara kelas yang satu dengan kelas yang lain.

Kasus dari ujaran kebencian kemudian dapat dipisahkan ke dalam 2 lingkup penelitian yaitu untuk membangun model 1 label dan lebih dari 1 model. Model 1 label ditujukan untuk membangun model yang dapat mengenali suatu teks mengandung unsur ujaran kebencian atau tidak. Sedangkan lingkup model lebih dari 1 label ditujukan untuk membangun model yang mampu melakukan klasifikasi terhadap teks dengan jumlah label lebih dari 1.

6. REFERENSI

[1] Kemp, S. 2021. *Digital in Indonesia: All the Statistics You Need in 2021 — DataReportal – Global Digital Insights*. https://datareportal.com/reports/digital-2021-indonesia (diakses 08 Mei 2021).

[2] Briliani, A., Irawan, B. & Setianingsih, C. 2019. 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*. 98–104. DOI: 10.1109/IoTaIS47347.2019.8980398.

[3] Hukmana, S. Y. 2021. 125 Akun Medsos Terjaring Virtual Police - Medcom.id. https://www.medcom.id/nasional/hukum/gNQ5RnwN-125-akun-medsos-terjaring-virtual-police (diakses 08 Mei 2021).

[4] Nurul, F. A., Nurhadi, N. & Pranawa, S. 2020. Konflik dan Ujaran Kebencian di Twitter (Studi Tentang Hashtag #2019TetapJokowi and #2019GantiPresiden Periode Januari-Februari 2019). *Jupiis J. Pendidik. Ilmu-Ilmu Sos.* 12, 1, 132. DOI: 10.24114/jupiis.v12i1.16083.

[5] Devita, P. 2021. 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 (diakses 08 Mei 2021).

[6] Alzami, F., Purinsyira, N. P., Anggi, R. P., Megantara, R. A. & Prabowo, D. P. 2020. Sentiment Analysis Untuk Deteksi Ujaran Kebencian Pada Domain Politik. *Science and Engineering National Seminar*. 5, 1, 213–218.

[7] Lyrawati, D. P. N. 2019. Deteksi Ujaran Kebencian Pada Twitter Menjelang Pilpres 2019 dengan Machine Learning. *J. Ilm. Mat.* 7, 2, 104–110.

[8] Herwanto, G. B., Ningtyas, A. M., Nugraha, K. E. & Trisna, I. N. P. 2019. Hate Speech and Abusive Language Classification using fastText. *2019 2nd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2019*. 69–72. DOI: 10.1109/ISRITI48646.2019.9034560.

[9] Saksesi, A. S., Nasrun, M. & Setianingsih, C. 2018. Analysis Text of Hate Speech Detection Using Recurrent Neural

Network. *Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018*. 242–248. DOI: 10.1109/ICCEREC.2018.8712104.

[10] Setyadi, N. A., Nasrun, M. & Setianingsih, C. 2018. Text Analysis for Hate Speech Detection Using Backpropagation Neural Network. *Proc. - 2018 Int. Conf. Control. Electron. Renew. Energy Commun. ICCEREC 2018*. 159–165. DOI: 10.1109/ICCEREC.2018.8712109.

[11] Prabowo, F. A., Ibrohim, M. O. & Budi, I. 2019. 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*. 1–5. DOI: 10.1109/ICITACEE.2019.8904425.

[12] Hana, K. M., Adiwijaya, Al Faraby, S. & Bramantoro, A. 2020. Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines. *Int. Conf. Data Sci. Its Appl. ICoDSA 2020*. DOI: 10.1109/ICoDSA50139. 2020.9212992.

[13] Ibrohim, M. O. & Budi, I. 2019. Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*. 46–57. DOI: 10.18653/v1/w19-3506.

[14] Alfina, I., Mulia, R., Fanany, M. I. & Ekanata, Y. 2018. Hate speech detection in the Indonesian language: A dataset and preliminary study. *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*. 2018-January, October, 233–237. DOI: 10.1109/ICACSIS.2017.8355039.

[15] Tjahyanti, L. P. A. S. 2020. Pendeteksian Bahasa Kasar (Abusive Language) Dan Ujaran Kebencian (Hate Speech) Dari Komentar Di Jejaring Sosial. *J. Chem. Inf. Model.* 7, 9, 1689–1699.