INFORMATICS AND DIGITAL EXPERT (INDEX) - VOL. 4 NO. 2 (2022) 46-52
Terbit *online* pada laman web jurnal : http://index.unper.ac.id
| ISSN (Print) 2775-2208 | ISSN (Online) 2715-0453 |

Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia

Rija Muhamad Yazid [1], Fajri Rakhmat Umbara[2], Puspita Nurul Sabrina[3]

[1,2,3]*Universitas Jenderal Achmad Yani Cimahi, Jl. Terusan Jend. Sudirman, Kota Cimahi 40531, Indonesia*
e-mail: rijamyazid17 @if.unjani.ac.id [1], fajri.rakhmat@lecture.unjani.ac.id[2], puspita.sabrina@lecture.unjani.ac.id[3]

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi : 13 Februari 2022
Revisi Akhir : 04 Juni 2022
Diterbitkan *Online* : 30 November 2022

Kata Kunci:
Ujaran Kebencian, Tingkat Ancaman, Naïve Bayes, N-Gram, Multi-Label

Korespondensi:
Telepon / Hp : 082319595671
E-mail : **rijamyazid17 @if.unjani.ac.id**

A B S T R A K

Ujaran kebencian adalah ungkapan atau bahasa yang digunakan untuk mengekspresikan kebencian terhadap seseorang atau sekelompok orang. Ujaran kebencian juga memiliki tingkatan ancaman, semakin tinggi tingkat ancaman ujaran kebencian maka akan semakin luas dan cepat penyebarannya sehingga dapat menimbulkan konflik antar individu sampai konflik antar kelompok. Tujuan penelitian ini adalah untuk mendeteksi dan mengklasifikasikan ujaran kebencian sekaligus tingkat ancamannya, data yang digunakan adalah dataset multi-label dari penelitian sebelumnya dengan menggunakan label yang masuk kedalam topik ujaran kebencian dan tingkat ancaman dengan total sebanyak 4 label. Dalam menyelesaikan permasalahan multi-label tersebut digunakan metode Naïve Bayes sebagai metode klasifikasi dan metode Label Power-set sebagai metode transformasi data, dalam penelitian ini juga digunakan pembobotan TF-IDF sekaligus melakukan beberapa skenario penelitian berdasarkan metode ekstraksi fitur n-gram. Hasil terbaik yang didapatkan berdasarkan hasil evaluasi F-score adalah sebesar 64,957% ketika menggunakan kombinasi metode ekstraksi fitur word unigram, word bigram dan character quadgram. Dari penelitian ini juga didapatkan bahwa semakin banyak fitur yang digunakan maka semakin baik nilai hasil evaluasinya terhadap jenis dataset yang digunakan.

1. PENDAHULUAN

Dalam beberapa tahun terakhir media sosial menjadi salah satu teknologi yang paling sering digunakan karena kemudahannya sebagai media untuk mendapatkan berita, penyebaran informasi dan komunikasi dari satu pengguna ke pengguna lain. Selain itu kebebasan berpendapat yang ditawarkan media sosial membuat semakin banyak diminati pengguna. Akan tetapi karena kemudahan dan kebebasan yang ditawarkan ini juga membuat pengguna menjadi lebih mudah dan leluasa dalam membuat dan menyebarkan ujaran kebencian [1].

Sampai saat ini masih belum ada definisi umum mengenai ujaran kebencian, berdasarkan penelitian [2] ujaran kebencian didefinisikan sebagai ungkapan atau bahasa yang digunakan untuk mengekspresikan kebencian dengan tujuan untuk menyerang dan menginspirasi seseorang atau sekelompok orang untuk menyakiti individu atau kelompok lain berdasarkan identitas yang dimiliki, dampak yang dapat ditimbulkan dari ujaran kebencian berbeda-beda tergantung dari seberapa besar ancaman dari ujaran kebencian tersebut. Dalam penelitian [3] ujaran kebencian ditentukan kedalam tiga tingkatan ancaman, yaitu ancaman tingkat rendah, menengah dan tinggi. Tingkatan ancaman ini ditentukan berdasarkan pola kata penulisan, banyaknya target, bentuk ujaran kebencian dan radius konflik yang dapat ditimbulkan. Semakin tinggi tingkat ancaman ujaran kebencian maka akan semakin cepat dan meluas konflik yang dapat ditimbulkan. Oleh karena itu diperlukan pendeteksian otomatis terhadap ujaran kebencian untuk menghindari peluasan penyebaran konflik sekaligus menentukan tingkat ancamannya untuk mempermudah pihak berwajib dalam memprioritaskan penangkapan penyebar ujaran kebencian.

Banyak penelitian sudah dilakukan untuk dapat mendeteksi ujaran kebencian ini, seperti yang dilakukan dalam penelitian [4][5] topik ujaran kebencian yang dibahas adalah topik ujaran kebencian dalam cakupan umum seperti agama, ras, suku dan jenis kelamin pada dataset yang didapatkan dari *Twitter* dengan kelas klasifikasi masuk kedalam ujaran kebencian atau bukan, dalam penelitian [6][7] topik yang dibahas tidak hanya mendeteksi ujaran kebencian namun juga mendeteksi perkataan kasar yang ada dalam teks, selanjutnya dalam penelitian [8] hampir sama dengan deteksi ujaran kebencian dalam penelitian ini dilakukan deteksi terhadap perkataan kasar, dalam penelitian tersebut dijelaskan bahwa tidak selalu setiap penggunaan bahasa kasar yang ada dalam teks adalah teks ujaran kebencian. Namun dalam penelitian-penelitian tersebut tidak dibahas mengenai topik tingkatan ancaman ujaran kebencian.

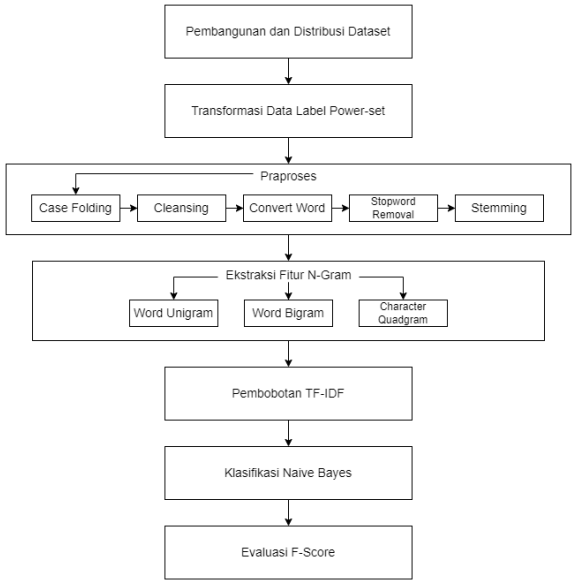
Untuk penelitian yang membahas mengenai ujaran kebencian sekaligus tingkatan ancamnnya masih belum banyak dilakukan, dalam hasil penelitian [1] mengenai trending topik ujaran kebencian dari tahun 1992 sampai tahun 2019 tidak ditemukan topik yang membahas mengenai tingkat ancaman ujaran kebencian. Sejauh ini hanya ditemukan beberapa penelitian yang membahas ujaran kebencian sekaligus tingkatan ancamannya, salah satunya dalam penelitian [9] dilakukan klasifikasi ujaran kebencian dengan tingkat ancamannya kedalam tiga kelas yaitu 'No_Hate', 'Weak_Hate' dan 'Stong_Hate' dalam penelitian tersebut digunakan metode klasifikasi SVM dan LSTM dengan hasil terbaik didapatkan ketika menggunakan SVM dengan akurasi sebesar 64.61%, selanjutnya dalam penelitian [3] digunakan dataset ujaran kebencian *multi-label* Twitter berbahasa Indonesia dengan 2 skenario penelitian, skenario pertama yaitu melakukan klasifikasi berdasarkan label ujaran kebencian dan perkataan kasar, skenario kedua melakukan klasifikasi berdasarkan label ujaran kebencian, perkataan kasar, target, kategori dan tingkatan ancaman ujaran kebencian dengan hasil terbaik didapatkan dengan menggunakan metode klasifikasi RFDT dan transformasi data Label Power-set pada kedua skenario. Namun dalam penelitian tersebut label perkataan kasar, target dan kategori juga masuk kedalam tahapan klasifikasi sehingga untuk klasifikasi yang secara khusus membahas mengenai ujaran kebencian dan tingkatan ancamannya masih belum dilakukan, jumlah label yang digunakan dapat mempengaruhi hasil akurasi akhir [10].

Dalam penelitian ini dilakukan klasifikasi terhadap ujaran kebencian dan tingkatan ancamannya dengan menggunakan metode klasifikasi Naïve Bayes, metode ini digunakan karena dapat memberikan akurasi yang baik dalam mengklasifikasikan teks ujaran kebencian [8][11]. Dataset yang digunakan dalam penelitian ini menggunakaan dataset *multi-label* ujaran kebencian *Twitter* yang digunakan dalam penelitian [3] dengan menggunakan label ujaran kebencian dan tingkat ancaman ujaran kebencian. Karena bentuk dataset yang digunakan menggunakan dataset *multi-label* sedangkan metode klasifikasi Naïve Bayes tidak bisa secara langsung melakukan klasifikasi terhadap dataset *multi-label* [3] maka dalam penelitian ini digunakan metode transformasi data Label Power-set untuk mentranformasikan data *multi-label* kedalam bentuk *single-label multi-class*, hal ini dilakukan karena dalam penelitian [8][11] klasifikasi dilakukan terhadap dataset berbentuk *single-label*.

Kemudian dalam penelitian ini juga dilakukan beberapa skenario penelitian berdasarkan penggunaan metode ekstraksi fitur *n-gram* untuk menemukan kombinasi metode ekstraksi fitur mana yang dapat memberikan hasil akurasi yang paling tinggi terhadap dataset yang digunakan, metode n-gram yang digunakan dalam penelitian ini yaitu *word n-gram* (*word unigram* dan *word bigram*) dan *chacarater n-gram* (*chracater quadgram*) [10], pembobotan kata menggunakan metode pembobotan kata TF-IDF dengan metode evaluasi menggunakan F-Score.

2. METODE PENELITIAN

Pada bagian ini dijelaskan mengenai urutan tahapan metode yang digunakan dalam penelitian, untuk gambar tahapan metode penelitian dapat dilihat pada Gambar 1.



Gambar 1. Metode Penelitian

2.1. Pembangunan dan Distribusi Dataset

Dataset yang digunakan dalam penelitian ini merujuk dari dataset yang digunakan dalam penelitian [3]. Dataset tersebut adalah dataset *multi-label* ujaran kebencian *Twitter* berbahasa Indonesia dengan total label sebanyak 12 label dengan penamaan dan kategorisasi label sebagai berikut, label *'HS'* adalah label yang masuk kedalam kategori ujaran kebencian, label *'Abusive'* adalah label yang masuk kedalam kategori perkataan kasar, label *'HS_Individual'*, *'HS_Group'* adalah label yang masuk kedalam kategori target ujaran kebencian, label *'HS_Religion'*, *'HS_Race'*, *'HS_Phsycal'*, *'HS_Gender'* dan *'HS_Other'* adalah label yang masuk kedalam kategori golongan ujaran kebencian, label *'HS_Weak'*, *'HS_Moderate'* dan *'HS_Strong'* adalah label yang masuk kedalam kategori tingkat ancaman kebencian. Setiap label memiliki kelas yang berbentuk *binary-class*.

Berdasarkan tujuan yang dilakukan dalam penelitian ini yaitu mendeteksi dan mengklasifikasikan ujaran kebencian sekaligus menentukan tingkatan ancamannya maka dari dataset referensi tersebut digunakan label *'HS'* untuk menentukan ujaran kebencian dan label *'HS_Weak'*, *'HS_Moderate'* dan *'HS_Strong'* untuk menentukan tingkat ancamannya.

Tabel 1. Jumlah dan Distribusi Dataset

Label				*Data Latih*	*Data Uji*
HS	***HS_ Weak***	***HS_ Moderate***	***HS_ Strong***		
0	0	0	0	4.654	1.995
1	0	0	1	317	136
1	0	1	0	1.095	470
1	1	0	0	2.199	943
Total				8.265	3.544

Setelah dataset dibangun kemudian dataset didistribusikan kedalam data latih dan data uji dengan perbandingan 7:3. Dari total keseluruhan 11.809 data sebanyak 8.265 data masuk sebagai data latih dan 3.544 data masuk sebagai data uji. Sedangkan untuk jumlah data setiap kombinasi label terdapat 6.649 data masuk kedalam label bukan ujaran kebencian (0, 0, 0, 0), 453 data masuk kedalam label ujaran kebencian dengan tingkat ancaman tinggi (1, 0, 0, 1), 1.565 data masuk kedalam label ujaran kebencian dengan tingkat ancaman menengah (1, 0, 1, 0) dan 3.142 data masuk kedalam label ujaran kebencian dengan tingkat ancaman rendah (1, 1, 0 ,0). Jumlah dan distribusi dataset dapat dilihat pada Tabel 1.

2.2. Transformasi Data Label Power-set

Dengan menggunakan metode label Power-set data *multi-label* ditransformasikan kedalam data yang berbentuk *single-label multi-class*. Proses transformasi dilakukan berdasarkan nilai kombinasi unik label pada data *multi-label*, dari kombinasi unik label tersebut dibuat label kelas baru yang dapat merepresentasikan setiap kombinasi label dalam data [12].

2.3. Praproses

Tahapan praproses dilakukan untuk mengubah data kedalam bentuk yang siap digunakan pada tahapan selanjutnya dengan cara menghilangkan atau mengurangi *noise* serta dengan melakukan normalisasi agar setiap data mempunyai format yang sama. Dalam penelitian ini terdapat lima tahapan praproses yaitu tahapan *case folding*, *cleansing*, *convert word*, *stopword removal* dan *stemming* [13][14].

Case Folding, tahapan ini digunakan untuk mentransformasikan keseluruhan teks dalam dokumen kedalam bentuk yang sama, sebagai contoh sistem akan menggolongkan kata “INDONESIA” dan “indonesia” kedalam hal yang berbeda, oleh karena itu teks tersebut perlu di normalisasikan terlebih dahulu kedalam bentuk yang sama baik kedalam bentuk huruf kapital (uppercase) atau huruf kecil (lowercase). Dalam penelitian ini setiap teks akan ditransformasikan kedalam huruf kecil.

Cleansing, tahapan ini digunakan untuk menghilangkan tanda baca (*punctuation*), spasi berlebih dan simbol dalam teks, tanda baca yang akan dihilangkan diantaranya yaitu tanda baca titik (.), tanda baca koma (,), tanda baca tanda tanya (?), tanda baca tanda seru (!) dan simbol simbol lain selain huruf, angka dan spasi (*whitespace*).

Convert Word, tahapan ini digunakan untuk melakukan normalisasi terhadap kata yang terdapat kesalahan penulisan (*typo*), terdapat singkatan dan kata gaul, sebagai contoh untuk kata “kamu” dalam media sosial sering disingkat menjadi “kmu, km, u”, kata-kata tersebut memiliki makna yang sama akan tetapi karena penulisannya berbeda maka menghasilkan fitur yang berbeda dan dapat mempengaruhi hasil klasifikasi.

Stopword Removal, tahapan ini digunakan untuk menghilangkan kata yang sering muncul pada dokumen sehingga tidak terlalu memberikan makna. Kata yang biasanya sering muncul dan dihilangkan diantaranya adalah kata-kata seperti “yang, di, dan, atau” dan kata lain yang serupa.

Stemming, tahapan ini digunakan untuk mentransformasikan teks kedalam bentuk dasarnya dengan cara menghilangkan imbuhan dari teks tersebut, terdapat beberapa jenis imbuhan yang dihilangkan yaitu imbuhan awalan (meng-, ber-, ter-, …), imbuhan akhiran (-an, -kan, -i, …), imbuhan awalan dan akhiran (ke-an, ber-an, pe-an, …) dan imbuhan sisipan (-el-, -em-, -se-, …).

2.4. Ekstraksi Fitur N-Gram

Proses ekstraksi fitur *n-gram* adalah tahapan yang dilakukan untuk mengambil ciri dari setiap kalimat dalam dokumen dan mentransformasikannya menjadi fitur yang digunakan dalam metode klasifikasi. Dalam penelitian ini digunakan tiga jenis metode ekstraksi yaitu *word n-gram* (*unigram* dan *bigram*) dan *character n-gram* (*quadgram*)

Word Unigram, metode ekstraksi fitur *word unigran* mengambil ciri dari teks dengan mengambil fitur berdasarkan kemunculan setiap kata pada teks.

Word Bigram, metode ekstraksi fitur *word bigram* mengambil ciri dari teks dengan mengambil fitur berdasarkan kemunculan setiap dua kata pada teks.

Character Quadgram, metode ekstraksi fitur *character quadgram* mengambil ciri dari teks dengan mengambil fitur berdasarkan kemunculan setiap empat karakter pada teks.

2.5. Pembobotan TF-IDF

Perhitungan *Term Frequency - Invers Document Frequency* (TF-IDF) didapatkan dari hasil perkalian antara persamaan *Term Frequency* dan persamaan *Invers Document Frequency*.

Dalam persamaan Te*rm Frequency* (TF) dihitung seberapa sering frekuensi suatu fitur (*term*) muncul dalam sebuah teks, semakin sering fitur tersebut muncul maka nilai TF akan semakin tinggi. Sedangkan dalam persamaan *Invers Document Frequency* (IDF) nilai IDF sebuah fitur akan semakin kecil apabila fitur tersebut sering muncul pada banyak teks dan nilainya akan semakin besar jika fitur tersebut hanya muncul pada satu atau sedikit teks [15].

Perhitungan TF-IDF yang digunakan dalam penelitian menggunakan persamaan (1).

$$TF_{IDF(ij)} = \text{tf}_{ij} * (\log\left(\frac{N}{n_i}\right) + 1) \quad (1)$$

Dimana tf_{ij} adalah jumlah kemunculan fitur i dalam teks j, N adalah jumlah banyaknya teks dan n_i adalah jumlah teks yang memiliki kemunculan fitur i minimal sebanyak satu kali.

2.6. Klasifikasi Naïve Bayes

Metode klasifikasi Naïve Bayes adalah metode klasifikasi statistik berdasarkan teorema Bayes yang dapat digunakan untuk memprediksi kelas sebuah data berdasarkan sekumpulan atribut dari kelas yang sudah ada. Dalam Naïve Bayes nilai dari setiap atribut dalam sebuah kelas tidak saling bergantung satu sama lain (*independent*) oleh karena itu metode klasifikasi ini sering disebut juga metode klasifikasi Naïve [16]. Rumus teorema Bayes dapat dilihat dalam persamaan (2).

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad (2)$$

Dimana nilai P(H|X) adalah probabilitas kelas H terhadap sekumpulan fitur X yang diuji (*posterior probability*), P(X | H) adalah probabilitas sekumpulan fitur X dalam kelas H (*conditional probability*), P(H) adalah probabilitas kemunculan kelas H dalam dataset (*class prior probability*) dan P(X) adalah probabilitas sekumpulan fitur (X) dalam dataset (*predictor prior probability*).

Karena nilai dari P(X) (*predictor prior probability*) yang digunakan akan selalu konstan dan sama dalam memprediksi setiap kelas maka yang perlu dikalkulasikan untuk mendapatkan nilai *maksimum posterior probability* hanyalah P(X | H) P(H) [16], sehingga rumusnya menjadi persamaan (3).

$$P(H|X) = P(X|H)P(H) \quad (3)$$

Untuk melakukan perhitungan *class prior probability* terhadap kelas yang diberikan digunakan persamaan (4).

$$P(H_i) = \frac{N_i}{N} \quad (4)$$

Dimana N_i adalah jumlah teks pada kelas *i* dan *N* adalah jumlah banyaknya dokumen.

Untuk menghitung conditional probability terhadap kelas yang diberikan, dalam Naïve Bayes nilai setiap atribut tidak bergantung dengan nilai pada atribut lainnya sehingga untuk menghitung nilai *P(X | H)* digunakan persamaan (5).

$$P(X|H_i) = \prod_{k=1}^{n} P(x_k|H_i) = P(x_1|H_i) \times \ldots \times P(x_n|H_i) \quad (5)$$

Dimana $P(x_n|H_i)$ adalah probabilitas fitur *x* yang ke n dalam kelas H_i. Untuk menghitung nilai setiap $P(x_n|H_i)$ digunakan *Laplace smoothing* untuk menghindari pembagian dengan nol [17] dengan menggunakan persamaan (6).

$$P(x_n|H_i) = \frac{T_{ni} + 1}{T_i + \beta} \quad (6)$$

Dimana T_{ni} adalah jumlah nilai fitur ke *n* dalam kelas H_i, T_i adalah jumlah keseluruhan nilai fitur dalam kelas H_i β adalah jumlah fitur dalam data latih.

Sehingga persamaan Naïve Bayes yang digunakan dalam penelitian adalah sebagai berikut :

$$P(H|X) = P(H) \prod_{k=1}^{n} P(x_k|H) \quad (7)$$

2.7. Evaluasi F-Score

F-score didefinisikan sebagai bobot rata-rata dari *precision* dan *recall* [16]. Perhitungan evaluasi F-score yang digunakan dalam penelitian menggunakan persamaan (8).

$$F - score: F_1 = 2 * \frac{P * R}{P + R} \quad (8)$$

Dimana P menyatakan *precision* dan R menyatakan *recall*. *Precision* adalah rasio hasil observasi yang diprediksi dengan benar t erhadap total observasi positif yang diprediksi. *Recall* adalah rasio hasil observasi yang diprediksi dengan benar terhadap semua data yang sebenarnya bernilai positif. Persamaan *precision* dan *recall* dapat dilihat dalam persamaan (9) dan (10).

$$P = \frac{tp}{tp + fp} \quad (9)$$

$$R = \frac{tp}{tp + fn} \quad (10)$$

Dimana *tp* adalah *true positives* yaitu jumlah hasil prediksi benar yang berarti jika kelas sebenarnya bernilai *true* maka nilai kelas prediksi juga bernilai *true*, *fp* adalah *false positives* yaitu jumlah hasil prediksi salah yang dimana jika kelas sebenarnya bernilai *false* tetapi nilai kelas prediksi bernilai *true*, *fn* adalah *false negatives* yaitu jumlah hasil prediksi salah yang dimana jika kelas sebenarnya bernilai *true* tetapi nilai kelas prediksi bernilai *false*.

3. HASIL DAN PEMBAHASAN

Berdasarkan hasil analisis yang dilakukan terhadap dataset, ditemukan 4 buah kombinasi label unik dari dataset *multi-label* yang digunakan yaitu kombinasi label (0, 0, 0, 0), (1, 0, 0, 1), (1, 0, 1, 0) dan (1, 1, 0, 0), dengan menggunakan metode label power-set dari kombinasi unik tersebut dibuat label kelas baru sebagai berikut {1=(0, 0, 0, 0), 2=(1, 0, 0, 1), 3=(1, 0, 1,

0), 4=(1, 1, 0, 0)}. Label hasil yang sudah ditransformasikan dapat dilihat pada Tabel 2.

Tabel 2. Label Hasil Transformasi

Label Lama				Label Baru
HS	HS_W	HS_M	HS_S	
0	0	0	0	1
1	0	0	1	2
1	0	1	0	3
1	1	0	0	4

Kemudian dalam penelitian ini dilakukan beberapa skenario penelitian berdasarkan kombinasi penggunaan metode ekstraksi fitur n-gram, hasil dari setiap skenario tersebut dapat dilihat pada Tabel 3.

Tabel 3. Hasil Evaluasi Setiap Kombinasi Metode Ekstraksi Fitur

Kombinasi Metode Ekstraksi Fitur	F-score (%)
Word unigram	60,777
Word bigram	53,698
Character quadgram	61,338
Word unigram + Word bigram	63,397
Word unigram + Character quadgram	62,096
Word bigram + Character quadgram	64,636
Word unigram + Word bigram + Character quadgram	64,957

Seperti yang terlihat pada Tabel 3 diantara penggunaan metode ekstraksi fitur *word unigram*, *word bigram* dan *character quadgram* didapatkan nilai F-score tertinggi yaitu ketika menggunakan metode ekstraksi fitur *character quadgram* dengan nilai F-score sebesar 61,338%, kemudian disusul dengan metode ekstraksi fitur *word unigram* dengan nilai F-score sebesar 60,777% dan terakhir metode ekstraksi fitur *word bigram* memberikan hasil yang paling rendah dengan nilai F-score yang didapatkan sebesar 53,698% perbedaanya mencapai sekitar 8% dari metode ekstraksi fitur *character quadgram*.

Untuk skenario dengan dua kombinasi metode ekstraksi fitur didapatkan nilai F-score tertinggi yaitu ketika menggunakan kombinasi metode ekstraksi fitur *word bigram* dan *character quadgram* dengan nilai F-score sebesar 64,636% disusul dengan penggunaan kombinasi metode ekstraksi fitur *word unigram* dan *word bigram* dengan nilai F-score sebesar 63,397% dan terakhir nilai F-score yang paling rendah untuk skenario dua kombinasi metode ekstraksi fitur didapatkan ketika mengkombinasikan metode ekstraksi fitur *word unigram* dan *character quadgram* dengan nilai F-score sebesar 62,096%. Sedangkan hasil terbaik untuk setiap skenario pengujian didapatkan ketika menggunakan ketiga metode ekstraksi fitur *word unigram*, *word bigram* dan *character quadgram* dengan nilai F-score yang didapatkan sebesar 64,957%.

Berdasarkan hasil evaluasi tersebut didapatkan bahwa semakin banyak fitur yang digunakan maka semakin baik nilai F-score yang didapatkan terhadap dataset ujaran kebencian dan tingkatan ancaman ini, hal ini terlihat dari hasil yang didapatkan ketika hanya menggunakan satu buah metode ekstraksi fitur nilai F-score terbaik hanya sebesar 61,338% sedangkan ketika beberapa metode ekstraksi fitur dikombinasikan nilai F-score terbaik naik menjadi sebesar 64,957%, bahkan hasil F-score terendah dari hasil kombinasi metode ekstraksi fitur lebih baik dibanding dengan penggunaan satu buah metode ekstraksi fitur saja dengan nilai F-score yang didapatkan sebesar 62,096%.

Meskipun sudah dilakukan beberapa skenario penelitian berdasarkan penggunaan metode ekstraksi fitur n-gram, hasil evaluasi F-score yang didapatkan hanya berkisar di 53% sampai 64%, untuk itu proses analisis dilanjutkan terhadap data uji yang disalah klasifikasikan, terutama untuk kelas 2, 3 dan 4 karena mempunyai sampel data latih yang paling sedikit, Tabel 4 menampilkan beberapa teks yang disalah klasifikasikan untuk skenario kombinasi metode ekstraksi fitur n-gram *word unigram*, *word bigram* dan *character quadgram*.

Tabel 4. Tabel Kesalahan Klasifikasi Hasil Uji

Teks yang sudah di praproses	Label Sebenarnya	Label Prediksi
Kelas 4 yang disalah Klasifikasikan Masuk Kedalam Kelas 2		
lengser jokowi dodo lebih tepat aman tentram negeri	4	2
bukan aksi turun jokowi biar selamat indonesia pak malu calon presiden	4	2
Kelas 3 yang disalah Klasifikasikan Masuk Kedalam Kelas 2		
semua fitnah bubar komisi berantas korupsi kalau bubar komisi berantas korupsi rakyat demo seluruh indonesia bubar dewan wakil rakyat makan gaji buta	3	2
komisi berantas korupsi butuh semua laku baik dewan wakil rakyat bubar	3	2
Kelas 4 yang disalah Klasifikasikan Masuk Kedalam Kelas 3		
sombong situ kampret cebong rendah hati rajin aji	4	3
perut buncit bukan banyak tidur cebong mana erti	4	3
Kegagalan Dalam Menormalisasikan Teks		
rohingya muslim rescued at sea off sumatra indonesia ekspat	1	2
berengsek lagu mantaaapppp	1	2
mantappppppp	1	2

Setelah menganalisis lebih dalam terhadap hasil uji klasifikasi, teks pada kelas 4 yang memiliki kemunculan kata “jokowi” cenderung sering disalah klasifikasikan masuk sebagai kelas 2, kemudian teks pada kelas 3 yang memiliki kemunculan kata “komisi”, “berantas” dan “korupsi” cenderung juga disalah klasifikasikan sebagai kelas 2. Untuk teks kelas 4 yang disalah klasifikasikan masuk kedalam kelas 3 rata-rata memiliki kemunculan kata “cebong”, “dungu” dan “rezim”.

Dari hasil analisis ini juga ditemukan beberapa teks yang menggunakan bahasa Inggris dari dataset yang mayoritas berbahasa Indonesia, hal ini menyebabkan sistem tidak bisa menghitung nilai probabilitas kelas secara akurat karena sampel yang digunakan untuk teks berbahasa Inggris sangatlah sedikit. Kemudian ditemukan beberapa kata yang gagal

dinormalisasikan karenya begitu banyaknya pola penulisan dari kata tersebut, contohnya ditemukan kata "mantap" akan tetapi terdapat berbagai macam pola penulisan seperti "mantaaapppp" dan "mantapppppppp" sehingga sistem tidak dapat menormalisasikan kata tersebut.

Selain itu nilai hasil evaluasi F-score juga dipengaruhi dari jumlah dataset yang digunakan, dataset yang tidak seimbang dapat mempengaruhi hasil klasifikasi. Dalam penelitian [11] yang membahas mengenai deteksi ujaran kebencian dibuktikan bahwa hasil klasifikasi yang dilakukan pada dataset yang tidak seimbang memiliki nilai hasil evaluasi yang lebih rendah dibanding dengan dataset yang seimbang untuk setiap metode klasifikasi yang digunakan pada penelitian tersebut.

4. KESIMPULAN

Dalam penelitian ini klasifikasi ujaran kebencian dan tingkat ancamannya dilakukan pada dataset *multi-label* Twitter berbahasa Indonesia dengan menggunakan metode klasifikasi Naïve Bayes, banyaknya label yang digunakan adalah sebanyak 4 label yaitu label '*HS*' untuk kategori jenis ujaran kebencian dan label '*HS_Weak*', '*HS_Moderate*' dan '*HS_Strong*' untuk kategori tingkatan ancaman ujaran kebencian. Selain dengan menggunakan metode klasifikasi Naïve Bayes tahapan klasifikasi juga digabungkan dengan beberapa metode tambahan lainnya seperti metode Label Power-set untuk mentransformasikan data dari *multi-label* kedalam bentuk *single-label multi-class*, ekstraksi fitur n-gram dan pembobotan kata TF-IDF. Hasil terbaik yang didapatkan yaitu sebesar 64,957% ketika menggunakan kombinasi metode ekstraksi fitur *word unigram*, *word bigram* dan *character quadgram* berdasarkan evaluasi dengan menggunakan perhitungan F-score.

Hasil terbaik selanjutnya yaitu ketika menggunakan kombinasi metode ekstraksi fitur *word bigram* dan *character quadgram* dengan nilai F-score yang didapatkan sebesar 64,636% untuk hasil terburuk didapatkan ketika hanya menggunakan *word bigram* dengan nilai F-score yang didapatkan sebesar 53,698%. Penggunaan kombinasi metode ekstraksi fitur memiliki nilai F-score yang lebih baik dibandingkan dengan hanya menggunakan satu buah metode ekstraksi fitur saja..

DAFTAR PUSTAKA

[1] A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella, "Thirty years of research into hate speech: topics of interest and their evolution," *Scientometrics*, vol. 126, no. 1, pp. 157–179, 2021, doi: 10.1007/s11192-020-03737-6.

[2] N. Chetty and S. Alathur, "Hate Speech Review in the Context of Online Social Networks," *Aggress. Violent Behav.*, vol. 40, pp. 108–118, 2018, doi: 10.1016/j.avb.2018.05.003.

[3] 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.

[4] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2017, vol. 2018-Janua, no. October, pp. 233–238, doi: 10.1109/ICACSIS.2017.8355039.

[5] M. Hakiem, M. A. Fauzi, and Indriati, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2443–2451, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4682.

[6] O. Oriola and E. Kotze, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.

[7] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, no. c, pp. 13825–13835, 2018, doi: 10.1109/ACCESS.2018.2806394.

[8] M. O. Ibrohim and I. Budi, "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media," *Procedia Comput. Sci.*, vol. 135, pp. 222–229, 2018, doi: 10.1016/j.procs.2018.08.169.

[9] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," *CEUR Workshop Proc.*, vol. 1816, no. January, pp. 86–95, 2017.

[10] 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.

[11] M. A. Fauzi and A. Yuniarti, "Ensemble Method for Indonesian Twitter Hate Speech Detection," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 11, no. 1, p. 294, Jul. 2018, doi: 10.11591/ijeecs.v11.i1.pp294-299.

[12] F. Herrera, F. Charte, A. J. Rivera, and M. J. Del Jesus, *Multilabel Classification: Problem Analysis, Metrics and Techniques*. 2016.

[13] S. Khomsah and A. S. Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam

Bahasa Indonesia," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 648–654, 2021, doi: 10.13140/RG.2.2.32319.74403.

[14] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[15] X. Tian and W. Tong, "An improvement to TF: Term distribution based term weight algorithm," *NSWCTC 2010 - 2nd Int. Conf. Networks Secur. Wirel. Commun. Trust. Comput.*, vol. 1, no. March 2011, pp. 252–255, 2011, doi: 10.1109/NSWCTC.2010.66.

[16] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[17] A. Rahman and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Ilm. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.