Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 3, Maret 2019, hlm. 2443-2451 http://j-ptiik.ub.ac.id

Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur *Information Gain*

Muhammad Hakiem[1], Mochammad Ali Fauzi[2], Indriati[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]muhammadhakiem@gmail.com, [2]moch.ali.fauzi@ub.ac.id, [3]indriati.tif@ub.ac.id

Abstrak

Ujaran kebencian atau *hate speech* adalah salah satu topik yang sering dibahas di bidang teknologi informasi. Ujaran kebencian banyak digunakan oleh orang-orang yang tidak suka atau benci terhadap seseorang maupun suatu kelompok. Orang menyatakan sebuah ujaran kebencian biasanya dilakukan dengan cara menulisnya di sosial media. Salah satu sosial media yang sering digunakan seseorang untuk menyebarkan ujaran kebencian adalah Twitter. Dibutuhkan klasifikasi ujaran kebencian agar dapat mengurangi penyebaran ujaran kebencian. Metode yang digunakan dalam penelitian ini adalah Naïve Bayes berbasis *N-gram* dan seleksi fitur Information Gain. Fitur *n-gram* yang digunakan pada penelitian ini adalah fitur *Unigram, Bigram,* dan kombinasi *unigram-bigram*. Data yang digunakan pada penelitian ini berjumlah 250 data berlabel ujaran kebencian dan 250 data berlabel bukan ujaran kebencian dengan perbandingan 80% untuk data latih dan 20% untuk data uji. Hasil akurasi terbaik yang didapat pada penelitian ini adalah dengan menggunakan fitur *Unigram* dan tanpa menggunakan seleksi fitur Information Gain. Hasil akurasi terbaik yang didapat adalah 84%, nilai *precision* 92%, nilai *recall* 79,31%, dan nilai *f-measure* 85,18%. Berdasarkan hasil yang didapat tersebut dapat diambil kesimpulan bahwa untuk melakukan klasifikasi ujaran kebencian pada Twitter menggunakan Naïve Bayes mendapat hasil paling bagus dengan fitur *Unigram* dan tanpa melakukan seleksi fitur Information Gain.

Kata kunci: ujaran kebencian, Twitter, *naïve bayes, n-gram, information gain*

Abstract

Hate speech is one of the topics that often discussed in information technology. Hate speech has been usually used by the people that don't like or hate with someone or a group. People stated their hate speech with post it in social media. One of the most used social media to spread the hate speech is Twitter. Hate speech identification is needed to decrease the spread of hate speech. The method used in this research is Naïve Bayes based on N-gram and feature selection Information Gain. N-gram features that used in this research are Unigram, Bigram, and combination unigram-bigram. 250 data are used in this research with hate speech label and 250 data with non hate speech label and have 80% proportion for data training and 20% for data testing. The best accuracy results in this research come from Unigram feature and without feature selection Information Gain. The best accuracy result is 84%, precision value 92%, recall value 79,31%, and f-measure value 85,18%. Based on the results obtained it can be concluded that to classify hate speech in Twitter using Naïve Bayes has the best result with Unigram feature and without using feature selection Information Gain.

Keywords: *hate speech*, Twitter, *naïve bayes, n-gram, information gain*

1. PENDAHULUAN

Ujaran kebencian adalah ucapan dan atau tulisan yang dibuat seseorang di muka umum untuk tujuan menyebar atau menyulut kebencian sebuah kelompok terhadap kelompok lain yang berbeda baik karena ras, agama keyakinan, gender, etnisitas, kecacatan, dan orientasi seksual (Greenawalt, 2016). Ujaran kebencian sangat banyak kita temui di kehidupan sehari-hari baik saat kita berkomunikasi secara langsung dengan orang-orang di sekitar kita maupun pada saat kita berseluncur di dunia maya yaitu di media sosial lebih utamanya. Kebebasan berpendapat yang dimiliki setiap orang malah seringkali disalahgunakan untuk menjatuhkan

ataupun mencemarkan nama baik orang lain.

Ujaran kebencian sangat banyak ditemui di media sosial yang dapat diakses oleh hampir setiap orang yang ada di Indonesia. Salah satu media sosial yang paling banyak digunakan untuk menyebarkan ujaran kebencian adalah Twitter. Jenis-jenis ujaran kebencian yang ada pada Twitter tersebut ada berbagai macam dan didasari oleh motif-motif tertentu seperti motif suku, agama, ras, dan antar golongan (SARA). Ujaran kebencian tersebut dapat menyebabkan terjadinya perpecahan antargolongan dan memecah belah persatuan bangsa Indonesia sehingga harus dapat ditangani dengan cepat sebelum berkembang menjadi sangat besar dan luas. Oleh karena itu klasifikasi ujaran kebencian diperlukan untuk mengurangi tersebarnya *tweet-tweet* yang berisikan ujaran kebencian tersebut. Klasifikasi ujaran kebencian yang dilakukan secara manual akan memakan waktu dan juga tenaga yang banyak, maka dari itu klasifikasi ujaran kebencian secara otomatis diperlukan untuk mempermudah proses klasifikasi tersebut sehingga kedepannya dapat digunakan oleh pihak yang membutuhkan seperti misalnya Kepolisian Negara Republik Indonesia (POLRI). Dengan banyaknya tweet yang berisikan ujaran kebencian maka dalam penelitian ini *tweet-tweet* tersebut akan digunakan sebagai data yang nantinya akan diolah sedemikian rupa sehingga dapat menghasilkan tujuan dari penelitian ini.

Permasalahan ujaran kebencian telah diangkat menjadi topik beberapa penelitian yang telah dilakukan sebelumnya, salah satunya yaitu deteksi ujaran kebencian di Twitter dalam Bahasa Indonesia (Fauzi, et al., 2018). Penelitian tersebut dilakukan untuk membandingkan beberapa metode algoritme yang ada dengan jumlah data yang seimbang dan yang tidak seimbang untuk mendeteksi ujaran kebencian. Metode yang digunakan adalah K–Nearest Neighbors (KNN), Naïve Bayes, Maximum Entropy, Random Forest, dan Support Vector Machine. Hasil dari penelitian tersebut bahwa metode Naïve Bayes menghasilkan nilai rata-rata F-measure yang paling tinggi dibandingkan metode lain yaitu 78,3% untuk data yang tidak seimbang dan 83,2% untuk data yang seimbang.

Penelitian lainnya yang dilakukan oleh Alfina (2017) mengenai ujaran kebencian di Twitter dalam Bahasa Indonesia yang berhubungan dengan Pemilihan Kepala Daerah (Pilkada) DKI Jakarta 2017 didapatkan hasil bahwa metode N-gram dapat menghasilkan nilai F-measure yang paling tinggi dibandingkan dengan metode lainnya seperti Char N-gram, Word Unigram, Word Bigram, Char Trigram, Char Quadragram dengan nilai 93,5%. Metode *N-gram* adalah gabungan dari *word unigram* dengan *word bigram* sehingga fitur yang didapat dari metode *word unigram* dan *word bigram* akan digabung menjadi satu. Fitur-fitur yang telah didapat tersebut akan digunakan dalam proses klasifikasi teks pada penelitian ini.

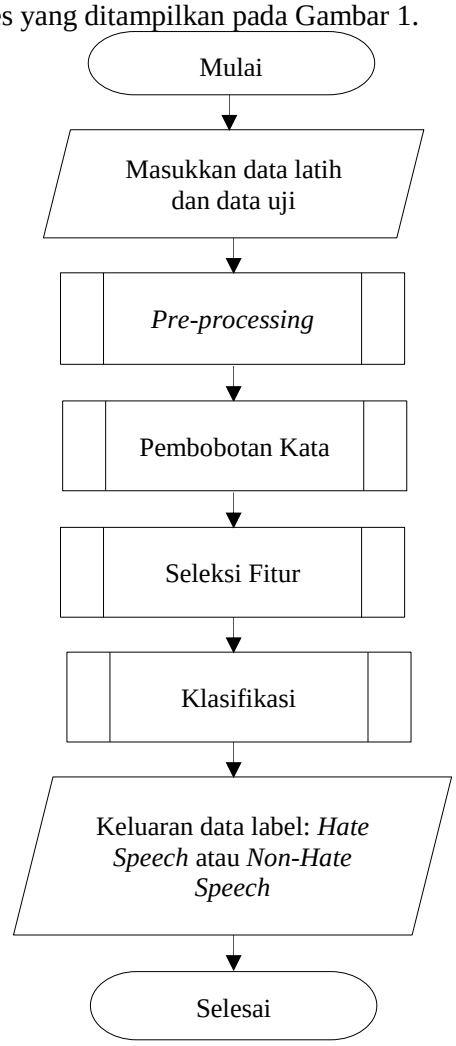
Data yang digunakan dalam penelitian ini didapat dari setiap tweet yang berisikan ujaran kebencian dalam Bahasa Indonesia sehingga dapat dilihat bahwa tata bahasa yang digunakan seseorang sangat beragam dalam menuliskan tweet tersebut, keberagaman itu membuat fitur yang dihasilkan melalui metode N-gram akan menjadi sangat banyak. Pengurangan dimensi fitur perlu dilakukan agar mendapatkan fitur-fitur yang relevan sehingga proses klasifikasi dapat dijalankan dengan lebih cepat dan mendapatkan nilai akurasi yang lebih tinggi. Cara untuk mengurangi dimensi fitur adalah salah satunya dengan melakukan teknik seleksi fitur. Pada penelitian Aini (2018) mengenai seleksi fitur Information Gain untuk klasifikasi penyakit jantung, didapatkan hasil bahwa dengan menggunakan Information Gain dapat meningkatkan nilai akurasi dari klasifikasi tersebut yaitu 84,62% tanpa menggunakan Information Gain menjadi 92,31% dengan menggunakan Information Gain. Dari hasil penelitian tersebut dapat dikatakan seleksi fitur Information Gain dapat digunakan untuk menyeleksi fitur-fitur yang akan digunakan untuk mendeteksi ujaran kebencian dengan metode Naïve Bayes berbasis N-gram dengan baik.

Berdasarkan dari penjelasan diatas bahwa pada penelitian ini diberi judul "Klasifikasi Ujaran Kebencian Pada Twitter Menggunakan Metode Naïve Bayes berbasis N-gram Dengan Seleksi Fitur Information Gain". Diharapkan metode Naïve Bayes berbasis N-gram dengan seleksi fitur Information Gain pada penelitian ini dapat mengklasifikasi ujaran kebencian dengan tepat dan sesuai. Sehingga maksud utama dari penelitian ini dapat terpenuhi yaitu dapat mengetahui pengaruh fitur N-gram dan seleksi fitur Information Gain terhadap metode Naive Bayes untuk klasifikasi ujaran kebencian pada Twitter.

2. METODE PENELITIAN

2.1 Deskripsi Umum Sistem

Sistem yang akan dibuat pada penelitian ini dijelaskan dengan melihat tahapan-tahapan alur proses yang ditampilkan pada Gambar 1.



Gambar 1. Diagram Alir Sistem

Alur dari tahapan-tahapan yang dilakukan dalam penyelesaian masalah pada penelitian ini digambarkan oleh diagram alir sistem pada Gambar 1. Tahapan pertama yang dilakukan adalah memasukkan data latih dan data uji untuk digunakan pada proses klasifikasi pada sistem ini. Tahapan selanjutnya adalah melakukan *pre-processing* terhadap data-data yang telah dimasukkan tersebut agar dapat diproses pada tahapan selanjutnya yaitu pembobotan kata. Pembobotan kata dilakukan untuk mengetahui berapa jumlah kemunculan kata yang telah didapat dari proses *pre-processing* pada dokumen yang dimiliki. Kemudian melakukan seleksi fitur pada fitur-fitur yang telah didapat dari proses sebelumnya agar mendapatkan fitur yang benar-benar berpengaruh pada proses klasifikasi yang akan dilakukan. Proses klasifikasi merupakan tahapan akhir yang dilakukan yaitu untuk mendapatkan hasil klasifikasi kelas dari data uji yang dimasukkan sehingga menghasilkan keluaran berupa label kelas *hate speech* ataupun *non-hate speech* dari data tersebut. Semua tahapan selesai dilakukan setelah mendapatkan keluaran berupa label kelas tersebut.

Data yang digunakan pada penelitian merupakan data sekunder yang didapat dari penelitian yang telah dilakukan sebelumnya oleh Alfina (2017). Data tersebut adalah 713 data sekunder berupa *tweets* dalam Bahasa Indonesia yang telah diberi label kelas yang terdiri dari HS (*Hate Speech*) dan Non_HS (*Non Hate Speech*). Data didapat menggunakan *Twitter Streaming* API yang dikumpulkan dari bulan Februari 2017 hingga bulan April 2017. Data tersebut berisi macam-macam bentuk *tweets* meliputi *emoticon*, *link*, *mention*, dan karakter lain. Dari data tersebut diambil sejumlah 500 data yang berisi 250 data berlabel HS (*Hate Speech*) dan 250 data berlabel Non_HS (*Non Hate Speech*). Data latih yang akan dipakai berjumlah 200 data berlabel HS (*Hate Speech*) dan 200 data berlabel Non_HS (*Non Hate Speech*) yang diambil dari 250 data yang ada untuk setiap labelnya. Data uji yang akan dipakai berjumlah 50 data berlabel HS (*Hate Speech*) dan 50 data berlabel Non_HS (*Non Hate Speech*) yang diambil dari sisa data yang tidak dipakai untuk data latih untuk setiap labelnya.

Untuk membangun sistem klasifikasi ujaran kebencian dengan menggunakan metode Naïve Bayes berbasis *n-gram* dengan seleksi fitur Information Gain ini, diperlukan teori pendukung untuk memperkuat dasar teori dari penelitian ini.

2.2 *Pre-processing* teks

Pre-processing teks merupakan tahapan awal sebelum mengolah data lebih lanjut dan kemudian digunakan dalam proses pembuatan sistem. Tahapan dalam *pre-processing* teks ada berbagai macam tahap dan tahapan *pre-processing* yang akan digunakan nantinya akan menyesuaikan dengan proses yang akan dilalui. *Pre-processing* teks dilakukan untuk mengubah data menjadi lebih mudah dan sesuai untuk diproses oleh sistem, dan juga untuk membuang data-data yang tidak diperlukan oleh sistem tersebut. *Pre-processing* teks sangat penting dalam melakukan analisis mengenai teks,

terutama untuk media sosial yang sebagian besar berisi kata-kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar (Mujilahwati, 2016). Tahapan pre-processing teks yang digunakan adalah *Cleaning*, *Case Folding*, Tokenisasi, *Stopword Removal*, dan *Stemming*.

2.3 Fitur *N-gram*

Fitur *n-gram* merupakan fitur yang dihasilkan dari proses ekstraksi fitur pada saat *pre-processing* teks. Fitur *n-gram* adalah fitur yang berisi kumpulan kata-kata yang akan menjadi acuan dalam proses-proses selanjutnya. Proses klasifikasi teks memerlukan kata-kata yang unik untuk menentukan kelas dari sebuah kalimat data uji yang ada. Setiap kata yang disimpan dalam fitur *n-gram* memiliki jumlah kemunculan yang berbeda-beda untuk setiap dokumen yang ada. Fitur *n-gram* dapat berupa beberapa jenis kumpulan kata seperti misalnya untuk kata tunggal yaitu disebut *unigram*, kata ganda disebut *bigram*, dan lain-lain. Kumpulan kata pada fitur *n-gram* tergantung dari proses ektraksi fitur yang dilakukan sebelumnya. Penggunaan fitur *n-gram* dapat membantu mengurangi kebutuhan analisis morfem pada setiap kalimat yang ada (Bozkir, et al., 2017).

2.4 Naïve Bayes

Naïve Bayes merupakan algoritme populer dan baik untuk data berdimensi tinggi seperti teks (Pratiwi, 2017). Algoritme ini didasarkan pada *teorema Bayes* yang menggunakan probabilitas bersyarat untuk mengklasifikasikan data ke dalam kelas yang ditentukan sebelumnya (Kumari, 2014). Dikatakan *Naïve* karena algoritme ini mengasumsikan pada setiap nilai atribut tidak ada hubungannya satu sama lain atau *independence*. Persamaan 2.1 adalah persamaan *Naïve Bayes*.

$$P(c|d) = \frac{\mathrm{P}(d|c)*P(C)}{P(d)} \quad (2.1)$$

Di mana *d* adalah kata, *c* adalah kelas, *P(d|c)* adalah peluang kata di kelas *c*, *P(c)* adalah peluang dari kelas *c* dan *P(d)* adalah peluang dari kata *d*.

Terdapat permasalahan yang sering terjadi pada algoritme *Naïve Bayes* di mana jika kemunculan kata tidak pernah terjadi maka akan membuat nilai *posterior* menjadi nol atau masalah ini biasa disebut *zero frequency problem* (Kikuchi et al., 2015). Untuk mengatasi ini, cara yang biasanya digunakan adalah *Laplace Smoothing*. *Laplace Smoothing* digunakan karena metode ini merupakan metode yang popular untuk mengatasi *zero frequency problem* (Kikuchi et al, 2015). Metode ini bekerja dengan menganggap kata yang tidak pernah muncul pernah muncul sekali sehingga di dalam persamaan terdapat penambahan 1 frekuensi pada frekuensi kemunculan kata. Persamaan *Laplace Smoothing* pada Persamaan 2.2.

$$P(word_i\,|class_j) = \frac{count(word_i,class_j)+1}{count(class_j)+|v|} \quad (2.2)$$

Keterangan:
$P(word_i\,|class_j)$ = nilai peluang yang diasumsikan bahwa semua kata yang muncul ditambah nilai 1 pada frekuensi kemunculan kata
$count(word_i\,|class_j)$ = jumlah kata i pada kelas j
$count(class_j)$ = jumlah seluruh kata pada kategori j
|v| = jumlah kata unik yang ada pada seluruh dokumen.

Fitur yang memiliki jumlah frekuensi kemunculan tinggi pada data latih suatu kategori akan menjadi nilai probabilitas yang tinggi pada setiap kemunculan fitur tersebut di data uji (Afriza, 2018). Hal tersebut menunjukkan bahwa fitur tersebut akan berpengaruh sekali untuk hasil klasifikasinya. Fitur-fitur seperti itulah yang akan menjadi kata kunci penting untuk menentukan hasil klasifikasi yang tepat pada setiap dokumennya. Sedangkan untuk fitur yang memiliki jumlah frekuensi kemunculan rendah pada data latih suatu kategori akan memiliki nilai probabilitas yang rendah juga sehingga fitur tersebut tidak memiliki pengaruh besar terhadap hasil klasifikasi yang ada.

2.5 Information Gain

Information Gain adalah jumlah informasi yang disediakan oleh fitur teks untuk kategori teks. Information Gain dihitung dengan berapa banyak term yang digunakan untuk klasifikasi informasi, untuk mengukur pentingannya lexical teks untuk diklasifikasi (Lei, 2002). Rumus Information Gain ditampilkan pada Persamaan 2.3.

$$IG(t) = -\sum_{i=1}^{|C|} P(Ci)logP(Ci) + P(t)\sum_{i=1}^{|C|} P(Ci|t)logP(Ci|t) + P(\bar{t})\sum_{i=1}^{|C|} P(Ci|\bar{t})logP(Ci|\bar{t}) \quad (2.3)$$

Keterangan:
P(Ci) = Peluang kategori i di seluruh dokumen
P(t) = Peluang *term* t muncul di dokumen
P(Ci|t) = Peluang *term* t muncul pada kategori i di dokumen
P($\bar{t}$) = Peluang *term* t tidak muncul di dokumen
P(Ci|$\bar{t}$) = Peluang *term* t tidak muncul pada kategori i di dokumen.

3. PENGUJIAN

3.1 Pengujian Variasi Fitur N-gram

Pada pengujian ini dilakukan dengan menggunakan fitur-fitur yang berbeda untuk mengetahui fitur mana yang merupakan fitur terbaik untuk melakukan klasifikasi pada sistem. Fitur yang akan digunakan adalah fitur unigram, bigram, dan kombinasi. Pada tabel 1 berikut ini akan ditunjukkan hasil pengujian fitur-fitur tersebut dengan melihat nilai akurasi, precision, recall, dan f-measure yang dihasilkannya.

Tabel 1. Hasil Pengujian Variasi Fitur N-gram

Evaluasi	Variasi Fitur (Dalam %)		
	Unigram	*Bigram*	Kombinasi
Akurasi	84	74	80
Precision	92	57,99	88
Recall	79,31	85,29	75,86
F-Measure	85,18	69,05	81,48

3.2 Pengujian Variasi Seleksi Fitur Information Gain

Pada pengujian ini dilakukan dengan menggunakan nilai *threshold* seleksi fitur Information Gain yang berbeda untuk mengetahui berapa nilai *threshold* yang terbaik untuk melakukan klasifikasi pada sistem. Pengujian juga dilakukan dengan tanpa menggunakan seleksi fitur Information Gain untuk mengetahui pengaruh seleksi fitur Information Gain terhadap hasil klasifikasi pada sistem. Nilai *threshold* yang akan digunakan pada pengujian ini adalah 20%, 40%,60%, 80%, dan 100% untuk tanpa menggunakan seleksi fitur Information Gain. Pada Tabel 2 berikut ini akan ditunjukkan hasil pengujian fitur-fitur tersebut pada Fitur *Unigram* dengan melihat nilai akurasi, *precision*, *recall*, dan *f-measure* yang dihasilkannya.

Tabel 2. Hasil Pengujian Variasi Seleksi Fitur Information Gain Pada FItur *Unigram*

Evaluasi	Nilai *Threshold* (Dalam %)				
	20%	40%	60%	80%	100%
Akurasi	78	75	79	78	84
Precision	84	90	90	90	92
Recall	75	69,23	73,77	72,58	79,31
F-Measure	79,24	78,26	81,08	80,36	85,18

Pada Tabel 2 ditunjukkan hasil akurasi, *precision*, *recall*, dan *f-measure* dari setiap pengujian nilai *threshold* seleksi fitur Information Gain yang dilakukan pada fitur *Unigram*. Hasil akurasi yang paling tinggi diraih oleh nilai *threshold* 100% dengan nilai 84%. Hal tersebut menunjukkan bahwa tanpa seleksi fitur Information Gain dapat memiliki nilai akurasi yang lebih tinggi dibandingkan dengan menggunakan seleksi fitur Information Gain. Hasil pengujian seleksi fitur Information Gain pada fitur *Bigram* ditampilkan pada Tabel 3 dengan melihat nilai akurasi, *prescision*, *recall*, dan *f-measure* yang dihasilkannya.

Tabel 3. Hasil Pengujian Variasi Seleksi Fitur Information Gain Pada Fitur *Bigram*

Evaluasi	Nilai *Threshold* (Dalam %)				
	20%	40%	60%	80%	100%
Akurasi	70	69	72	71	74
Precision	44	44	52	54	57,99
Recall	91,67	88	86,67	81,82	85,29
F-Measure	59,46	58,67	65	65,06	69,05

Pada Tabel 3 ditunjukkan hasil akurasi, *precision*, *recall*, dan *f-measure* dari setiap pengujian nilai *threshold* seleksi fitur Information Gain yang dilakukan pada fitur Bigram. Hasil akurasi yang paling tinggi diraih oleh nilai *threshold* 100% dengan nilai 74%. Hal tersebut menunjukkan bahwa tanpa seleksi fitur Information Gain dapat memiliki nilai akurasi

yang lebih tinggi dibandingkan dengan menggunakan seleksi fitur Information Gain. Hasil pengujian seleksi fitur Information Gain pada fitur Kombinasi ditampilkan pada Tabel 4 dengan melihat nilai akurasi, *prescision*, *recall*, dan *f-measure* yang dihasilkannya.

Tabel 4. Hasil Pengujian Variasi Seleksi Fitur Information Gain Pada Fitur Kombinasi

Evaluasi	Nilai *Threshold* (Dalam %)				
	20%	40%	60%	80%	100%
Akurasi	79	70	78	82	80
Precision	88	90	92	88	88
Recall	74,58	64,29	71,87	78,57	75,86
F-Measure	80,73	75	80,70	83,02	81,48

Pada Tabel 4 ditunjukkan hasil akurasi, *precision*, *recall*, dan *f-measure* dari setiap pengujian nilai *threshold* seleksi fitur Information Gain yang dilakukan pada fitur kombinasi *unigram-bigram*. Hasil akurasi yang paling tinggi diraih oleh nilai *threshold* 80% dengan nilai 82%. Hal tersebut menunjukkan bahwa seleksi fitur Information Gain dapat memiliki nilai akurasi yang lebih tinggi dibandingkan dengan tanpa menggunakan seleksi fitur Information Gain.

4. ANALISIS HASIL PENGUJIAN

Pengujian yang telah dilakukan dengan berbagai parameter telah menghasilkan nilai akurasi, precision, recall, dan f-measure yang berbeda-beda. Dari hasil pengujian tersebut dapat dianalisis sehingga mendapatkan kesimpulan dari hasil penelitian ini.

4.1 Analisis Metode Naïve Bayes

Metode yang digunakan pada penelitian ini adalah metode klasifikasi Naïve Bayes yang merupakan salah metode terbaik untuk melakukan klasifikasi teks. Pada penelitian ini metode Naïve Bayes dapat melakukan klasifikasi dengan baik yang dapat dilihat dari hasil akurasi terbaik pada fitur Unigram yaitu 84%, pada fitur Bigram yaitu 74%, dan pada fitur Kombinasi unigram-bigram yaitu 82%. Metode klasifikasi Naïve Bayes dapat memiliki nilai akurasi yang tinggi karena pada metode Naïve Bayes melakukan perhitungan probabilitas kemunculan setiap kata yang ada pada setiap kelasnya untuk menentukan hasil klasifikasinya. Pada penelitian ini telah dilakukan proses training dengan 400 data latih sehingga telah memiliki nilai probabilitas untuk setiap katanya. Setiap kelas pasti memiliki sejumlah kata-kata yang sering muncul pada kelas tersebut sehingga menjadi kata-kata pembeda dari kelas lainnya. Apabila kata tersebut muncul pada data uji maka nilai probabilitas yang dihasilkan pada data uji tersebut akan lebih besar pada kelas yang kata tersebut sering muncul di data latih. Pada akhirnya metode klasifikasi Naïve Bayes dapat melakukan klasifikasi dengan baik yang disebabkan oleh kemunculan kata-kata yang menjadi pembeda pada setiap kelasnya.

4.2 Analisis Pengujian Variasi Fitur *N-gram*

Pengujian dilakukan terhadap variasi fitur n-gram dan variasi seleksi fitur Information Gain. Pengujian yang dilakukan pada sistem ini telah menghasilkan bahwa fitur Unigram dengan tanpa melakukan seleksi fitur Information Gain memiliki nilai akurasi yang paling tinggi yaitu 84%. Hal tersebut dikarenakan untuk melakukan klasifikasi ujaran kebencian di Twitter diperlukan fitur-fitur yang dapat mencakup setiap kata kunci yang digunakan orang untuk melakukan tweet ujaran kebencian tersebut.

Pada fitur Unigram berisi berisi kata-kata tunggal yang dianggap tidak memiliki hubungan antara satu kata dengan kata yang lainnya sehingga suatu kata yang bersebelahan akan tidak berpengaruh dan dianggap sama saja apabila bersebelahan dengan kata yang lainnya. Dalam fitur Unigram dapat mencakup kata-kata tunggal yang sering digunakan orang dalam membuat tweet ujaran kebencian. Setiap orang yang membuat tweet ujaran kebencian memiliki gaya bahasa dan cara penulisan yang berbeda-beda sehingga akan memiliki kata tunggal yang banyak dan berbeda pada setiap tweet ujaran kebencian yang ada. Hal tersebut mengakibatkan untuk fitur Bigram dan fitur kombinasi unigram-bigram akan memiliki nilai akurasi yang lebih rendah yaitu 74% untuk fitur bigram dan 82% untuk fitur kombinasi unigram-bigram. Nilai akurasi yang lebih rendah tersebut karena pada fitur Bigram dan kombinasi unigram-bigram berisi gabungan dari kata-kata tunggal yang bersebelahan sehingga suatu kata yang bersebelahan akan berpengaruh pada setiap kemunculannya di dalam dokumen. Pada fitur Bigram hanya berisi gabungan dari kata-kata tunggal yang bersebelahan sehingga kemungkinan kata tunggal yang merupakan kata pembeda antara dua kelas pada data latih

tersebut muncul bersebelahan dengan kata-kata yang lain akan sedikit ditemukan pada dokumen uji. Gabungan kata-kata tunggal yang akan sering muncul pada data latih dan juga pada data uji adalah kata umum yang sering digunakan untuk membuat sebuah kalimat biasa pada umumnya sehingga kata tersebut akan memiliki kemunculan yang banyak pada kedua kelas yang ada. Hal itu menyebabkan metode Naïve Bayes akan susah melakukan klasifikasi kelas dari data karena akan kekurangan kata yang menjadi pembeda antara dua kelas yang ada dan hanya memiliki kata-kata umum yang muncul pada kedua kelas tersebut.

Pada fitur kombinasi unigram-bigram juga akan terjadi hal yang sama seperti pada fitur bigram karena fitur kombinasi unigram-bigram berisi fitur unigram dan bigram sehingga kekurangan yang dimiliki pada fitur bigram juga akan dimiliki pada fitur kombinasi unigram-bigram ini. Pada fitur kombinasi unigram-bigram ini juga akan mengalami kesalahan klasifikasi yang dikarenakan kata-kata dari fitur bigram lebih banyak terjadi kemunculan kata umum yang dimiliki oleh kedua kelas daripada kata yang menjadi pembeda dari kedua kelas tersebut. Setelah melihat hasil pengujian dan analisis tersebut dapat diambil kesimpulan bahwa untuk kasus klasifikasi ujaran kebencian ini fitur unigram dapat melakukan klasifikasi dengan lebih baik karena berisi kata tunggal yang akan menjadi pembeda untuk setiap kelasnya.

4.3 Analisis Pengujian Variasi Seleksi Fitur Information Gain

Pengujian yang dilakukan pada seleksi fitur Information Gain dilakukan pada setiap fitur yang ada yaitu fitur unigram, bigram, dan kombinasi unigram-bigram. Hasil yang ditunjukkan dari pengujian seleksi fitur pada setiap fitur yang diujikan memiliki hasil yang berbeda dan ditunjukkan dengan nilai akurasi, precision, recall, dan f-measure masing-masing pengujian yang dilakukan.

Pada pengujian seleksi fitur Information Gain untuk fitur Unigram dihasilkan nilai akurasi tertinggi yang didapat adalah pada nilai *threshold* 100% dengan nilai akurasi 84%. Hal tersebut menunjukkan bahwa fitur Unigram dapat melakukan klasifikasi dengan lebih baik tanpa seleksi fitur Information Gain. Itu disebabkan oleh fitur Unigram memerlukan semua fitur yang dihasilkan oleh proses ekstraksi fitur untuk melakukan proses klasifikasi. Pada fitur Unigram yang berisi kata-kata tunggal akan lebih baik apabila memiliki setiap fitur yang ada karena dari kata-kata tunggal tersebut menjadi pembeda antara kedua kelas walaupun hanya muncul dalam jumlah yang sangat sedikit. Apabila dilakukan seleksi fitur Information Gain dapat menyebabkan sebagian kata dengan kemunculan yang sangat sedikit menjadi tidak terseleksi dan mengurangi keakuratan dari proses klasifikasi tersebut. Oleh karena itu tanpa seleksi fitur adalah parameter yang terbaik untuk fitur Unigram.

Pada pengujian seleksi fitur Information Gain untuk fitur Bigram dihasilkan nilai akurasi tertinggi yang didapat adalah pada nilai *threshold* 100% dengan nilai akurasi 74%. Hal tersebut menunjukkan bahwa fitur Bigram dapat melakukan klasifikasi dengan lebih baik tanpa seleksi fitur Information Gain. Itu disebabkan oleh fitur Bigram memerlukan semua fitur yang dihasilkan oleh proses ekstraksi fitur untuk melakukan proses klasifikasi. Karena pada fitur Bigram yang berisi gabungan kata-kata tunggal yang bersebelahan tersebut akan memiliki jumlah kata pembeda antar kelas yang sangat sedikit. Sehingga apabila dilakukan seleksi fitur Information Gain dapat menyebabkan sebagian kata dengan kemunculan yang sangat sedikit namun merupakan kata pembeda antar kelas menjadi tidak terseleksi dan mengurangi keakuratan dari proses klasifikasi tersebut. Oleh karena itu tanpa seleksi fitur adalah parameter yang terbaik untuk fitur Bigram.

Pada pengujian seleksi fitur Information Gain untuk fitur kombinasi unigram-bigram dihasilkan nilai akurasi tertinggi yang didapat adalah pada nilai *threshold* 80% dengan nilai akurasi 82%. Hal tersebut menunjukkan bahwa fitur Bigram dapat melakukan klasifikasi lebih baik dengan menggunakan seleksi fitur Information Gain. Itu disebabkan oleh fitur kombinasi unigram-bigram tidak memerlukan semua fitur yang dihasilkan oleh proses ekstraksi fitur untuk melakukan proses klasifikasi. Karena pada fitur kombinasi unigram-bigram yang berisi gabungan fitur Unigram dan Bigram tersebut akan memilik jumlah kata yang sangat banyak. Dari jumlah kata yang sangat banyak tersebut juga banyak berisi kata-kata bukan merupakan pembeda antar kelas. Sehingga apabila dilakukan seleksi fitur Information Gain dapat menyebabkan sebagian kata yang bukan merupakan kata pembeda antar kelas dan memiliki kemunculan yang sedikit akan tidak

terseleksi dan dapat meningkatkan keakuratan dari proses klasifikasi tersebut. Oleh karena itu dengan nilai *threshold* 80% pad seleksi fitur Information Gain adalah parameter yang terbaik untuk fitur kombinasi unigram-bigram.

Pengujian yang telah dilakukan tersebut menghasilkan kesimpulan bahwa fitur Unigram merupakan fitur yang terbaik untuk melakukan klasifikasi karena dapat mencakup banyak kata kunci dalam setiap tweet ujaran kebencian. Nilai *threshold* 100% pada seleksi fitur Information Gain merupakan nilai *threshold* yang terbaik untuk Unigram dan Bigram, dan nilai *threshold* 80% pada seleksi fitur Information Gain merupakan nilai terbaik untuk fitur kombinasi unigram-bigram. Hal ini membuktikan bahwa fitur n-gram dan seleksi fitur Information Gain yang digunakan dapat mempengaruhi hasil klasifikasi ujaran kebencian yang dilakukan oleh sistem ini.

5. KESIMPULAN

Hasil dari penelitian ini dapat diambil kesimpulan untuk klasifikasi ujaran kebencian pada Twitter menggunakan metode Naïve Bayes berbasis n-gram dengan seleksi fitur Information Gain. Kesimpulan tersebut didapat dari analisis hasil-hasil pengujian yang telah dilakukan pada penelitian ini.

Pada proses klasifikasi yang dilakukan di dalam penelitian ini digunakan fitur n-gram yaitu fitur unigram, bigram, dan kombinasi. Fitur-fitur tersebut diuji dan didapatkan hasil bahwa fitur *Unigram* memiliki nilai akurasi yang paling tinggi yaitu 84% dan diikuti dengan nilai precision, recall, dan f-measure yaitu 92%, 79,31%, dan 85,18%. Hasil pengujian tersebut menunjukkan bahwa fitur *Unigram* merupakan fitur yang paling cocok untuk melakukan klasifikasi ujaran kebencian pada Twitter. Fitur *unigram* berisi kata-kata tunggal yang dianggap tidak memiliki hubungan dengan kata disebelahnya sehingga lebih banyak memiliki kata-kata yang sering digunakan orang-orang dalam membuat sebuah tweet ujaran kebencian. Hal tersebut membuat fitur *Unigram* menjadi yang paling cocok untuk digunakan pada klasifikasi tersebut.

Pada penelitian ini dilakukan seleksi fitur Information Gain untuk mendapatkan fitur-fitur yang paling berpengaruh dan membuang fitur-fitur yang dianggap kurang berpengaruh untuk proses klasifikasi pada sistem ini. Pengujian pada seleksi fitur Information Gain dilakukan dengan nilai *threshold* 20%, 40%, 60%, dan 80%. Pengujian juga dilakukan dengan tanpa menggunakan seleksi fitur Information Gain yaitu dengan nilai *threshold* 100%. Hasil pengujian yang didapat adalah nilai *threshold* 100% pada seleksi fitur Information Gain merupakan nilai *threshold* yang terbaik untuk *Unigram* dan *Bigram*, dan nilai *threshold* 80% pada seleksi fitur Information Gain merupakan nilai terbaik untuk fitur kombinasi *unigram-bigram*. Hasil pengujian tersebut menunjukkan bahwa seleksi fitur Information Gain dapat menurunkan nilai akurasi dari klasifikasi pada sistem ini karena fitur *Unigram* dan *Bigram* memerlukan semua kata yang dihasilkan dari proses ekstraksi fitur. Namun juga dapat meningkatkan nilai akurasi dari klasifikasi pada sistem ini karena kata yang terlalu banyak dan tidak berpengaruh pada proses klasifikasi dapat mengurangi akurasi yang akan didapat.

DAFTAR PUSTAKA

Afriza, A., Adisantoso, J., 2018. Metode Klasifikasi Rocchio untuk Analisis Hoax.

Aini, S. H. A., Sari, Y. A., Arwan, A., 2018. Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes.

Alfina, I., Mulia, R., Fanany, M. I. & Ekanata, Y., 2017. Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study. 9th International Conference on Advanced Computer Science and Information Systems 2017 (ICACSIS).

Bozkir, A. S., Sezer, E. A., Aydos, M., 2017. Spam E-Mail Classification by Utilizing N-Gram Features of Hyperlink Texts.

Greenawalt, R. K., 2016. Legal Enforcement of Morality.

Fauzi, M. A., Yuniarti, A., 2018. Ensemble Method for Indonesian Twitter Hate Speech Detection. Indonesian Journal of Electrical Engineering and Computer Science

Kikuchi, Masato., Yoshida, Mitsui., Okabe, Masayuki., 2015, Confidence Interval of Probability Estimator of Laplace Smoothing.

Kumari, Anjana., 2014, Study on Naive Bayesian Classifier and its relation to Information Gain, International Journal on Recent and Innovation Trends in Computing and Communication Vol II, pp. 601-603.

Lei, S., 2012, A Feature Selection Method Based on Information Gain and Genetic Algorithm, International Conference on Computer Science and Electronics Engineering.

Mujilahwati, S., 2016, Pre-processing Text Mining Pada Data Twitter. Seminar Nasional Teknologi Informasi dan Komunikasi 2016 (SENTIKA 2016).

Pratiwi, I.Y.R., Asmara, R.A., Rahutomo, F., 2017. Study of Hoax News Detection Using Naive Bayes Classifier in Indonesia Language.