METODE MULTINOMIAL NAÏVE BAYES UNTUK KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA

(*Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia*)

Alif Sabrani*, I Gede Putu Wirarama Wedashwara W., Fitri Bimantoro
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Mataram
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: alifsabrani@gmail.com, [wirarama, bimo]@unram.ac.id

Abstract

Artikel online tentang gempa bumi dapat dikelompokan ke dalam kategori ekonomi, kesehatan, dan pariwisata. Pengelompokan artikel dalam jumlah besar dapat menguras waktu dan tenaga apabila dilakukan secara manual.Text classification dapat membantu proses klasifikasi artikel ini. Pada penelitian ini, dilakukan pengujian pada performa dari metode probabilistik multinomial Naïve Bayes dalam mengelompokan artikel online tentang gempa bumi di Indonesia. Pembobotan dilakukan dengan menggunakan teknik TF-IDF. Pengujian dilakukan dengan 2 jenis feature yaitu unigram dan bigram, serta penggabungan dari keduanya. Selain itu, pengujian juga dilakukan dengan menghilangkan stemming dan stopwords removal dari tahap preprocessing. F-measure tertinggi yang didapatkan adalah sebesar 95.20% yaitu pada skenario pengujian dengan menggabungkan feature unigram dan bigram serta melewati tahap stemming dan stopwords removal pada preprocessing.

Keywords: *Gempa, Artikel Online, Klasifikasi Teks, TF-IDF, Preprocessing, Multinomial Naïve Baye*

**Penulis Korespondensi*

1. PENDAHULUAN

Secara geologi, Indonesia berada di pertemuan tiga lempeng utama dunia, yaitu Eurasia, Indoaustralia dan Pasifik. Selain itu, Indonesia juga dikenal berada di Cincin Api Pasifik (*Ring of Fire*) yaitu daerah "tapal kuda" sepanjang 40.000 km yang sering mengalami gempa bumi dan letusan gunung berapi yang mengelilingi cekungan Samudra Pasifik. Sekitar 90% dari gempa bumi yang terjadi dan 81% dari gempa bumi terbesar terjadi di sepanjang Cincin Api ini. Menurut Dr. Daryono, kepala bidang informasi gempa bumi dan peringatan dini tsunami Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), kondisi ini menyebabkan gempa bumi sering terjadi di Indonesia [1].

Terdapat begitu banyak artikel *online* yang berhubungan dengan gempa bumi di Indonesia yang tersebar di berbagai *website*. Artikel *online* pasca gempa dapat berupa kondisi, dampak, maupun aktivitas yang dilakukan di lokasi terjadinya gempa dan dapat dikelompokan ke dalam kategori ekonomi, kesehatan, atau pariwisata. Artikel yang telah dikelompokan dapat membantu untuk memudahkan pembaca dalam mencari informasi.

Pengelompokan artikel dalam jumlah besar dapat menguras waktu dan tenaga apabila dilakukan secara manual. Untuk mempermudah proses pengelompokan ini, diperlukan suatu teknik yang tepat. Teknik yang digunakan untuk klasifikasi dokumen secara otomatis oleh komputer dikenal dengan istilah *text classification* atau klasifikasi teks [2]. Metode yang dapat digunakan dalam klasifikasi teks adalah metode probabilistik *Naïve Bayes*. Metode *Naïve Bayes* terbukti dapat memberikan hasil yang cukup memuaskan ketika digunakan untuk klasifikasi teks [3]. Salah satu model dari *Naïve Bayes* yang sering digunakan dalam klasifikasi teks adalah *multinomial Naïve Bayes* [4].

Pada Bagian 2, dibahas uraian-uraian pustaka terbaru yang berkaitan dengan topik artikel. Bagian 3 memuat penjelasan tentang tahapan proses penelitian dengan urutan logis untuk mendapatkan hasil penelitian sesuai dengan harapan. Bagian 4 berisi hasil dan pembahasan penelitian. Dan Bagian 5 berisi pernyataan atas temuan yang dihasilkan dari penelitian dan pernyataan jawaban atas masalah di ingin diselesaikan beserta rencana penelitian di masa mendatang.

2. TINJAUAN PUSTAKA

Telah dilakukan penelitian untuk mengklasifikasikan pengaduan dan pelaporan masyarakat menggunakan metode *multinomial Naïve Bayes* [2]. Metode *multinomial Naïve Bayes* dipilih karena dikenal memiliki

tingkat akurasi yang tinggi dengan perhitungan sederhana. *Dataset* yang digunakan pada penelitian ini adalah sebanyak 113 pelaporan. Data dikelompokkan menjadi 3 kelompok yaitu Informasi, Kamtibnas, dan Tindak Pidana. Penelitian yang dilakukan menghasilkan rata-rata akurasi yang tinggi, yaitu *recall* 93%, *precission* 90 %,dan *f-measure* 92%.

Dilakukan juga penelitian untuk menguji akurasi metode *multinomial Naïve Bayes* dalam mengelompokkan pesan di ruang percakapan maya dengan [5]. Pesan dikategorikan menjadi 5 kategori yaitu Dalam himpunan, Luar himpunan, Berita duka, Ulang tahun, dan Percakapan lainnya. Hasil yang didapatkan cukup baik dengan *F-measure* mencapai 90,57% untuk kategori Dalam himpunan.

Penelitian untuk membandingkan metode *multinomial Naïve Bayes* dan *k-nearest neighbor* dalam pengelompokan jurnal juga telah dilakukan [6]. Metode *multinomial Naïve Bayes* dipilih karena hanya memerlukan sejumlah kecil data latih untuk menentukan parameter *mean* dan *varians* dari variabel yang diperlukan untuk klasifikasi. Terdapat 4 kategori jurnal yaitu Pendidikan Ekonomi, Pendidikan Bisnis dan Manajemen, Akuntansi Aktual, dan Ekonomi Bisnis. Jumlah data yang digunakan adalah 40 jurnal yang dibagi masing-masing 10 jurnal per kategori. Hasil yang didapatkan menunjukkan metode *Naïve Bayes* memiliki kinerja yang lebih baik dengan tingkat akurasi 70%, sedangkan metode *k-nearest neighbor* memiliki tingkat akurasi yang cukup rendah yaitu 40%.

Penelitian untuk mengklasifikasikan dokumen bahasa Bali menggunakan metode *Naïve Bayes* dengan model *multinomial* juga telah dilakukan sebelumnya [7]. Metode *multinomial Naïve Bayes* sering digunakan dalam penelitian tentang klasifikasi teks karena kesederhanaan dan efektivitasnya yang menggunakan ide dasar probabilitas gabungan dari kata-kata dan kategori untuk memperkirakan probabilitas kategori pada suatu dokumen. Setelah dilakukan *preprocessing* pada dokumen, dilakukan pula seleksi fitur dengan metode *information gain*. Dokumen dikelompokkan ke dalam kategori seni budaya dan upacara, dengan jumlah data sejumlah 100 dokumen untuk masing-masing kategori. Penelitian menghasilkan nilai rata-rata akurasi dari 10 *fold cross validation* sebesar 95,22%.

Dilakukan penelitian tentang klasifikasi konten *e-government* dengan *Naïve Bayes classifier* menggunakan pembobotan TF-IDF (*term frequency-inverse document frequency*) [8]. Metode ini dipilih karena memiliki kinerja yang baik terhadap pengklasifikasian data dokumen yang mengandung angka maupun teks. Dokumen diklasifikasikan ke dalam kategori ekonomi dan politik. Penelitian menghasilkan akurasi yang cukup baik yaitu sebesar 85%.

Telah dilakukan pula penelitian tentang *sentiment analisys* di jejaring sosial *Twitter* menggunakan algoritma *naïve Bayes* dengan seleksi fitur *mutual information* [9]. Metode ini dipilih karena sederhana, memiliki kecepatan yang cukup tinggi, dan menghasilkan akurasi yang baik dalam *sentiment analysis*. Data yang digunakan adalah sejumlah 500 *tweet* tentang pariwisata Lombok. Data dikelompokan ke dalam 2 kategori yaitu sentimen positif dan sentimen negatif. Akurasi yang didapatkan melalui pengujian 10-*fold cross validation* adalah 96.2% tanpa menggunakan seleksi fitur *mutual information* dan 97.9% dengan menggunakan seleksi fitur *mutual information*.

Telah dilakukan perbandingan terhadap beberapa jenis *multinomial naïve Bayes* dalam mengelompokan dokumen [10]. *Dataset* yang digunakan dalam penilitian ini adalah 20 *newsgroups*, *industry sector*, WebKB, dan *Reuters*-21578. *Dataset* tersebut merupakan *dataset* yang sering digunakan dalam penilitian tentang klasifikasi teks. Penilitian membuktikan bahwa modifikasi *transformed weight-normalized complement naïve Bayes* (TWNBC) tidak diperlukan untuk mendapatkan hasil yang optimal untuk beberapa *dataset*. Akan tetapi, penggunaan TF-IDF dalam pembobotan kata terbukti dapat meningkatkan akurasi secara signifikan pada sebagian besar *dataset.* Selain itu, penggunaan normalisasi panjang dokumen dapat mengurangi akurasi dari pembobotan dengan TF-IDF.

Berdasarkan berbagai penelitian yang telah dijelaskan sebelumnya, dapat disimpulkan bahwa metode *multinomial naïve Bayes* serta metode TF-IDF memiliki hasil yang baik ketika digunakan untuk mengelompokan artikel. Oleh karena itu, penelitian untuk klasifikasi artikel *online* tentang gempa di Indonesia dapat dilakukan dengan menggunakan kedua metode tersebut.

2.1. *Text Mining*

Text mining merupakan teori tentang pengolahan kumpulan teks dengan tujuan untuk mengetahui dan mengekstrak informasi bermanfaat dari kumpulan teks tersebut. Informasi didapatkan dengan cara identifikasi dan eksplorasi pola yang menarik dari sumber data. *Text mining* merupakan bidang khusus dari *data mining* dimana data yang digunakan adalah data tekstual yang tidak terstruktur [8]. Bagian – bagian dari *text mining* meliputi *classification* (klasifikasi), *clustering*, dan *association* [3].

2.2. Klasifikasi Teks

Klasifikasi teks merupakan salah satu aplikasi dari *text mining*. Klasifikasi teks adalah proses pengelompokan teks berdasarkan kata, frase, atau kombinasinya untuk menentukan kategori yang telah ditetapkan sebelumnya (*supervised learning*) [2].

2.3. *Text Preprocessing*

Text preprocessing merupakan proses untuk mentransformasikan teks ke dalam kumpulan kata. Teks merupakan data yang tidak terstruktur, yang mana cukup sulit untuk diproses dengan komputer. Operasi numerik pun tidak dapat diaplikasikan pada data teks. Oleh karena itu, perlu dilakukan *preprocessing* pada teks untuk mendapatkan data yang dapat diolah menggunakan komputer. Terdapat 3 langkah mendasar yang dilakukan dalam *text preprocessing*, yaitu *tokenization*, *stemming*, dan *stopword removal* [3].

2.3.1. *Tokenization*

Tokenization adalah proses untuk memotong teks menjadi kata / *token* yang dipisahkan oleh spasi atau tanda baca. Proses *tokenization* menerima teks sebagai *input* dan menghasilkan kumpulan *token* sebagai *output*. Selanjutnya, *token* yang mengandung karakter spesial atau angka akan dihilangkan, lalu *token* akan diubah menjadi *lowercase* [3].

2.3.2. *Stemming*

Proses selanjutnya dalam *text preprocessing* adalah *stemming*. Pada tahap ini, *token* yang didapatkan dari proses tokenization diubah menjadi bentuk dasarnya. Proses *stemming* biasanya dilakukan pada kata benda, kata kerja, dan kata sifat [3].

2.3.3. *Stop-word removal*

Pada proses *stop-word removal*, dilakukan penghapusan *stop word* dari daftar *token* atau kata yang sudah diproses dengan tahap *stemming*. *Stop word* merupakan kata yang tidak berhubungan dengan konteks dari teks, sehingga perlu dihilangkan untuk meningkatkan efisiensi dari proses *training* atau klasifikasi [3]. Contoh dari *stop word* dalam bahasa Indonesia adalah "di" dan "ke". Kata – kata tersebut tidak dapat mewakili konteks dari dokumen karena terdapat pada hampir seluruh dokumen.

2.4. *Multinomial Naïve Bayes Classifier*

Naïve Bayes merupakan salah metode pembelajaran mesin probabilistik. Seperti namanya, metode ini mengasumsikan bahwa setiap atribut dari data tidak bergantung satu sama lain. Pada dasarnya, asumsi bahwa setiap kata tidak bergantung satu dengan yang lain pada metode *Naïve Bayes* ini berlawanan dengan keadaan sebenarnya. Hal ini dikarenakan suatu dokumen atau teks perlu memiliki kata yang saling berhubungan agar dokumen tersebut memiliki makna. Akan tetapi, metode ini terbukti mampu memberikan hasil yang cukup memuaskan apabila diterapkan di bidang klasifikasi teks [3].

Salah satu model dari *Naïve Bayes* yang sering digunakan dalam klasifikasi teks adalah *multinomial Naïve Bayes* [4]. *Multinomial Naïve Bayes* merupakan metode *supervised learning*, sehingga setiap data perlu diberikan label sebelum dilakukan *training*. Probabilitas suatu dokumen *d* berada di kelas *c* dapat dihitung menggunakan Persamaan (1) [4].

$$P(c|d) \propto P(c) \prod_{k=1}^{n} P(t_k|c) \qquad (1)$$

dimana :

$P(c|d)$: Probabilitas dokumen d berada di kelas c

$P(c)$: Prior probability suatu dokumen berada di kelas c

$\{t_1, t_1, t_1, \dots, t_n\}$: Token dalam dokumen d yang merupakan bagian dari vocabulary dengan jumlah n

$P(t_k|c)$: Probabilitas bersyarat term t_k berada di dokumen pada kelas c

Klasifikasi dokumen bertujuan untuk menentukan kelas terbaik untuk suatu dokumen. Kelas terbaik dalam klasifikasi *Naïve Bayes* ditentukan dengan mencari *maximum a posteriori* (MAP) kelas c_{map} melalui Persamaan (2).

$$c_{\text{map}} = \arg\max_{c \in C} \hat{P}(c) \prod_{k=1}^{n} \hat{P}(t_k|c) \qquad (2)$$

P ditulis dengan $\hat{P}$ karena nilai sebenarnya dari $P(c|d)$ dan $P(t_k|c)$ belum diketahui, yang akan dihitung pada saat proses *training* [4].

Pada Persamaan (2), terdapat banyak probabilitas bersyarat yang dikalikan. Hal ini dapat menyebabkan *floating point underflow*. Karena itu, proses perhitungan akan lebih baik apabila dilakukan penjumlahan pada logaritma dari probabilitas. Kelas dengan logaritma dari probabilitas tertinggi merupakan kelas dengan probabilitas terbaik untuk dokumen; log(*xy*) = log(*x*) + log(*y*). Persamaan (2) yang menggunakan logaritma dari probabilitas dapat dinyatakan dalam Persamaan (3) [4].

$$c_{\text{map}} = \arg\max_{c \in C} \left[\log \hat{P}(c) + \sum_{1 \le k \le n} \log \hat{P}(t_k|c)\right] \quad (3)$$

$\hat{P}(c)$ dan $\hat{P}(t_k|c)$ didapatkan dengan menghitung *maximum likelihood*, yang merupakan frekuensi relatif dari parameter. Untuk *prior*, dapat digunakan Persamaan (4).

$$\hat{P}(c) = \frac{N_c}{N} \quad (4)$$

dimana :

$\hat{P}(c)$: Prior probability suatu dokumen berada di kelas c
N_c : Jumlah dokumen dengan kelas c
N : Jumlah seluruh dokumen

$\hat{P}(t|c)$ merupakan probabilitas frekuensi relatif *term t* dalam dokumen berada di kelas *c*, yang dapat dihitung menggunakan Persamaan (5).

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \quad (5)$$

dimana :

$\hat{P}(t|c)$: Probabilitas bersyarat term t berada di dokumen pada kelas c
T_{ct} : Jumlah kemunculan term t pada dokumen dengan kategori c
$\sum_{t' \in V} T_{ct'}$: Jumlah frekuensi seluruh term pada kelas c

Perhitungan *maximum likelihood* memiliki kelemahan, yaitu suatu kata dalam kelas yang tidak terlihat pada data *training* akan memiliki nilai 0. Hal ini menyebabkan perhitungan $P(c|d)$ menghasilkan nilai 0, karena setiap bilangan yang dikalikan dengan 0 akan menghasilkan 0. Untuk mengatasi masalah ini, diterapkan teknik *add-one* atau *Laplace smoothing*, sehingga Persamaan (5) berubah menjadi Persamaan (6).

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'})+B'} \quad (6)$$

dimana :

B : Jumlah seluruh term pada vocabulary

Sedangkan untuk rumus *multinomial naïve Bayes* dengan menggunakan pembobotan TF-IDF dapat dilihat pada persamaan (7) [11].

$$\hat{P}(t|c) = \frac{W_{ct}+1}{(\sum_{w' \in V} W_{ct'})+B'} \quad (7)$$

dimana :

W_{ct} : Bobot TF-IDF term t pada dokumen dengan kategori c
$\sum_{w' \in V} W_{ct'}$: Jumlah bobot TF-IDF seluruh *term* pada kelas *c*
B' : Jumlah IDF seluruh *term* pada *vocabulary*

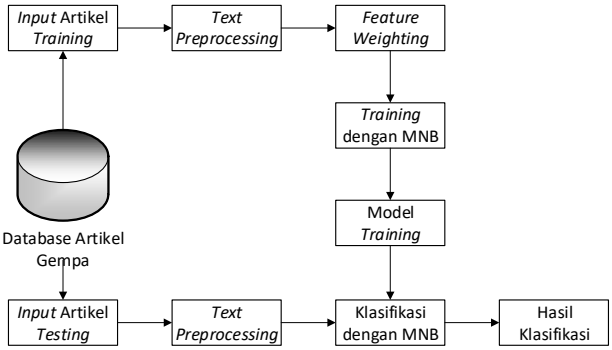
2.5. *K-fold Cross Validation*

K-fold cross validation merupakan salah satu teknik validasi silang dengan cara membagi data menjadi *k* bagian dengan ukuran yang sama. Pelatihan dan pengujian dilakukan sebanyak *k* kali. Pada percobaan pertama, *subset* S1 diberlakukan sebagai data pengujian, dan *subset* lainnya digunakan sebagai data *training*. Pada percobaan ke-2, *subset* S2 diberlakukan sebagai data pengujian, kemudian *subset* lainnya digunakan sebagai data *training*. Proses ini dilakukan sampai *k* kali dimana *subset* Sk dijadikan data pengujian [12].

3. METODE PENELITIAN

Pada penelitian ini terdapat 6 kategori klasifikasi artikel yang digunakan yaitu kategori Ekonomi Gempa, Kesehatan Gempa, Pariwisata Gempa, Ekonomi Non-gempa, Kesehatan Non-gempa, dan Pariwisata Non-gempa. Pengumpulan data diarahkan oleh pakar Ilmu Komunikasi dengan kualifikasi Master. Artikel Kesehatan Gempa dan Kesehatan Non-gempa dikumpulkan dari kanal *health.detik.com* serta *www.liputan6.com/health*. Artikel Ekonomi Gempa dan Ekonomi Non-gempa dikumpulkan dari kanal *finance.detik.com*, *economy.okezone.com*, serta *ekonomi.kompas.com*. Sedangkan untuk artikel Pariwisata Gempa dan Pariwisata Non-gempa, dikumpulkan dari kanal *travel.detik.com* serta *travel.kompas.com*. Artikel yang dikumpulkan adalah sejumlah 1000 artikel. Dari 1000 artikel yang terkumpul, terdapat 100 artikel untuk label Kesehatan Gempa, 100 artikel untuk label Ekonomi Gempa, 100 artikel untuk label Pariwisata Gempa, 230 artikel untuk label Kesehatan Non-gempa, 230 artikel untuk label Ekonomi Non-gempa, 240 artikel untuk label Pariwisata Non-gempa.

Rancangan dari sistem klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes text classification* terdiri dari beberapa tahapan, yang dapat dilihat pada Gambar 1.



Gambar 1. Rancangan sistem klasifikasi artikel tentang gempa di indonesia.

3.1. Input *Artikel* Training *dan* Testing

Pada tahap ini, artikel yang telah dikumpulkan dibagi menjadi 2 yaitu artikel *training* dan artikel *testing*. Artikel *training* digunakan untuk membuat model klasifikasi sedangkan artikel *testing* digunakan untuk menguji model yang telah dibuat.

3.1.1. *Artikel training*

Artikel *training* yang sebelumnya telah diberi label kategori dimasukan ke dalam sistem untuk diproses. Artikel yang didapatkan dari *subdomain health.detik.com*, dan *www.liputan6.com/health* diberi kategori Kesahatan Gempa dan Kesehatan Non-gempa. Artikel yang didapatkan dari *subdomain finance.detik.com*, *economy.okezone.com*, dan *ekonomi.kompas.com* diberi kategori Ekonomi Gempa dan Ekonomi Non-gempa. Sedangkan untuk artikel yang didapatkan dari *subdomain travel.detik.com* dan *travel.kompas.com* diberi kategori Pariwisata dan Pariwisata Non-gempa. Setelah artikel diberi label, dilakukan *preprocessing* dan pembobotan pada artikel, yang kemudian di-*training* menggunakan *naïve Bayes classifier*.

3.1.2. *Artikel testing*

Artikel *testing* merupakan artikel yang diambil dari *dataset* tetapi tidak diberi label seperti artikel *training*. Artikel *testing* dimasukan ke sistem untuk diprediksikan kategorinya. Sebelum diklasifikasikan, artikel *testing* juga terlebih dahulu melewati tahap *preprocessing*.

3.2. *Text Preprocessing*

Text preprocessing yang dilakukan pada penelitian ini dibagi menjadi 3 tahap, yaitu tahap *tokenization*, *stemming*, dan *stop-word removal*.

3.2.1. *Tokenization*

Tokenization merupakan untuk mentransformasikan artikel menjadi kumpulan kata yang disebut *terms*. Pada *tokenization* juga dilakukan penghilangan tanda baca. Hal ini dilakukan karena tanda baca tidak dapat digunakan sebagai *terms* karena terdapat pada hampir seluruh dokumen. Sebelum proses *tokenization*, terlebih dahulu dilakukan proses *case folding* atau mengubah setiap kata menjadi huruf kecil. Tujuannya adalah agar tidak terjadi kesalahan interpretasi oleh komputer ketika ada dua kata yang sama tapi dianggap berbeda karena perbedaan huruf besar dan huruf kecil.

3.2.2. *Stemming*

Proses *stemming* dilakukan dengan menggunakan algoritma Nazief dan Adriani karena artikel yang digunakan pada penelitian merupakan artikel berbahasa Indonesia. Selain itu, algoritma Nazief dan Adriani terbukti dapat memberikan akurasi yang lebih akurat dibandingkan dengan algoritma lainnya seperti Arifin dan Setiono, Vega, serta Tala [13]. Algoritma Nazief dan Adriani melakukan stemming dengan menghilangkan *inflection suffixes* (“-lah”, “-kah”, “-ku”, “-mu”, atau “-nya”), *possesive pronouns* (“-ku”, “-mu”, atau “-nya”), *derivation suffixes* (“-i”, “-an” atau “-kan”) dan *derivation prefixes* (“di-”,“ke-”,“se-”,“te-”, “be-”, “me-”, atau “pe-”), kemudian mencocokan kata dengan kata yang ada di kamus [14].

3.2.3. *Stop-word removal*

Stop-word removal merupakan proses menghilangkan *stop words* yang tidak dapat mewakili isi artikel. Proses ini dilakukan untuk meningkatkan efisiensi dalam proses *training* maupun klasifikasi. *Stop words* yang dihillangkan didasarkan dari *stop words* pada *library* Sastrawi.

3.3. *Feature Weighting*

Feature weighting merupakan suatu proses untuk menghitung serta memberi bobot pada suatu *feature* sebagai derajat kepentingannya. *Term frequency* dan pembobotan TF-IDF merupakan metode yang sering digunakan dalam pembobotan kata[3].

Metode TF-IDF (*Term Frequency – Inverse Document Frequency*) merupakan suatu metode yang menggabungkan 2 cara untuk memberikan bobot pada kata, yaitu dengan menghitung *term frequency* dan melakukan perhitungan *invers* dari jumlah dokumen yang mengandung kata tersebut (IDF) [6]. Karena dilakukan pula perhitungan IDF, maka metode TF-IDF membutuhkan referensi dari seluruh dokumen (*corpus*) [3]. Perhitungan TF dan IDF dapat dilakukan dengan Persamaan (8) dan (9) [4].

$$TF(d,t) = f(d,t) \quad (8)$$

$$IDF(t) = log\left(\frac{N_d}{df(t)}\right) \quad (9)$$

dimana :
$TF(d,t)$: Term frequency
$f(d,t)$: Frekuensi *term* t pada dokumen d
$IDF(t)$: Inverse document frequency
N_d : Jumlah dokumen keseluruhan
$df(t)$: Jumlah dokumen yang mengandung *term* t

Sehingga untuk perhitungan TF-IDF dari suatu kata pada dokumen dapat dilakukan dengan Persamaan.

$$TF - IDF = TF(d,t) \,.\, IDF(t) \quad (10)$$

3.4. *Training*

Pada penelitian ini, *training* dan klasifikasi dilakukan menggunakan metode *Naïve Bayes* dengan model *multinomial*. Proses *training* diawali dengan menghitung probabilitas *prior* dari setiap kategori menggunakan Persamaan (4). Setelah didapatkan probabilitas *prior* dari tiap kategori, proses *training* dilanjutkan dengan menghitung probabilitas suatu *feature* terdapat pada suatu kategori. Proses perhitungan dilakukan dengan Persamaan (7).

Pada tahap klasifikasi, ada kemungkinan terdapat suatu *feature* pada artikel *testing* yang tidak pernah muncul di artikel *training*. Apabila ditemukan kasus seperti ini, maka $P(t|c)$ dari *feature* tersebut adalah nilai $P(t|c)$ paling kecil dari tiap kategori.

3.5. Klasifikasi

Tujuan dari proses klasifikasi adalah untuk mengetahui kategori dari suatu artikel. Proses ini memanfaatkan model *Naïve Bayes* yang telah didapatkan saat *training* untuk melakukan perhitungan probabilistik pada artikel yang ingin diklasifikasikan. Penentuan kategori suatu artikel dilakukan menggunakan Persamaan (3). Pada Persamaan (3), dilakukan perbandingan terhadap probabilitas suatu artikel berada di suatu kategori *c* untuk tiap kategori. Sebelum dilakukan klasifikasi, artikel *testing* terlebih dahulu melewati proses *preprocessing*.

3.6. Evaluasi

Pengujian yang dilakukan dalam penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes text classification* adalah dengan teknik *k-fold cross validation*. Nilai *k* yang digunakan adalah 5 karena jumlah *dataset* yang cukup terbatas, dimana dataset untuk artikel gempa sendiri hanya terdapat 100 artikel untuk tiap kategori.

Teknik yang digunakan untuk melakukan evaluasi dalam klasifikasi pada penelitian ini adalah dengan menghitung *recall*, *precision*, dan *f-measure*. Teknik ini menggunakan *confusion matrix* sebagai acuan perhitungan [11].

Recall untuk kelas *c* merupakan perbandingan dari jumlah dokumen yang diklasifikasikan benar pada kelas *c* dengan jumlah seluruh dokumen yang sebenarnya berada pada kelas *c*. Perhitungan *recall* pada suatu kelas *c* dapat dilakukan dengan Persamaan (11) [11].

$$Recall_c = \frac{TP(Kelas-c)}{Total(Kelas-c)} \quad (11)$$

Precision untuk kelas *c* merupakan perbandingan dari jumlah dokumen yang diklasifikasikan benar pada kelas *c* dengan jumlah dokumen yang diklasifikasikan sebagai kelas *c*. *Precision* pada suatu kelas *c* dapat dihitung dengan menggunaan Persamaan (12) [11].

$$Precision_c = \frac{TP(Kelas-c)}{Prediksi(Kelas-c)} \quad (12)$$

Sedangkan *f-measure* merupakan nilai yang mewakili seluruh kinerja sistem yang merupakan penggabungan nilai *recall* dan *rrecision*. *F-measure* dapat dihitung menggunakan Persamaan (13) [2].

$$F - measure_c = \frac{2PR}{P+R} \quad (13)$$

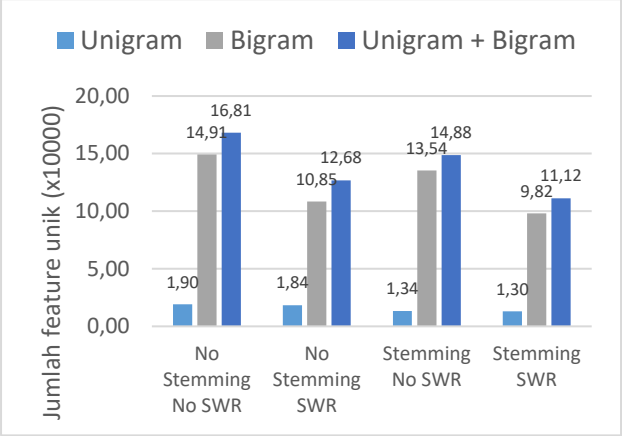
Pada setiap percobaan, evaluasi dilakukan dengan menghitung *recall*, *precision*, dan *f-measure* dari model. Nilai *recall*, *precision*, dan *f-measure* untuk tiap percobaan didapatkan dengan mencari nilai rata-rata dari *recall*, *precision*, dan *f-measure* per kategori. Performa model secara keseluruhan didapatkan dengan menghitung nilai rata-rata *recall*, *precision*, dan *f-measure* dari seluruh percobaan.

Terdapat beberapa skema pengujian yang dilakukan pada penelitian ini, antara lain adalah klasifikasi hanya dengan menggunakan *feature unigram*, klasifikasi hanya dengan menggunakan *feature bigram*, dan klasifikasi dengan *feature unigram* dan *bigram*. Selain itu, dilakukan juga pengujian dengan *stemming*, tanpa *stemming*, dengan *stopwords removal*, dan tanpa *stopwords removal*.

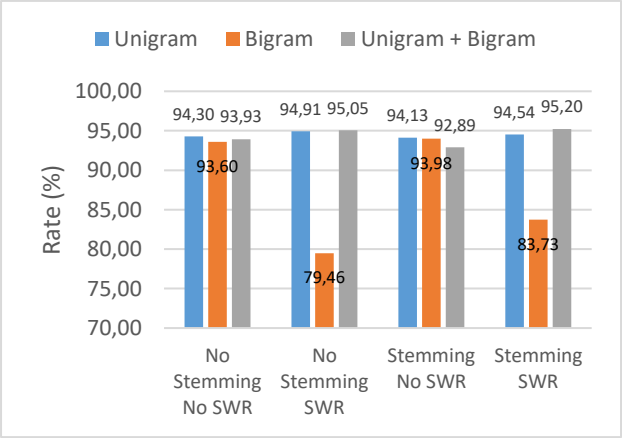
4. HASIL DAN PEMBAHASAN

Berdasarkan beberapa skema pengujian yang dilakukan dengan 5 perulangan 5-*fold cross validation*, didapatkan nilai *precision* dan *recall* yang kemudian diwakilkan oleh nilai *f-measure*. Selain itu, didapatkan juga panjang *vocabulary* dari masing-masing pengujian.

Grafik perbandingan panjang *vocabulary* dan perbandingan nilai *f-measure* untuk tiap pengujian dapat dilihat pada Gambar 2 dan Gambar 3.



Gambar 2. Pengaruh jenis pengujian terhadap ukuran *vocabulary*.



Gambar 3. Pengaruh jenis pengujian terhadap nilai *f-measure*.

Pada Gambar 2, ukuran *vocabulary* yang didapatkan dari tiap jenis *feature* berbeda-beda satu sama lain. *Feature unigram* memiliki ukuran *vocabulary* jauh lebih kecil dibandingkan dengan *feature bigram*. Hal ini disebabkan oleh pasangan kata yang bervariasi dalam tiap dokumen. Ukuran *vocabulary* dari penggunaan *feature unigram* bersamaan dengan *feature bigram* sendiri merupakan jumlah dari ukuran 2 jenis *feature* tersebut. Selain itu, dapat dilihat juga penggunaan *stemming* dan *stopwords removal* juga dapat mengurangi ukuran dari *vocabulary*.

Pada Gambar 3, dapat dilihat bahwa penggunaan *stemming* mengurangi nilai *f-measure* dari pengujian menggunakan *feature unigram*. Hal ini juga berlaku pada pengujian yang menggunakan *feature unigram* sekaligus *bigram* dan menyertakan *stopwords*. Akan tetapi, nilai *f-measure* yang didapatkan pada pengujian dengan menggunakan *stemming* justru mengalami peningkatan pada *feature bigram* Nilai *f-measure* dari *feature unigram* sekaligus *bigram* yang tidak menyertakan *stopwords* juga mengalami peningkatan ketika dilakukan *stemming*.

Pada Gambar 3 juga dapat dilihat bahwa penghilangan *stopwords* dapat meningkatkan nilai *f-measure* untuk pengujian dengan *feature unigram* dan juga *feature unigram* sekaligus *bigram*. Akan tetapi, *stopwords removal* pada pengujian dengan *feature bigram* justru dapat mengurangi nilai *f-measure* secara signifikan.

4.1. Pengujian tanpa *Stemming* dan *Stopwords Removal*

Pada skema pengujian ini, tidak dilakukan *stemming* maupun *stopwords removal* pada data *training*. Hasil *training* menggunakan seluruh *dataset* menghasilkan *vocabulary* dengan panjang 19016 untuk *feature unigram* dan 149125 untuk *feature bigram*. 5 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel I, Tabel II, Tabel III, dan Tabel IV.

TABLE I. FEATURE UNIGRAM TANPA STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL GEMPA

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.123	kesehatan	241.043	gili	290.8
rp	191.195	korban	232.627	lombok	264.151
bank	190.703	gempa	168.47	gempa	220.717
bantuan	181.44	lombok	150.315	pariwisata	207.101
lombok	176.101	palu	149.352	wisatawan	203.849

TABLE II. FEATURE UNIGRAM TANPA STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
saham	304.75	jantung	369.96	pantai	243.68
rp	291.20	anda	343.13	gili	227.26
harga	265.44	kanker	266.28	wisata	223.99
banjir	233.93	serangan	254.43	pengunjung	218.87
tol	227.72	seks	245.70	gunung	214.86

TABLE III. FEATURE BIGRAM TANPA STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.2	korban gempa	122.9	gili trawangan	116.1
rp triliun	88.87	rumah sakit	99.16	gunung rinjani	92.9
sri mulyani	88.33	lombok utara	85.1	gili air	90.98
di palu	86.87	kementerian kesehatan	82.44	di lombok	83.45
di lombok	82.63	rsud tanjung	76.60	di gili	76.52

TABLE IV. FEATURE BIGRAM TANPA STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
jalan tol	170.61	serangan jantung	221.91	kompas com	108.7
rp triliun	146.24	dikutip dari	90.658	gili trawangan	106.8
menjadi rp	115.77	bunuh diri	88.874	gunung rinjani	105.1
harga rokok	109.39	nyeri dada	88	gili air	90.98
rp ribu	92.947	getah bening	81.83	di bali	76.52

Pada tabel yang disajikan, dapat dilihat beberapa *feature* yang sama pada kategori ekonomi, kesehatan, dan pariwisata untuk artikel tentang gempa maupun artikel non-gempa. Hal ini disebabkan karena artikel dengan kategori yang sama, baik artikel tersebut merupakan artikel gempa maupun artikel non-gempa, memiliki beberapa *feature* kunci yang sama. Akan tetapi, *feature-feature* tersebut tidak terlalu berpengaruh dalam menentukan apakah suatu artikel termasuk ke dalam artikel gempa atau non gempa, karena terdapat beberapa *feature* yang hanya terdapat pada artikel gempa. *Feature-feature* tersebut memiliki peran yang penting dalam mengelompokan suatu artikel ke dalam kategori gempa. Selain dari *feature* yang hanya ada untuk artikel gempa tersebut, *feature* lain di dalam *vocabulary* sangat beragam untuk masing-masing kategori, sehingga menghasilkan akurasi yang tinggi saat dilakukan klasifikasi artikel.

Hasil yang didapatkan dari skema pengujian ini melalui 5-*fold cross validation* dapat dilihat pada Tabel V.

TABLE V. HASIL PENGUJIAN TANPA STEMMING DAN STOPWORDS REMOVAL

Jenis Feature	Precision	Recall	F-Measure	Standar Deviasi
Unigram	95.21%	93.66%	94.30%	1.09%
Bigram	93.12%	94.46%	93.60%	1.69%
Unigram dan *Bigram*	95.21%	93.04%	93.93%	1.43%

4.2. Pengujian tanpa *Stemming* dengan *Stopwords Removal*

Pada skema pengujian ini, tidak dilakukan *stemming*, akan tetapi dilakukan *stopwords removal* pada data *training*. Hasil *training* menggunakan seluruh *dataset* menghasilkan *vocabulary* dengan panjang 18374 untuk *feature unigram* dan 108474 untuk *feature bigram*. 5 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel VI, Tabel VII, Tabel VIII, dan Tabel IX.

TABLE VI. FEATURE UNIGRAM TANPA STEMMING DAN STOPWORDS PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.13	kesehatan	241	gili	290.8
rp	191.2	korban	232.6	lombok	264.2
bank	190.7	gempa	168.5	gempa	220.7
bantuan	181.44	lombok	150.3	pariwisata	207.1
lombok	176.1	palu	149.4	wisatawan	203.8

TABLE VII. FEATURE UNIGRAM TANPA STEMMING DAN STOPWORDS PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
saham	304.75	jantung	369.96	pantai	243.7
rp	291.2	kanker	266.28	gili	227.3
harga	265.44	serangan	254.43	wisata	224
banjir	233.93	seks	245.7	pengunjung	218.9
tol	227.72	pneumonia	236.05	gunung	214.9

TABLE VIII. FEATURE BIGRAM TANPA STEMMING DAN STOPWORDS PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.18	korban gempa	123.85	gili trawangan	116.1
rp triliun	88.87	rumah sakit	99.161	gunung rinjani	92.9
sri mulyani	88.327	lombok utara	84.706	gili air	90.98
rp juta	79.624	kementerian kesehatan	82.442	jalur pendakian	65.41
kantor cabang	74.912	rsud tanjung	76.604	gempa bumi	64.21

TABLE IX. FEATURE BIGRAM TANPA STEMMING DAN STOPWORDS PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
jalan tol	170.61	serangan jantung	221.91	kompas com	108.7
rp triliun	146.24	nyeri dada	90	gili trawangan	108.1
harga rokok	111.28	getah bening	81.83	gunung rinjani	105.1
rp ribu	92.947	kelenjar getah	79.683	gili air	90.98
dollar as	82.62	paru paru	71.132	kembang api	75.69

Pada tabel yang disajikan, terdapat beberapa *feature* yang dihilangkan baik untuk *feature unigram* maupun *feature bigram* dibandingkan dengan pengujian tanpa menghilangkan *stop words*. Untuk

feature unigram, pengaruh dari penghilangan *stop words* tidak terlalu tinggi terhadap proses klasifikasi. Hal ini dikarenakan pemberian bobot *feature* dilakukan dengan TF-IDF yang mana mempertimbangkan kemunculan *feature* diseluruh dokumen pada *corpus*. *Stop words* merupakan kata yang umum dan muncul di hampir seluruh dokumen, sehingga IDF-nya lebih rendah dibandingkan dengan *feature* lainnya, dan menyebabkan bobot-nya tidak terlalu tinggi.

Akan tetapi, penghilangan *stop words* pada *feature bigram* memiliki pengaruh yang cukup signifikan. Hal ini terjadi karena penghilangan *stop words* akan menyebabkan 2 *terms* yang dipisahkan oleh *stop words* dianggap bersebelahan. Konsep dasar dari *bigram* yang menyatukan 2 *terms* yang bersebalahan akan diabaikan. Imbasnya, *feature* yang terdapat pada *vocabulary* menjadi tidak representatif terhadap dokumen. Contoh *feature* penting dengan bobot tinggi yang dihilangkan pada proses *stop words removal* adalah "gempa di" dan "para korban" untuk artikel gempa serta "menjadi rp" dan "di bali" untuk artikel non-gempa.

Hasil yang didapatkan dari skema pengujian ini melalui 5-*fold cross validation* dapat dilihat pada Tabel X.

TABLE X. HASIL PENGUJIAN TANPA STEMMING DAN STOPWORDS.

Jenis Feature	Precision	Recall	F-Measure	Standar Deviasi
Unigram	94.58%	95.51%	94.91%	1.54%
Bigram	79.93%	84.96%	79.46%	2.99%
Unigram dan *Bigram*	94.84%	95.46%	95.05%	1.65%

4.3. Pengujian dengan *Stemming* tanpa *Stopwords Removal*

Pada skema pengujian ini, dilakukan *stemming* tanpa *stopwords removal* pada data *training*. Hasil *training* menggunakan seluruh *dataset* menghasilkan *vocabulary* dengan panjang 13354 untuk *feature unigram* dan 135399 untuk *feature bigram*. 5 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel XI, Tabel XII, Tabel XIII, dan Tabel XIV.

TABLE XI. FEATURE UNIGRAM DENGAN STEMMING DAN STOPWORDS PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.13	korban	231.91	gili	290.8
rp	191.2	sehat	227.37	lombok	264.2
bank	190.7	gempa	168.47	gempa	221.8
lombok	176.1	lombok	150.31	pariwisata	211.2
bencana	175.91	palu	149.35	wisatawan	203.8

TABLE XII. FEATURE UNIGRAM DENGAN STEMMING DAN STOPWORDS PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
saham	310.34	jantung	373.9	pantai	248.1
rp	291.2	anda	339	wisata	227.5
harga	256.82	sakit	298.4	gili	227.3
banjir	233.91	kanker	271	gunung	220.5
tol	225.15	serang	249.5	unjung	216.6

TABLE XIII. *FEATURE BIGRAM* DENGAN *STEMMING* DAN *STOPWORDS* PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.18	korban gempa	122.87	gili trawangan	116.1
rp triliun	88.87	rumah sakit	100.18	gunung rinjani	92.9
sri mulyani	88.327	menteri sehat	86.387	gili air	90.98
di palu	86.87	lombok utara	85.099	di lombok	83.45
di lombok	82.634	rsud tanjung	76.604	di gili	76.52

TABLE XIV. FEATURE BIGRAM DENGAN STEMMING DAN STOPWORDS PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
jalan tol	170.61	serang jantung	221.91	kompas com	108.7
rp triliun	146.24	orang yang	101.25	gili trawangan	106.8
jadi rp	118.24	kutip dari	90.378	gunung rinjani	105.1
harga rokok	109.39	bunuh diri	88.874	gili air	90.98
rp ribu	92.328	nyeri dada	88	ada di	88.11

Pada tabel yang disajikan, terdapat beberapa *feature* yang berbeda dibandingkan dengan pengujian tanpa melakukan *stemming*. Proses *stemming* untuk *feature unigram* dapat mengurangi akurasi dari *model* walaupun tidak terlalu signifikan. Hal ini dapat dilihat dari *terms* yang penting untuk suatu kategori yang telah diubah menjadi kata dasarnya memiliki bobot yang lebih rendah dibandingkan dengan sebelum dilakukan *stemming*. Contoh *term* penting yang mengalami penurunan bobot akibat *stemming* dapat dilihat pada Tabel XV.

TABLE XV. FEATURE UNIGRAM PENTING YANG MENGALAMI PERUBAHAN BOBOT SETELAH STEMMING.

Term		Kategori
Sebelum *stemming*	Setelah *stemming*	
bantuan (TF = 259, IDF = 0.5686)	bantu (TF = 204, IDF = 0.8894)	Ekonomi Gempa
pendaki (TF = 24, IDF = 1.585), pendakian (TF = 65, IDF = 1.6198)	daki (TF = 92, IDF = 1.4685)	Pariwisata Gempa

Untuk *feature bigram*, penerapan *stemming* justru meningkatkan akurasi dari model. Hal ini dikarenakan terdapat beberapa pasangan *terms* yang memiliki relevansi cukup tinggi pada suatu kategori dianggap berbeda karena berbeda imbuhan. Hal ini merupakan *general phenomenon* pada *text behavior* dimana suatu pasangan kata memiliki relevansi yang tinggi terhadap suatu dokumen terlepas dari ada atau tidaknya imbuhan. Oleh karena itu, penerapan *stemming* untuk *feature bigram* dapat meningkatkan bobot dari pasangan *term* yang merupakan *feature* penting untuk suatu kategori. Contoh pasangan *term* yang mengalami kenaikan bobot setelah dilakukan *stemming* dapat dilihat pada Tabel XVI.

TABLE XVI. FEATURE BIGRAM PENTING YANG MENGALAMI PENINGKATAN BOBOT SETELAH STEMMING.

Feature		Kategori
Sebelum *stemming*	Setelah *stemming*	
kementerian keuangan (w = 52.3059)	menteri uang (w = 73.188)	Ekonomi Gempa
kementerian kesehatan (w = 82.44171)	menteri sehat (w = 86.3875)	Kesehatan Gempa
dapat menyebabkan (w = 65.531)	dapat sebab (w = 68.8886)	Kesehatan Non-gempa

Hasil yang didapatkan dari skema pengujian ini melalui 5-*fold cross validation* dapat dilihat pada Tabel XVIII.

TABLE XVII. HASIL PENGUJIAN DENGAN STEMMING DAN STOPWORDS.

Jenis Feature	Precision	Recall	F-Measure	Standar Deviasi
Unigram	94.93%	93.63%	94.13%	1.32%
Bigram	93.68%	94.57%	93.98%	1.99%
Unigram dan *Bigram*	94.66%	91.71%	92.89%	1.66%

4.4. Pengujian dengan *Stemming* dan *Stopwords Removal*

Pada skema pengujian ini, dilakukan *stemming* dan *stopwords removal* pada data *training*. Hasil *training* menggunakan seluruh *dataset* menghasilkan *vocabulary* dengan panjang 13033 untuk *feature unigram* dan 98154 untuk *feature bigram*. 10 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel XVIII, Tabel XIX, Tabel XX, dan Tabel XXI.

TABLE XVIII. FEATURE UNIGRAM DENGAN STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.13	korban	231.91	gili	290.8
rp	191.2	sehat	227.37	lombok	264.2
bank	190.7	gempa	168.47	gempa	221.8
lombok	176.1	lombok	150.31	pariwisata	211.2
bencana	175.91	palu	149.35	wisatawan	203.8

TABLE XIX. FEATURE UNIGRAM DENGAN STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
saham	310.34	jantung	373.94	pantai	248.1
rp	291.2	sakit	298.43	wisata	227.5
harga	256.82	kanker	271.04	gili	227.3
banjir	233.91	serang	249.51	gunung	220.5
tol	225.15	seks	247.44	unjung	216.6

TABLE XX. FEATURE BIGRAM DENGAN STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL GEMPA.

Ekonomi Gempa		Kesehatan Gempa		Pariwisata Gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.18	korban gempa	123.9	gili trawangan	116.1
rp triliun	88.87	rumah sakit	100.2	gunung rinjani	92.9
sri mulyani	88.327	menteri sehat	86.39	gili air	90.98
rp juta	77.739	lombok utara	84.71	menteri pariwisata	73.52
kantor cabang	74.912	rsud tanjung	76.6	jalur daki	65.41

TABLE XXI. FEATURE BIGRAM DENGAN STEMMING DAN STOPWORDS REMOVAL PADA ARTIKEL NON-GEMPA.

Ekonomi Non-gempa		Kesehatan Non-gempa		Pariwisata Non-gempa	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
jalan tol	170.61	serang jantung	221.91	kompas com	108.7
rp triliun	146.24	nyeri dada	90	gili trawangan	106.8
harga rokok	111.28	getah bening	81.83	gunung rinjani	105.1
rp ribu	92.328	kelenjar getah	79.683	gili air	90.98
dollar as	82.62	paru paru	75.427	daki gunung	78.43

Pada tabel yang telah disajikan, dapat dilihat beberapa *feature* yang telah dihilangkan dari pengujian tanpa penggunaan *stemming* maupun *stopwords removal*. *Feature* yang dihilangkan adalah *stopwords* dan *feature-feature* berimbuhan yang digabung menjadi satu *feature* yang berupa kata dasar. Untuk *feature unigram*, tidak terdapat perubahan bobot *feature* yang cukup besar untuk tiap kategori. Sehingga penerapan *stemming* bersamaan dengan *stopwords removal* tidak memiliki pengaruh yang cukup besar terhadap akurasi dari model.

Akan tetapi, penggunaan *stemming* bersamaan dengan *stopwords removal* pada *feature bigram* dapat meningkatkan akurasi secara signifikan dibandingkan dengan penerepan *stopwords removal* tanpa *stemming*. Hal ini dikarenakan suatu *feature* yang terdiri dari 2 kata yang sebenarnya tidak bersebelahan dikarenakan penghilangan *stopwords* memiliki bobot yang lebih tinggi ketika *feature* yang sama namun berbeda imbuhan dijadikan satu *feature* melalui tahap *stemming*. Proses ini tidak menghilangkan pelanggaran terhadap konsep *bigram*, tetapi dapat meningkatkan akurasi dari model dengan cukup signifikan.

Hasil yang didapatkan dari skema pengujian ini melalui 5-*fold cross validation* dapat dilihat pada Tabel XXI.

TABLE XXII. HASIL PENGUJIAN DENGAN STEMMING DAN STOPWORDS REMOVAL.

Jenis Feature	Precision	Recall	F-Measure	Standar Deviasi
Unigram	94.22%	95.15%	94.54%	1.66%
Bigram	83.27%	88.00%	83.73%	2.62%
Unigram dan *Bigram*	95.12%	95.47%	95.20%	1.58%

Pada Gambar 2, ukuran *vocabulary* yang didapatkan dari tiap jenis *feature* berbeda-beda satu sama lain. *Feature unigram* memiliki ukuran *vocabulary* jauh lebih kecil dibandingkan dengan *feature bigram*. Hal ini disebabkan oleh pasangan kata yang bervariasi dalam tiap dokumen. Ukuran *vocabulary* dari penggunaan *feature unigram* bersamaan dengan *feature bigram* sendiri merupakan jumlah dari ukuran 2 jenis *feature* tersebut. Selain itu, dapat dilihat juga penggunaan *stemming* dan *stopwords removal* juga dapat mengurangi ukuran dari *vocabulary*.

Pada Gambar 3, dapat dilihat bahwa penggunaan *stemming* mengurangi nilai *f-measure* dari pengujian menggunakan *feature unigram*. Hal ini juga berlaku pada pengujian yang menggunakan *feature unigram* sekaligus *bigram* dan menyertakan *stopwords*. Akan tetapi, nilai *f-measure* yang didapatkan pada pengujian dengan menggunakan *stemming* justru mengalami peningkatan pada *feature bigram* Nilai *f-measure* dari *feature unigram* sekaligus *bigram* yang tidak menyertakan *stopwords* juga mengalami peningkatan ketika dilakukan *stemming*.

Pada Gambar 3 juga dapat dilihat bahwa penghilangan *stopwords* dapat meningkatkan nilai *f-measure* untuk pengujian dengan *feature unigram* dan juga *feature unigram* sekaligus *bigram*. Akan tetapi, *stopwords removal* pada pengujian dengan *feature bigram* justru dapat mengurangi nilai *f-measure* secara signifikan.

5. PENUTUP

Berdasarkan hasil penelitian yang telah didapatkan, dapat disimpulkan bahwa:

1. Metode *multimonial naïve Bayes* dapat diterapkan pada kasus klasifikasi artikel dimana pada penilitian ini didapatkan akurasi yang cukup tinggi yaitu dengan nilai *f-measure* mencapai 95.20% dan standar deviasi sebesar 1.58% melalui pengujian dengan 5 perulangan 5-*fold cross validation*.
2. Penggunaan *stemming* mengurangi akurasi pada pengujian dengan *feature unigram* karena terdapat kata berimbuhan yang merupakan *feature* unik dari suatu kategori.
3. Penggunaan *stemming* pada pengujian dengan *feature bigram* dapat meningkatkan akurasi dari percobaan. Hal ini merupakan *general phenomenon* pada *text behavior* dimana suatu pasangan kata memiliki relevansi yang tinggi terhadap dokumen terlepas dari imbuhan-nya.
4. Penggunaan *stopwords removal* pada *feature bigram* dapat menurunkan akurasi dari model secara signifikan karena hilangnya suatu kata dapat menghilangkan pasangan kata yang berkontribusi positif terhadap dokumen.
5. Pada penilitian ini, hasil pengujian terbaik didapatkan dengan menggunakan *feature unigram* sekaligus *bigram* dan dengan melewati tahap *stemming* dan *stopwords removal*. Sedangkan hasil pengujian dengan nilai *f-measure* terendah didapatkan dari pengujian dengan *feature bigram* yang tidak melewati tahap *stemming* tetapi melewati tahap *stopwords removal*.

Kemudian terdapat beberapa catatan saran untuk dapat diperbaiki serta dikembangkan pada penelitian serupa selanjutnya yaitu melakukan *feature selection* pada model dan menambah *feature trigram* dalam skema pengujian.

UCAPAN TERIMA KASIH

Ucapan terima kasih diberikan kepada pakar ilmu komunikasi, Shinta Desiyana Fajarica, S.IP., M.Si., yang telah mengarahkan dalam pengumpulan data.

DAFTAR PUSTAKA

[1] N. Aziz, “Mengapa gempa terus terjadi di Indonesia?,” BBC, 7 Agustus 2018. [Online]. Available: https://www.bbc.com. [Diakses 4 Desember 2018].

[2] F. Handayani and F. S. Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *Jurnal Teknik Elektro,* vol. 7, no. 1, pp. 19-24, 2015.

[3] T. Jo, "Text Mining : Concepts, Implementation, and Big Data Challenge", vol. 45, Cham: Springer International Publishing AG, 2019.

[4] C. D and P. R. H. S. Manning, Introduction to Information Retrieval, New York: Cambridge University Press, 2008.

[5] M. S. Islam, M. I. Fauzan P. P. N. and M. T. Pratama, "Penggunaan Naive Bayes Classifier untuk Pengelompokan Pesan Pada Ruang Percakapan Maya dalam Lingkungan Kemahasiswaan," *Jurnal Computech & Bisnis,* vol. 11, no. 2, pp. 87-97, 2012.

[6] R. N. Devita, H. W. Herwanto and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa Indonesia," *Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. 5, no. 4, pp. 427-434, 2018.

[7] I. B. G. W. Putra, M. Sudarma and I. N. S. Kumara, "Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naïve Bayes Classifier," *Teknologi Elektro,* vol. 15, no. 2, pp. 81-86, 2016.

[8] A. P. Wijaya and H. A. Santoso, "Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government," *Journal of Applied Intelligent System,* vol. 1, no. 1, pp. 48-55, 2016.

[9] M. A. Ulfa, B. Irmawati and A. Y. Husodo, "Twitter Sentiment Analysis using Naive Bayes Classifier with Mutual Information Feature Selection," *J-COSINE,* vol. 2, no. 2, pp. 106-111, 2018.

[10] A. M. Kibriya, E. Frank, B. Pfahringer and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited," in *Australian Joint Conference on Artificial Intelligence*, Cairns, 2004.

[11] A. Rahman, W. and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART: Jurnal Ilmiah Teknologi dan Informasi,* vol. 6, no. 1, pp. 32-38, 2017.

[12] F. Tempola, M. Muhammad and A. Khairan, "Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation," *Jurnal Teknologi Informasi dan Ilmu Komputer,* vol. 5, no. 5, pp. 577-584, 2018.

[13] M. S. H. Simarangkir, "Studi Perbandingan Algoritma-Algoritma Stemming untuk Dokumen Teks Bahasa Indonesia," *Jurnal Inkofar,* vol. 1, no. 1, pp. 40-46, 2017.

[14] H. R. Pramudita, "Penerapan Algoritma Stemming Nazief & Adriani dan Similarity pada Penerimaan Judul Thesis," *Jurnal Ilmiah DASI,* vol. 15, no. 4, pp. 15-19, 2014.