KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA MENGGUNAKAN MULTINOMIAL NAÏVE BAYES

Tugas Akhir
Untuk memenuhi sebagian persyaratan
mencapai derajat Sarjana S-1 Program Studi Teknik Informatika

Oleh :
Alif Sabrani
F1D 015 006

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNIK
UNIVERSITAS MATARAM
2020

TUGAS AKHIR

KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA MENGGUNAKAN MULTINOMIAL NAÏVE BAYES

Oleh :

Alif Sabrani

F1D015006

Telah diperiksa dan disetujui oleh:

1. Pembimbing Utama

Dr.Eng. I Gede Putu Wirarama W.W., S.T., M.T. Tanggal: 6 Maret 2020

NIP: 19840919 201803 1 001

2. Pembimbing Pendamping

Fitri Bimantoro, S.T., M.Kom. Tanggal: 6 Maret 2020

NIP. 19860622 201504 1 002

Mengetahui,

Ketua Program Studi Teknik Informatika

Fakultas Teknik

Universitas Mataram

Prof. D.Eng. I Gede Pasek Suta Wijaya S.T., M.T.

NIP. 19731130 200003 1 001

TUGAS AKHIR

KLASIFIKASI ARTIKEL ONLINE TENTANG GEMPA DI INDONESIA MENGGUNAKAN MULTINOMIAL NAÏVE BAYES

Oleh :

Alif Sabrani

F1D015006

Telah dipertahankan di depan Dewan Penguji
Pada Tanggal 3 Maret 2020
dan dinyatakan telah memenuhi syarat mencapai derajat S-1
Program Studi Teknik Informatika
Susunan Tim Penguji

1. Penguji I

Dr. Eng. Budi Irmawati S.Kom., M.T. Tanggal : 6 Maret 2020
NIP. 19721019 199903 2 001

2. Penguji II

Prof. D.Eng. I Gede Pasek Suta Wijaya S.T., M.T. Tanggal : 6 Maret 2020
NIP. 19731130 200003 1 001

3. Penguji III

Ramaditia Dwiyansaputra S.T., M.Eng Tanggal : 6 Maret 2020
NIP. -

Mataram, Maret 2020
Dekan Fakultas Teknik
Universitas Mataram

Akmaluddin, ST, M.Sc.(Eng.), Ph.D
NIP : 19681231 199412 1 001

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya yang bertanda tangan di bawah ini bahwa dalam Skripsi ini tidak terdapat karya yang pernah di ajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Mataram, Maret 2020

Alif Sabrani

PRAKATA

Puji syukur penulis hantarkan ke hadirat Allah SWT, Tuhan semesta alam, karena berkat limpahan rahmat dan karunia-Nya lah penulis dapat menyelesaikan Laporan Tugas Akhir dengan judul "Klasifikasi Artikel *Online* tentang Gempa di Indonesia menggunakan *Multinomial Naïve Bayes*" ini sesuai dengan ketentuan yang telah ditetapkan.

Dalam penulisan Tugas Akhir ini tentunya tidak luput dari kekurangan, baik aspek kualitas maupun aspek kuantitas dari materi penelitian yang disajikan. Semua ini didasarkan dari keterbatasan yang dimiliki penulis. Penulis menyadari bahwa Tugas Akhir ini masih jauh dari kata sempurna sehingga penulis membutuhkan kritik dan saran yang membangun untuk kemajuan teknologi di masa yang akan datang.

Akhir kata semoga tidaklah terlampau berlebihan, bila penulis berharap agar karya ini dapat bermanfaat bagi pembaca.

Mataram, Maret 2020

Alif Sabrani

UCAPAN TERIMA KASIH

Tugas Akhir ini dapat diselesaikan tidak terlepas dari bantuan dari berbagai pihak dan berkat bimbingan, dukungan ilmiah maupun materi. Oleh karena itu penulis menyampaikan ucapan terima kasih kepada :

1. Allah SWT atas segala kesempatan, kesehatan dan anugerah yang telah diberikan selama pembuatan Tugas Akhir ini.
2. Kedua orang tua yang telah memberikan semangat untuk menyelesaikan tugas akhir ini.
3. Bapak Dr. Eng. I Gede Putu Wirarama W. W.,S.T., M.T. dan Bapak Fitri Bimantoro, S.T., M.Kom. selaku dosen pembimbing yang telah memberikan bimbingan dan arahan kepada penulis selama penyusunan Tugas Akhir ini, sehingga dapat terselesaikan dengan baik.
4. Dosen penguji yang telah memberikan kritik dan saran yang bersifat membangun dalam penyelesaian Tugas Akhir ini.
5. Shinta Desiyana Fajarica, S.IP., M.Si selaku pakar yang telah membimbing penulis selama mengumpulkan data.
6. Semua pihak yang tidak dapat penulis sebutkan satu persatu, yang telah memberikan bimbingan kepada penulis dalam menyelesaikan Tugas Akhir ini.

Semoga Tuhan Yang Maha Esa memberikan imbalan yang setimpal atas bantuan yang diberikan kepada penulis.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Indonesia merupakan negara yang rawan gempa bumi. Hal ini menyebabkan banyaknya pemberitaan tentang gempa bumi oleh berbagai media massa. Salah satu cara penyampaian informasi yang cukup populer adalah melalui artikel *online*. Artikel *online* tentang gempa bumi dapat dikelompokan ke dalam kategori ekonomi, kesehatan, dan pariwisata. *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*.

Kata kunci: Gempa, Artikel *Online*, Klasifikasi Teks, TF-IDF, *Preprocessing*, Multinomial *Naïve Bayes*

BAB I
PENDAHULUAN

1.1 Latar Belakang

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].

Pada tahun 2018, di Indonesia telah terjadi 2 gempa bumi dengan dampak yang cukup besar. Salah satunya adalah gempa bumi yang terjadi di Pulau Lombok, Nusa Tenggara Barat. Gempa Lombok terjadi secara beruntun selama bulan Agustus 2018. Peristiwa ini telah mengakibatkan kerusakan yang cukup parah dimana lebih dari 160.000 bangunan mengalami kerusakan, 1.584 orang mengalami luka-luka, dan jumlah korban jiwa mencapai 564 orang [2]. Gempa Lombok kemudian disusul oleh gempa yang mengguncang Kota Palu, Kabupaten Donggala, dan Kabupaten Sigi. Tercatat 2.113 orang meninggal dunia dan 4.612 orang mengalami luka berat akibat gempa yang mengguncang sejumlah daerah di Sulawesi Tengah ini [3].

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 yang terkait dengan gempa di bidang tertentu.

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 klasifikasi sebuah dokumen secara otomatis oleh komputer dikenal dengan istilah *text classification* atau klasifikasi teks [4]. 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 [5]. Salah satu model dari *Naïve Bayes* yang sering digunakan dalam klasifikasi teks adalah *multinomial Naïve Bayes* [6].

Penulis mengajukan sebuah penelitian untuk merancang sebuah model untuk mengklasifikasikan artikel *online* pasca bencana gempa bumi di Indonesia menggunakan *multinomial Naïve Bayes*. Keluaran dari penilitian ini diharapkan dapat menghasilkan model yang mampu mengklasifikasikan artikel *online* pasca gempa sehingga proses klasifikasi dapat dilakukan dengan lebih cepat.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dijelaskan sebelumnya, rumusan masalah yang ingin di jawab pada penelitian ini adalah sebagai berikut.

7. Bagaimana membangun sebuah model klasifikasi artikel *online* pasca gempa di Indonesia ke dalam kategori ekonomi, pariwisata dan kesehatan menggunakan *multinomial Naïve Bayes*?
8. Bagaimana akurasi dari metode *multinomial Naïve Bayes* dalam mengelompokkan artikel *online* pasca gempa di Indonesia ke dalam kategori ekonomi, pariwisata, dan kesehatan?

1.3 Batasan Masalah

Penelitian ini memiliki batas masalah agar pengerjaan dapat dilakukan dengan lebih fokus. Adapun batasan masalah yang ditetapkan adalah sebagai berikut.

1. Klasifikasi artikel dibagi menjadi 6 kelas yaitu ekonomi, pariwisata dan kesehatan untuk masing-masing kejadian gempa dan non-gempa.
2. Pengelompokan didasarkan pada isi dari artikel.
3. Data yang digunakan dalam penelitian adalah artikel *online* pasca gempa di Indonesia serta artikel ekonomi, kesehatan, dan pariwisata yang tidak berhubungan dengan gempa yang didapatkan dari *www.kompas.com*, *www.detik.com*, *www.okezone.com*, dan *www.liputan6.com*.

1.4 Tujuan

Tujuan yang diharapkan dapat dicapai melalui penelitian ini adalah sebagai berikut.

1. Menghasilkan sebuah model klasifikasi artikel *online* pasca gempa di Indonesia ke dalam kategori ekonomi, pariwisata dan kesehatan menggunakan *multinomial Naïve Bayes*.
2. Mengetahui akurasi dari metode *multinomial Naïve Bayes* dalam mengelompokkan artikel *online* pasca gempa ke dalam kategori ekonomi, pariwisata, dan kesehatan.

1.5 Manfaat

Manfaat dari penelitian ini secara umum dapat diperoleh oleh dua subjek antara lain.

1. Bagi Penulis
   a. Dapat menerapkan ilmu yang didapatkan selama perkuliahan terutama yang berhubungan dengan kecerdasan buatan.
   b. Dapat menambah pengetahuan di bidang *machine learning* khususnya *text classification*.
2. Bagi Pembaca
   a. Dapat mengetahui cara untuk membangun sebuah model yang dapat mengklasifikasikan artikel ke dalam kategori ekonomi, pariwisata, dan kesehatan.
   b. Dapat dijadikan rujukan dalam pengembangan model klasifikasi teks agar mendapat akurasi yang lebih baik.
   c. Dapat mengetahui dampak dari gempa bumi di segi ekonomi, kesehatan, dan pariwisata secara kuantitatif.

1.6 Sistematika Penulisan

Sistematika penulisan dari penelitian ini disajikan dalam beberapa bab antara lain sebagai berikut.

1. Bab I Pendahuluan

   Bab ini menjelaskan dasar-dasar dari penulisan laporan tugas akhir, yang terdiri dari latar belakang, rumusan masalah, batasan masalah, tujuan, serta sistematika penulisan laporan tugas akhir.

2. Bab II Tinjauan Pustaka dan Dasar Teori

   Bab ini membahas tentang penelitian-penelitian terdahulu yang mengimplementasikan metode *multinomial Naïve Bayes* serta teori-teori sebagai referensi penulis ketika melakukan penelitian.

3. Bab III Metodologi Penelitian

   Bab ini membahas tentang metodologi yang digunakan untuk membangun model klasifikasi artikel ke dalam kategori Ekonomi Gempa, Ekonomi Non-gempa, Kesehatan Gempa, Kesehatan Non-gempa, Pariwisata Gempa dan Pariwisata Non-gempa.

4. Bab IV Hasil dan Pembahasan

   Memuat tentang hasil dan pembahasan yang diperoleh berdasarkan hasil pengukuran dan perhitungan.

5. Bab V Kesimpulan dan Saran

   Bab ini memaparkan kesimpulan yang telah didapatkan dari hasil penelitian tentang klasifikasi teks melalui *platform web* menggunakan *python* dengan metode *multinomial Naïve Bayes* dan saran-saran yang diberikan untuk menyempurnakan hasil penelitian kedepannya.

BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Telah dilakukan penelitian untuk mengklasifikasikan pengaduan dan pelaporan masyarakat menggunakan metode *multinomial Naïve Bayes* [4]. *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 mengelompokkan pesan dalam ruang percakapan maya dengan metode *multinomial Naïve Bayes* [7]. 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 [8]. 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 [9]. 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*) [10]. 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* [11]. 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 [12]. *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.

2.2 Dasar Teori

2.2.1 Teks

Teks merupakan data tidak terstruktur yang disusun oleh kumpulan kata. Kata – kata perlu memiliki arti tersendiri serta disusun berdasarkan aturan tertentu untuk dapat membentuk sebuah teks. Aturan yang digunakan dalam penyusunan kata dalam teks ini disebut *grammar* [5].

2.2.2 *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 [10]. Bagian – bagian dari *text mining* meliputi *classification* (klasifikasi), *clustering*, dan *association* [5].

2.2.3 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*). Klasifikasi teks melibatkan dua proses utama, yakni pertama ekstraksi fitur yang menjadi kata kunci yang efektif dalam tahap pelatihan atau *training* dan kemudian proses kedua yakni klasifikasi dokumen [4]. Langkah awal yang perlu dilakukan dalam klasifikasi teks adalah dengan memisahkan data menjadi *training* dan data uji. Data *training* kemudian dikelompokan / diberi label yang sesuai dengan isi teks. Lalu data *training* diproses dengan algoritma tertentu untuk menghasilkan model yang digunakan dalam proses klasifikasi. Model inilah yang akan digunakan untuk memprediksikan kelas dari data uji [5].

2.2.4 *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* [5].

a. *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* [5].

b. *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 [5].

c. *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 [5]. 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.

2.2.5 *Feature weighting* dengan TF-IDF

Term weighting merupakan suatu proses untuk menghitung serta memberi bobot pada suatu kata sebagai derajat kepentingannya. *Term frequency* dan pembobotan TF-IDF merupakan metode yang sering digunakan dalam pembobotan kata. Bobot kata dapat digunakan untuk memberi nilai matematis pada suatu kata agar dapat diproses oleh komputer [5].

Term frequency adalah jumlah kemunculan suatu kata dalam dokumen. Terdapat 2 jenis *term frequency* yang dapat digunakan untuk pemberian bobot kata, yaitu *absolute term frequency* dan *relative term frequency*. *Absolute term frequency* merupakan kemunculan kata dalam dokumen, sedangkan *relative term frequency* adalah rasio kemunculan kata dalam dokumen terhadap jumlah seluruh kata dalam dokumen [5].

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) [8]. Karena dilakukan pula perhitungan IDF, maka metode TF-IDF membutuhkan referensi dari seluruh dokumen (*corpus*) [5]. Perhitungan TF dan IDF dapat dilakukan dengan Persamaan (2-1) dan (2-2) [6].

$$TF(d,t) = f(d,t) \tag{2-1}$$

$$IDF(t) = log\left(\frac{N_d}{df(t)}\right) \tag{2-2}$$

dimana :

$TF(d,t)$: *Term frequency*

$f(d,t)$: Frekuensi kemunculan *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 (2-3).

$$TF - IDF = TF(d, t) \, . IDF(t) \tag{2-3}$$

2.2.6 *Multinomial Naïve Bayes*

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 [5].

Salah satu model dari *Naïve Bayes* yang sering digunakan dalam klasifikasi teks adalah *multinomial Naïve Bayes* [6]. *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 (2-4) [6].

$$P(c|d) \propto P(c) \prod_{k=1}^{n} P(t_k|c) \tag{2-4}$$

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 (2-5).

$$c_{\text{map}} = \arg\max_{c \in C} \hat{P}(c|d) = \arg\max_{c \in C} \hat{P}(c) \prod_{k=1}^{n} \hat{P}(t_k|c) \tag{2-5}$$

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* [6].

Pada Persamaan (2-5), 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 (2-5) yang menggunakan logaritma dari probabilitas dapat dinyatakan dalam Persamaan (2-6) [6].

$$c_{\text{map}} = \arg\underset{c \in C}{max} \left[\log \hat{P}(c) + \sum_{1 \leq k \leq n} \log \hat{P}(t_k|c)\right] \quad (2\text{-}6)$$

$\hat{P}(c)$ dan $\hat{P}(t_k|c)$ didapatkan dengan menghitung *maximum likelihood*, yang merupakan frekuensi relatif dari parameter. Untuk *prior*, dapat digunakan Persamaan (2-7).

$$\hat{P}(c) = \frac{N_c}{N} \quad (2\text{-}7)$$

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 (2-8).

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \quad (2\text{-}8)$$

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 (2-8) berubah menjadi Persamaan (2-9).

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'})+B'} \quad (2\text{-}9)$$

dimana :

B : Jumlah seluruh *term* pada *vocabulary*

Sedangkan untuk rumus *multinomial naïve Bayes* dengan menggunakan pembobotan TF-IDF dapat dilihat pada persamaan (2-10) [13].

$$\hat{P}(t|c) = \frac{W_{ct}+1}{(\sum_{w\prime\in V} W_{ct\prime})+B\prime} \quad (2\text{-}10)$$

dimana :

W_{ct} : Bobot TF-IDF *term t* pada dokumen dengan kategori *c*

$\sum_{w\prime\in V} W_{ct\prime}$: Jumlah bobot TF-IDF seluruh *term* pada kelas *c*

B' : Jumlah IDF seluruh *term* pada *vocabulary*

2.2.7 *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 [14].

2.2.8 *Recall, precision,* dan *f-measure*

Salah satu teknik yang digunakan untuk melakukan evaluasi dalam klasifikasi adalah dengan menghitung *recall*, *precision*, dan *f-measure*. Teknik ini menggunakan *confusion matrix* sebagai acuan perhitungan. Tabel *confusion matrix* untuk data dengan lebih dari 2 kelas dapat dilihat pada Tabel 2.1 [13].

Tabel 2.1 Tabel *confusion matrix*.

Realita	Sistem				
	Kelas-1	Kelas-2	....	Kelas-n	Total
Kelas-1	*True Positive*	*Error*	....	*Error*	Total Kelas-1
Kelas-2	*Error*	*True Positive*	...	*Error*	Total Kelas-2
...	*Error*	*Error*	...	*Error*	...
Kelas-n	*Error*	*Error*	...	*True Positive*	Total Kelas-n
	Prediksi Kelas-1	Prediksi Kelas-2	...	Prediksi Kelas-n	

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 (2-11) [13].

$$Recall_c = \frac{TP(Kelas-c)}{Total(Kelas-c)} \quad (2\text{-}11)$$

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 (2-12) [13].

$$Precision_c = \frac{TP(Kelas-c)}{Prediksi(Kelas-c)} \quad (2\text{-}12)$$

Sedangkan *f-measure* merupakan nilai yang mewakili seluruh kinerja sistem yang merupakan penggabungan nilai *recall* dan *precision*. *F-measure* dapat dihitung menggunakan Persamaan (2-13) [4].

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

BAB III
METODOLOGI PENELITIAN

3.1 Alat dan Bahan Penelitian

3.1.1 Alat penelitian

Dalam penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes*, digunakan beberapa alat yang terdiri perangkat keras dan perangkat lunak. Alat – alat tersebut adalah sebagai berikut.

a. Perangkat Keras

Perangkat keras yang digunakan dalam penelitian adalah satu unit laptop dengan spesifikiasi sebagai berikut.

1. Processor Intel® Core™ i3-5005U
2. Memori RAM 4096MB

b. Perangkat Lunak

Perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Sistem Operasi *Windows* 10 *Pro*
2. *Visual Studio Code*
3. Bahasa Pemrograman *Python* versi 3.7.1

3.1.2 Bahan penelitian

Bahan penelitian yang digunakan dalam penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes* ini adalah artikel online tentang gempa di Indonesia serta artikel *online* yang tidak berhubungan dengan gempa yang dikumpulkan dari situs *www.kompas.com*, *www.detik.com*, *www.liputan6.com*, dan *www.okezone.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.

3.2 Studi Literatur

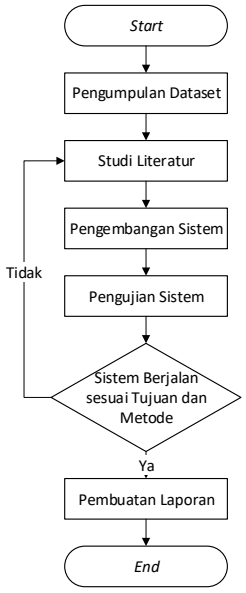
Studi literatur yang dilakukan untuk mendukung penelitian adalah mempelajari buku elektronik, jurnal – jurnal penelitian, serta berbagai sumber lainnya yang berkaitan

dengan topik penelitian, yaitu klasifikasi dokumen. Lebih spesifik, materi yang dipelajari adalah *text preprocessing*, *text classification* serta pemanfaatan metode probabilistik *naïve Bayes classifier* dalam melakukan klasifikasi dokumen. Jurnal – jurnal yang dipelajari membahas berbagai studi kasus tentang klasifikasi dokumen dengan metode *naïve Bayes classifier*.

3.3 Rancangan Penelitian

Penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes* dilakukan dalam beberapa tahapan. Tahap pertama yang dilakukan pengumpulan *dataset* dari *website www.kompas.com*, *www.detik.com*, *www.liputan6.com*, dan *www.okezone.com*. Kemudian dilakukan studi literatur untuk mendapatkan pengetahuan serta gambaran akan penelitian yang dilakukan. Literatur yang dipelajari berupa jurnal penelitian serta buku yang membahas tentang klasifikasi teks menggunakan metode *naïve Bayes classifier*. Kemudian dilakukan pengembangan sistem yang dilandaskan pada literatur yang telah dipelajari sebelumnya. Setelah sistem telah selesai dikembangkan, dilakukan pengujian pada sistem. Apabila sistem berjalan sesuai dengan tujuan dan metode yang digunakan, penelitian dilanjutkan dengan pembuatan laporan. Apabila sistem tidak berjalan sesuai dengan yang diharapkan, maka akan dilakukan kembali studi literatur untuk memperbaiki kesalahan – kesalahan yang menyebabkan kurang optimalnya sistem yang telah dibangun. Target akurasi minimal yang diharapkan pada penelitian ini adalah sekurang-kurangnya 50%.

Diagram alir penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 Diagram alir rancangan penelitian.

3.4 Kebutuhan Sistem

Dalam penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes*, analisis kebutuhan sistem dibagi menjadi 3 jenis yaitu analisis pengguna, serta analisis perangkat keras dan perangkat lunak yang digunakan dalam penelitian.

a. Analisis pengguna

Pengguna dari sistem ini adalah masyarakat umum yang ingin mengetahui kategori dari artikel *online* yang memuat berita tentang gempa di Indonesia. Pengguna dapat mengetahui bagaimana dampak dari gempa di Indonesia dari segi ekonomi, kesehatan, serta pariwisata.

b. Analisis perangkat keras

Perangkat keras yang digunakan dalam pembangunan sistem, pelatihan data, serta pengujian sistem merupakan elemen penting dalam penelitian ini. Perangkat keras yang mumpuni dapat membantu mempercepat proses – proses yang dilakukan seperti pelatihan data yang membutuhkan sumber daya cukup tinggi. Perangkat keras yang digunakan dalam penelitian ini memiliki spesifikasi seperti yang terdapat pada Tabel 3.1.

Tabel 3.1 Kebutuhan perangkat keras.

No.	Nama Perangkat	Spesifikasi
1	*Processor*	Intel® Core™ i3-5005U
2	Memori	Memori RAM 4096MB

c. Analisis perangkat lunak

Selain perangkat keras, perangkat lunak juga memiliki peranan penting dalam proses pengembangan sistem. Penggunaan perangkat lunak yang tepat dapat membantu mempercepat proses penelitian. Perangkat lunak yang digunakan dalam penelitian ini dapat dilihat pada Tabel 3.2.

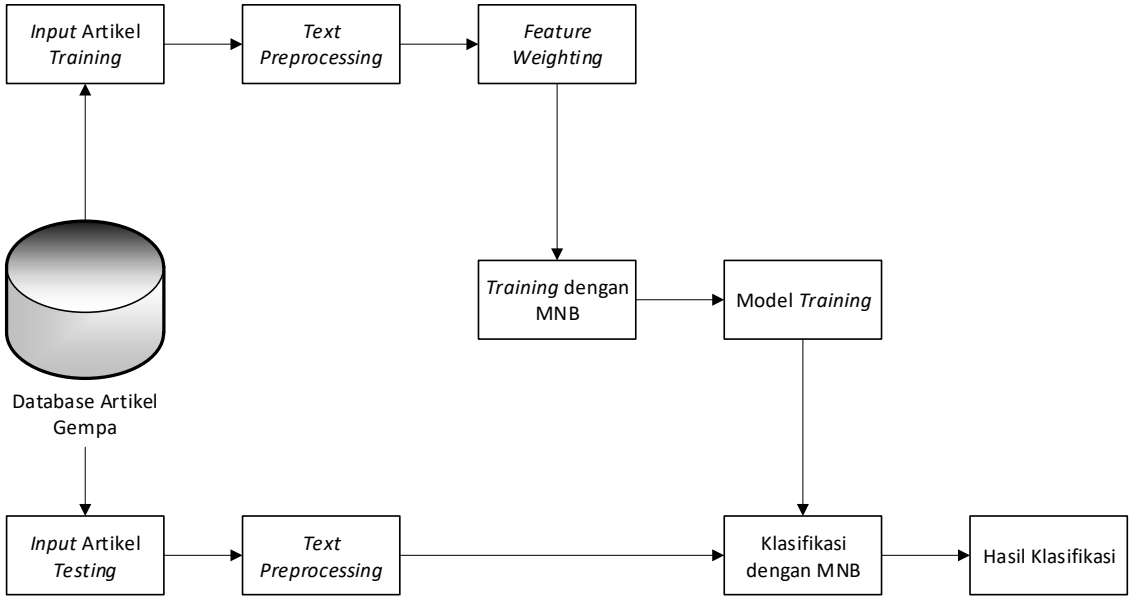
Tabel 3.2 Kebutuhan perangkat lunak.

No.	Nama Perangkat	Spesifikasi
1	Sistem Operasi	Windows 10
2	*Text Editor*	*Visual Studio Code*
3	*Microsoft Office*	*Office* 2016

4	Bahasa pemrograman *python*	*Python* 3.7.1
5	*Library* NLTK	*Python* NLTK 3.4
6	*Library scikit learn*	*Python scikit learn* 0.21.2
7	*Library* Sastrawi	*Python* Sastrawi
8	*Library* BeautifulSoup4	*Python* BeautifulSoup4
9	*Web Browser*	*Google Chrome*

3.5 Rancangan Sistem

Rancangan dari sistem klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes* terdiri dari beberapa tahapan, yang dapat dilihat pada Gambar 3.2.



Gambar 3.2 Rancangan sistem klasifikasi artikel gempa.

3.5.1 *Input* artikel *training* dan *testing*

Pada tahap ini, artikel yang telah dikumpulkan dari situs *www.kompas.com*, *www.detik.com*, *www.liputan6.com*, *www.cnnindonesia.com*, dan *www.okezone.com* 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.

a. *Input* 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*.

Contoh artikel yang digunakan sebagai artikel *training* untuk tiap kategori dapat dilihat pada Tabel 3.3.

Tabel 3.3 Contoh artikel *training.*

Kategori Dokumen	Isi Dokumen
Kesehatan Gempa	Viral di media sosial postingan netizen yang tinggal di Lombok. Dalam postingan tersebut, ia mengaku mengalami halusinasi di mana tanah terasa bergoyang padahal tidak terjadi gempa. Psikolog dari Personal Growth, Veronica Adesla, mengatakan trauma psikologis memang rentan terjadi pada korban bencana alam seperti di Lombok. Dalam pedoman Diagnostic and Statistical Manual of Mental Disorders Fifth Edition (DSM-5), disebutkan bahwa salah satu ciri trauma psikologis adalah merasakan kejadian yang membuat trauma.
Ekonomi Gempa	Gempa mengguncang Lombok, NTB sejak akhir Agustus hingga awal September 2018. Menurut catatan pemerintah nilai kerusakan akibat gempa yang berlangsung dari 29 Agustus hingga 9 September 2018 mencapai Rp 10 triliun Kepala Badan Nasional Penanggulangan Bencana (BNPB) Willem Rampangilei mengatakan total kerusakan per 5 September 2018 tersebut diakumulasi dari tujuh wilayah, yakni Lombok Utara, Lombok Timur, Sumbawa Barat, Sumbawa, Mataram, Lombok Barat, Lombok Tengah.

Pariwisata Gempa	Kawasan wisata Pantai Senggigi di Kabupaten Lombok Barat lengang pasca gempa 7,0 Skala Richter melanda Lombok. Biasanya kawasan ini ramai. Senggigi yang merupakan pusat dan destinasi andalan penyumbang PAD pariwisata terbesar di Lombok Barat ini terpantau lengang sejak Senin pagi (6/8/2018). Sebabnya, hampir semua wisatawan mancanegara telah dievakuasi untuk meninggalkan wilayah Senggigi akibat adanya gempa bumi dahsyat yang berpusat di Pulau Lombok.
Kesehatan Non-gempa	Hingga saat ini belum ada kasus positif virus corona 2019-nCoV (novel coronavirus) di Indonesia. Khususnya di Jakarta, ada 3 RS yang siap menangani jika ada kasus yang terkonfirmasi. Tiga rumah sakit di Jakarta yang menjadi pusat rujukan nasional adalah Rumah Sakit Pusat Angkatan Darat (RSPAD) Gatot Soebroto, RS Pusat Infeksi Sulianti Saroso, dan RS Pusat Persahabatan. Ketiganya termasuk dalam 100 RS yang disiagakan untuk menangani virus corona.
Ekonomi Non-gempa	Bank Indonesia (BI) meyakini pertumbuhan ekonomi Indonesia sepanjang 2019 tumbuh di level 5,1%. Hal itu ditopang dari kecenderungan kondisi siklus ekonomi yang kian membaik. Gubernur BI Perry Warjiyo mengatakan, pertumbuhan ekonomi Indonesia tetap berdaya tahan ditopang perbaikan ekspor dan konsumsi rumah tangga yang tetap baik.
Pariwisata Non-gempa	Asosiasi Hotel Mataram bekerja sama dengan organisasi kerja sama internasional milik pemerintah Jerman (GIZ) untuk membangun pariwisata berkelanjutan di Pulau Lombok, Nusa Tenggara Barat. Penandatanganan nota kesepahaman (MoU) antara Ketua Asosiasi Hotel Mataram (AHM) Reza Bouvierd dengan Team Leader Tourism and Investment GIZ, Oliver Oehms, berlangsung di Kota Mataram, Kamis (15/10/2015) lalu.

b. *Input* 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*.

3.5.2 *Text preprocessing*

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

a. *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. Contoh artikel yang telah melewati proses *tokenization* dengan *feature unigram* dapat dilihat pada Tabel 3.4.

Tabel 3.4 Contoh artikel yang telah melewati proses *tokenization* dengan *feature unigram*.

Kategori Dokumen	*Token* / *term*
Kesehatan Gempa	viral, di, media, sosial, postingan, netizen, yang, tinggal, di, lombok, dalam, postingan, tersebut, ia, mengaku, mengalami, halusinasi, di, mana, tanah, terasa, bergoyang, padahal, tidak, terjadi, gempa, psikolog, dari, personal, growth, veronica, adesla, mengatakan, trauma, psikologis, memang, rentan, terjadi, pada, korban, bencana, alam, seperti, di, lombok, dalam, pedoman, diagnostic, and, statistical, manual, of, mental, disorders, fifth, edition, dsm, disebutkan, bahwa, salah, satu, ciri, trauma, psikologis, adalah, merasakan, kejadian, yang, membuat, trauma
Ekonomi Gempa	gempa, mengguncang, lombok, ntb, sejak, akhir, agustus, hingga, awal, september, menurut, catatan, pemerintah, nilai, kerusakan, akibat, gempa, yang, berlangsung, dari, agustus, hingga, september, mencapai, rp, triliun, kepala, badan, nasional, penanggulangan,

	bencana, bnpb, willem, rampangilei, mengatakan, total, kerusakan, per, september, tersebut, diakumulasi, dari, tujuh, wilayah, yakni, lombok, utara, lombok, timur, sumbawa, barat, sumbawa, mataram, lombok, barat, lombok, tengah
Pariwisata Gempa	kawasan, wisata, pantai, senggigi, di, kabupaten, lombok, barat, lengang, pasca, gempa, skala, richter, melanda, lombok, biasanya, kawasan, ini, ramai, senggigi, yang, merupakan, pusat, dan, destinasi, andalan, penyumbang, pad, pariwisata, terbesar, di, lombok, barat, ini, terpantau, lengang, sejak, senin, pagi, sebabnya, hampir, semua, wisatawan, mancanegara, telah, dievakuasi, untuk, meninggalkan, wilayah, senggigi, akibat, adanya, gempa, bumi, dahsyat, yang, berpusat, di, pulau, lombok
Kesehatan Non-gempa	hingga, saat, ini, belum, ada, kasus, positif, virus, corona, ncov, novel, coronavirus, di, indonesia, khususnya, di, jakarta, ada, rs, yang, siap, menangani, jika, ada, kasus, yang, terkonfirmasi, tiga, rumah, sakit, di, jakarta, yang, menjadi, pusat, rujukan, nasional, adalah, rumah, sakit, pusat, angkatan, darat, rspad, gatot, soebroto, rs, pusat, infeksi, sulianti, saroso, dan, rs, pusat, persahabatan, ketiganya, termasuk, dalam, rs, yang, disiagakan, untuk, menangani, virus, corona
Ekonomi Non-gempa	bank, indonesia, bi, meyakini, pertumbuhan, ekonomi, indonesia, sepanjang, tumbuh, di, level, hal, itu, ditopang, dari, kecenderungan, kondisi, siklus, ekonomi, yang, kian, membaik, gubernur, bi, perry, warjiyo, mengatakan, pertumbuhan, ekonomi, indonesia, tetap, berdaya, tahan, ditopang, perbaikan, ekspor, dan, konsumsi, rumah, tangga, yang, tetap, baik
Pariwisata Non-gempa	asosiasi, hotel, mataram, bekerja, sama, dengan, organisasi, kerja, sama, internasional, milik, pemerintah, jerman, giz, untuk, membangun, pariwisata, berkelanjutan, di, pulau, lombok, nusa, tenggara, barat, penandatanganan, nota, kesepahaman, mou, antara, ketua, asosiasi, hotel, mataram, ahm, reza, bouvierd, dengan, team,

	leader, tourism, and, investment, giz, oliver, oehms, berlangsung, di, kota, mataram, kamis, lalu

b. *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 [15]. 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 [16]. *Stemming* bertujuan agar suatu kata yang memiliki imbuhan yang berbeda dapat diartikan sebagai kata yang sama. Contoh artikel yang telah melewati proses *stemming* dengan *feature unigram* dapat dilihat pada Tabel 3.5.

Tabel 3.5 Contoh artikel yang telah melewati proses *stemming* dengan *feature unigram*.

Kategori Dokumen	*Token* / *term*
Kesehatan Gempa	viral, di, media, sosial, postingan, netizen, yang, tinggal, di, lombok, dalam, postingan, sebut, ia, aku, alami, halusinasi, di, mana, tanah, asa, goyang, padahal, tidak, jadi, gempa, psikolog, dari, personal, growth, veronica, adesla, kata, trauma, psikologis, memang, rentan, jadi, pada, korban, bencana, alam, seperti, di, lombok, dalam, pedoman, diagnostic, and, statistical, manual, of, mental, disorders, fifth, edition, dsm, sebut, bahwa, salah, satu, ciri, trauma, psikologis, adalah, rasa, jadi, yang, buat, trauma
Ekonomi Gempa	gempa, guncang, lombok, ntb, sejak, akhir, agustus, hingga, awal, september, turut, catat, perintah, nilai, rusa, akibat, gempa, yang, langsung, dari, agustus, hingga, september, capai, rp, triliun, kepala, badan, nasional, tanggulang, bencana, bnpb, willem, rampangilei, kata, total, rusa, per, september, sebut, akumulasi, dari, tujuh, wilayah, yakni, lombok, utara, lombok, timur, sumbawa, barat, sumbawa, mataram, lombok, barat, lombok, tengah

Pariwisata Gempa	kawasan, wisata, pantai, senggigi, di, kabupaten, lombok, barat, lengang, pasca, gempa, skala, richter, landa, lombok, biasa, kawasan, ini, ramai, senggigi, yang, rupa, pusat, dan, destinasi, andal, sumbang, pad, pariwisata, besar, di, lombok, barat, ini, pantau, lengang, sejak, senin, pagi, sebab, hampir, semua, wisatawan, mancanegara, telah, evakuasi, untuk, tinggal, wilayah, senggigi, akibat, ada, gempa, bumi, dahsyat, yang, pusat, di, pulau, lombok
Kesehatan Non-gempa	hingga, saat, ini, belum, ada, kasus, positif, virus, corona, ncov, novel, coronavirus, di, indonesia, khusus, di, jakarta, ada, rs, yang, siap, tangan, jika, ada, kasus, yang, konfirmasi, tiga, rumah, sakit, di, jakarta, yang, jadi, pusat, rujuk, nasional, adalah, rumah, sakit, pusat, angkat, darat, rspad, gatot, soebroto, rs, pusat, infeksi, sulianti, saroso, dan, rs, pusat, sahabat, tiga, masuk, dalam, rs, yang, siaga, untuk, tangan, virus, corona
Ekonomi Non-gempa	bank, indonesia, bi, yakin, tumbuh, ekonomi, indonesia, panjang, tumbuh, di, level, hal, itu, topang, dari, cenderung, kondisi, siklus, ekonomi, yang, kian, baik, gubernur, bi, perry, warjiyo, kata, tumbuh, ekonomi, indonesia, tetap, daya, tahan, topang, baik, ekspor, dan, konsumsi, rumah, tangga, yang, tetap, baik
Pariwisata Non-gempa	asosiasi, hotel, mataram, kerja, sama, dengan, organisasi, kerja, sama, internasional, milik, perintah, jerman, giz, untuk, bangun, pariwisata, lanjut, di, pulau, lombok, nusa, tenggara, barat, penandatanganan, nota, paham, mou, antara, ketua, asosiasi, hotel, mataram, ahm, reza, bouvierd, dengan, team, leader, tourism, and, investment, giz, oliver, oehms, langsung, di, kota, mataram, kamis, lalu

c. *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 digunakan didasarkan dari *stop words* pada *library* Sastrawi.

Contoh artikel yang telah melewai proses *stop-word removal* dengan *feature unigram* dapat dilihat pada Tabel 3.6.

Tabel 3.6 Contoh artikel yang telah melalui proses *stop-word removal* dengan *feature unigram*.

Kategori Dokumen	*Token* / *term*
Kesehatan Gempa	viral, media, sosial, postingan, netizen, tinggal, lombok, postingan, alami, halusinasi, tanah, asa, goyang, gempa, psikolog, personal, growth, veronica, adesla, trauma, psikologis, rentan, korban, bencana, alam, lombok, pedoman, diagnostic, and, statistical, manual, of, mental, disorders, fifth, edition, dsm, salah, ciri, trauma, psikologis, trauma
Ekonomi Gempa	gempa, guncang, lombok, ntb, agustus, september, catat, perintah, nilai, rusa, akibat, gempa, agustus, september, capai, rp, triliun, kepala, badan, nasional, tanggulang, bencana, bnpb, willem, rampangilei, total, rusa, september, akumulasi, tujuh, wilayah, lombok, utara, lombok, timur, sumbawa, barat, sumbawa, mataram, lombok, barat, lombok, tengah
Pariwisata Gempa	kawasan, wisata, pantai, senggigi, kabupaten, lombok, barat, lengang, pasca, gempa, skala, richter, landa, lombok, kawasan, ramai, senggigi, pusat, destinasi, andal, sumbang, pad, pariwisata, lombok, barat, pantau, lengang, senin, pagi, wisatawan, mancanegara, evakuasi, tinggal, wilayah, senggigi, akibat, gempa, bumi, dahsyat, pusat, pulau, lombok
Kesehatan Non-gempa	positif, virus, corona, ncov, novel, coronavirus, indonesia, jakarta, rs, tangan, konfirmasi, rumah, sakit, jakarta, pusat, rujuk, nasional, rumah, sakit, pusat, angkat, darat, rspad, gatot, soebroto, rs, pusat, infeksi, sulianti, saroso, rs, pusat, sahabat, rs, siaga, tangan, virus, corona
Ekonomi Non-gempa	bank, indonesia, bi, tumbuh, ekonomi, indonesia, tumbuh, level, topang, cenderung, kondisi, siklus, ekonomi, kian, gubernur, bi, perry, warjiyo, tumbuh, ekonomi, indonesia, daya, tahan, topang, ekspor, konsumsi, rumah, tangga

Pariwisata Non-gempa	asosiasi, hotel, mataram, organisasi, internasional, milik, perintah, jerman, giz, bangun, pariwisata, pulau, lombok, nusa, tenggara, barat, penandatanganan, nota, paham, mou, ketua, asosiasi, hotel, mataram, ahm, reza, bouvierd, team, leader, tourism, and, investment, giz, oliver, oehms, kota, mataram, kamis

Sedangkan untuk contoh artikel yang telah melewati *preprocessing* dengan *feature bigram* dapat dilihat pada Tabel 3.7.

Tabel 3.7 Contoh artikel yang telah melewati *preprocessing* dengan *feature bigram*.

Kategori Dokumen	*Token*
Kesehatan Gempa	viral media, media sosial, sosial postingan, postingan netizen, netizen tinggal, tinggal lombok, lombok postingan, postingan alami, alami halusinasi, halusinasi tanah, tanah asa, asa goyang, goyang gempa, gempa psikolog, psikolog personal, personal growth, growth veronica, veronica adesla, adesla trauma, trauma psikologis, psikologis rentan, rentan korban, korban bencana, bencana alam, alam lombok, lombok pedoman, pedoman diagnostic, diagnostic and, and statistical, statistical manual, manual of, of mental, mental disorders, disorders fifth, fifth edition, edition dsm, dsm salah, salah ciri, ciri trauma, trauma psikologis, psikologis trauma
Ekonomi Gempa	gempa guncang, guncang lombok, lombok ntb, ntb agustus, agustus september, september catat, catat perintah, perintah nilai, nilai rusa, rusa akibat, akibat gempa, gempa agustus, agustus september, september capai, capai rp, rp triliun, triliun kepala, kepala badan, badan nasional, nasional tanggulang, tanggulang bencana, bencana bnpb, bnpb willem, willem rampangilei, rampangilei total, total rusa, rusa september, september akumulasi, akumulasi tujuh, tujuh wilayah, wilayah lombok, lombok utara, utara lombok, lombok timur, timur sumbawa, sumbawa barat, barat sumbawa, sumbawa mataram, mataram lombok, lombok barat, barat lombok, lombok tengah

Pariwisata Gempa	kawasan wisata, wisata pantai, pantai senggigi, senggigi kabupaten, kabupaten lombok, lombok barat, barat lengang, lengang pasca, pasca gempa, gempa skala, skala richter, richter landa, landa lombok, lombok kawasan, kawasan ramai, ramai senggigi, senggigi pusat, pusat destinasi, destinasi andal, andal sumbang, sumbang pad, pad pariwisata, pariwisata lombok, lombok barat, barat pantau, pantau lengang, lengang senin, senin pagi, pagi wisatawan, wisatawan mancanegara, mancanegara evakuasi, evakuasi tinggal, tinggal wilayah, wilayah senggigi, senggigi akibat, akibat gempa, gempa bumi, bumi dahsyat, dahsyat pusat, pusat pulau, pulau lombok
Kesehatan Non-gempa	positif virus, virus corona, corona ncov, ncov novel, novel coronavirus, coronavirus indonesia, indonesia jakarta, jakarta rs, rs tangan, tangan konfirmasi, konfirmasi rumah, rumah sakit, sakit jakarta, jakarta pusat, pusat rujuk, rujuk nasional, nasional rumah, rumah sakit, sakit pusat, pusat angkat, angkat darat, darat rspad, rspad gatot, gatot soebroto, soebroto rs, rs pusat, pusat infeksi, infeksi sulianti, sulianti saroso, saroso rs, rs pusat, pusat sahabat, sahabat rs, rs siaga, siaga tangan, tangan virus, virus corona
Ekonomi Non-gempa	bank indonesia, indonesia bi, bi tumbuh, tumbuh ekonomi, ekonomi indonesia, indonesia tumbuh, tumbuh level, level topang, topang cenderung, cenderung kondisi, kondisi siklus, siklus ekonomi, ekonomi kian, kian gubernur, gubernur bi, bi perry, perry warjiyo, warjiyo tumbuh, tumbuh ekonomi, ekonomi indonesia, indonesia daya, daya tahan, tahan topang, topang ekspor, ekspor konsumsi, konsumsi rumah, rumah tangga
Pariwisata Non-gempa	asosiasi hotel, hotel mataram, mataram organisasi, organisasi internasional, internasional milik, milik perintah, perintah jerman, jerman giz, giz bangun, bangun pariwisata, pariwisata pulau, pulau lombok, lombok nusa, nusa tenggara, tenggara barat, barat penandatanganan, penandatanganan nota, nota paham, paham mou, mou ketua, ketua asosiasi, asosiasi hotel, hotel mataram, mataram ahm, ahm reza, reza bouvierd, bouvierd team, team leader, leader

	tourism, tourism and, and investment, investment giz, giz oliver, oliver oehms, oehms kota, kota mataram, mataram kamis

3.5.3 *Feature weighting*

Feature weighting dilakukan dengan menggunakan metode TF-IDF. Tahap *feature weighting* dibagi menjadi 3 yaitu perhitungan *term frequency* (TF), perhitungan *inverse document frequency* (IDF) dan terakhir perhitungan TF-IDF.

a. *Term Frequency* (TF)

Pada tahap ini, dilakukan perhitungan *term frequency* pada seluruh kategori. Daftar 10 *feature unigram* dan 5 *feature bigram* dengan frekuensi paling tinggi pada contoh dokumen *training* dapat dilihat pada Tabel 3.8.

Tabel 3.8 Contoh TF pada artikel yang telah melewati tahap *preprocessing*.

Token / Term	TF					
	Ekonomi Gempa	Kesehatan Gempa	Pariwisata Gempa	Ekonomi Non-gempa	Kesehatan Non-gempa	Pariwisata Non-gempa
lombok	5	2	4	0	0	1
pusat	0	0	2	0	4	0
barat	2	0	2	0	0	1
gempa	2	1	2	0	0	0
indonesia	0	0	0	3	1	0
mataram	1	0	0	0	0	3
rs	0	0	0	0	4	0
ekonomi	0	0	0	3	0	0
rumah	0	0	0	1	2	0
senggigi	0	0	3	0	0	0
lombok barat	1	0	2	0	0	0
agustus september	2	0	0	0	0	0
akibat gempa	1	0	1	0	0	0
asosiasi hotel	0	0	0	0	0	2
ekonomi indonesia	0	0	0	2	0	0

b. *Inverse Document Frequency* (IDF)

IDF merupakan *invers* dari *document frequency*. *Document frequency* merupakan jumlah dokumen yang mengandung suatu *feature*. Untuk menghitung IDF, digunakan

persamaan (2-2). Daftar 10 *feature unigram* dan 5 *feature bigram* pada contoh artikel *training* dengan IDF terendah dapat dilihat pada Tabel 3.9.

Tabel 3.9 Contoh IDF pada artikel yang telah melewati tahap *preprocessing.*

Token / Term	IDF
lombok	0,176091
barat	0,30103
gempa	0,30103
pusat	0,477121
indonesia	0,477121
mataram	0,477121
rumah	0,477121
akibat	0,477121
and	0,477121
bencana	0,477121
lombok barat	0,477121
akibat gempa	0,477121
pulau lombok	0,477121
agustus september	0,778151
asosiasi hotel	0,778151

c. TF-IDF

Setelah didapatkan TF dan IDF untuk setiap *term*, kemudian dilakukan perhitungan TF-IDF masing-masing *term* untuk tiap kategori. Untuk menghitung TF-IDF, digunakan persamaan (2-3). Daftar 10 *feature unigram* dan 5 *feature bigram* pada contoh artikel *training* dengan TF-IDF tertinggi dapat dilihat pada Tabel 3.10.

Tabel 3.10 Contoh TF-IDF pada artikel yang telah melewati tahap *preprocessing.*

Token / Term	TF-IDF					
	Ekonomi Gempa	Kesehatan Gempa	Pariwisata Gempa	Ekonomi Non-gempa	Kesehatan Non-gempa	Pariwisata Non-gempa
rs	0	0	0	0	3,112605	0
pusat	0	0	0,954243	0	1,908485	0
ekonomi	0	0	0	2,334454	0	0
senggigi	0	0	2,334454	0	0	0
september	2,334454	0	0	0	0	0
trauma	0	2,334454	0	0	0	0
tumbuh	0	0	0	2,334454	0	0
lombok	0,880456	0,352183	0,704365	0	0	0,176091
indonesia	0	0	0	1,431364	0,477121	0
mataram	0,477121	0	0	0	0	1,431364

agustus september	1,556303	0	0	0	0	0
asosiasi hotel	0	0	0	0	0	1,556303
ekonomi indonesia	0	0	0	1,556303	0	0
hotel mataram	0	0	0	0	0	1,556303
rs pusat	0	0	0	0	1,556303	0

Setelah tahap *term weighting*, didapatkan jumlah kata unik dari artikel contoh sejumlah 160 *feature unigram* dan 212 *feature bigram*.

3.5.4 *Training* dengan MNB

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 (2-7). Dari contoh data *training* yang telah melewati tahap *preprocessing* pada Tabel 3.6, setiap kategori terdiri dari 1 artikel. Oleh karena itu, probabilitas prior dari masing-masing kategori adalah:

$$P(ekonomi\ gempa) = \frac{1}{6} = 0{,}16667$$

$$P(kesehatan\ gempa) = \frac{1}{6} = 0{,}16667$$

$$P(pariwisata\ gempa) = \frac{1}{6} = 0{,}16667$$

$$P(ekonomi\ non\ gempa) = \frac{1}{6} = 0{,}16667$$

$$P(kesehatan\ non\ gempa) = \frac{1}{6} = 0{,}16667$$

$$P(pariwisata\ non\ gempa) = \frac{1}{6} = 0{,}16667$$

Setelah didapatkan probabilitas *prior* dari tiap kategori, proses *training* dilanjutkan dengan menghitung probabilitas suatu *feature* terdapat pada suatu kategori. Proses perhitungan dilakukan dengan Persamaan (2-10). Dari tahap *feature weighting*, jumlah kata unik (*vocabulary*) yang didapatkan adalah 160 *feature unigram* dan 212 *feature bigram* dengan jumlah IDF sebesar 283,0995. Sedangkan jumlah TF-IDF *feature* yang terdapat pada kategori Ekonomi Gempa, Kesehatan Gempa, Pariwisata Gempa, Ekonomi Non-gempa, Kesehatan Non-gempa, dan Pariwisata Non-gempa berturut-Turut adalah 58.8158, 62,0022, 56.9585, 41,5942, 55.9531 dan 54.8739. Berikut adalah contoh perhitungan $P(t|c)$ dari kata “lombok”:

$$P(lombok \mid ekonomi\ gempa) = \frac{0{,}880456+1}{58{,}8158+283{,}0995} = 0{,}005499772$$

$$P(lombok \mid kesehatan\ gempa) = \frac{0{,}352183+1}{62{,}0022+283{,}0995} = 0{,}003918215$$

$$P(lombok \mid pariwisata\ gempa) = \frac{0{,}704365+1}{56{,}9585+283{,}0995} = 0{,}005011984$$

$$P(lombok \mid ekonomi\ non\ gempa) = \frac{0+1}{41{,}5942+283{,}0995} = 0{,}003079826$$

$$P(lombok \mid kesehatan\ non\ gempa) = \frac{0+1}{55{,}9531+283{,}0995} = 0{,}002949395$$

$$P(lombok \mid pariwisata\ non\ gempa) = \frac{0{,}176091+1}{54{,}8739+283{,}0995} = 0{,}003479834$$

P(*c*) dan *P*(*t*|*c*) dari untuk tiap *feature* dan kategori inilah yang akan menjadi model *training* dari metode *Naïve Bayes*. Model ini kemudian digunakan untuk mengklasifikasikan suatu dokumen yang tidak diketahui kategori-nya.

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 untuk tiap kategori adalah:

$$P(t \mid ekonomi\ gempa) = \frac{0+1}{58{,}8158+283{,}0995} = 0{,}002924701$$

$$P(t \mid kesehatan\ gempa) = \frac{0+1}{62{,}0022+283{,}0995} = 0{,}002897696$$

$$P(t \mid pariwisata\ gempa) = \frac{0+1}{56{,}9585+283{,}0995} = 0{,}002940675$$

$$P(t \mid ekonomi\ non\ gempa) = \frac{0+1}{41{,}5942+283{,}0995} = 0{,}003079826$$

$$P(t \mid kesehatan\ non\ gempa) = \frac{0+1}{55{,}9531+283{,}0995} = 0{,}002949395$$

$$P(t \mid pariwisata\ non\ gempa) = \frac{0+1}{54{,}8739+283{,}0995} = 0{,}002958813$$

3.5.5 Klasifikasi dengan MNB

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 (2-6). Pada Persamaan (2-6), dilakukan perbandingan terhadap probabilitas suatu artikel berada di suatu kategori *c* untuk tiap kategori. Sebelum dilakukan klasifikasi, artikel *testing* terlebih dahulu melewati proses *preprocessing*. Contoh artikel *testing* yang telah melewati tahap *preprocessing* dapat dilihat pada Tabel 3.11.

Tabel 3.11 Contoh artikel *testing*.

Artikel *Testing*	*Feature unigram*	*Feature bigram*
Pemerintah mengaku tengah mengupayakan percepatan pencairan anggaran sebesar Rp1 triliun bagi masyarakat terdampak gempa di Lombok, Nusa Tenggara Barat.	perintah, tengah, upaya, cepat, cair, anggar, rp, triliun, masyarakat, dampak, gempa, lombok, nusa, tenggara, barat	perintah tengah, tengah upaya, upaya cepat, cepat cair, cair anggar, anggar rp, rp triliun, triliun masyarakat, masyarakat dampak, dampak gempa, gempa lombok, lombok nusa, nusa tenggara, tenggara barat

Hasil klasifikasi merupakan kategori dengan probabilitas tertinggi. $P(t|c)$ untuk tiap *feature* pada artikel *testing* di tiap kategori dapat dilihat pada Tabel 3.12.

Tabel 3.12 Contoh $P(t|c)$ untuk dokumen *testing*.

Term (*t*)	$P(t\|c)$					
	c = Ekonomi Gempa	*c* = Kesehatan Gempa	*c* = Pariwisata Gempa	*c* = Ekonomi Non-gempa	*c* = Kesehatan Non-gempa	*c* = Pariwisata Non-gempa
barat	0,004686	0,002898	0,004711	0,00308	0,002949	0,00385
gempa	0,004686	0,00377	0,004711	0,00308	0,002949	0,002959
lombok	0,0055	0,003918	0,005012	0,00308	0,002949	0,00348
lombok nusa	0,002925	0,002898	0,002941	0,00308	0,002949	0,005261
nusa	0,002925	0,002898	0,002941	0,00308	0,002949	0,005261
nusa tenggara	0,002925	0,002898	0,002941	0,00308	0,002949	0,005261
perintah	0,00432	0,002898	0,002941	0,00308	0,002949	0,004371
rp	0,005201	0,002898	0,002941	0,00308	0,002949	0,002959

rp triliun	0,005201	0,002898	0,002941	0,00308	0,002949	0,002959
tengah	0,005201	0,002898	0,002941	0,00308	0,002949	0,002959
tenggara	0,002925	0,002898	0,002941	0,00308	0,002949	0,005261
tenggara barat	0,002925	0,002898	0,002941	0,00308	0,002949	0,005261
triliun	0,005201	0,002898	0,002941	0,00308	0,002949	0,002959
anggar	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
anggar rp	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
cair	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
cair anggar	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
cepat	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
cepat cair	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
dampak	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
dampak gempa	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
gempa lombok	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
masyarakat	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
masyarakat dampak	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
perintah tengah	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
tengah upaya	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
triliun masyarakat	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
upaya	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959
upaya cepat	0,002925	0,002898	0,002941	0,00308	0,002949	0,002959

Pada artikel *testing*, terdapat beberapa *feature* yang tidak pernah ditemukan di dokumen training. Oleh karena itu, $P(t|c)$ dari kata-kata tersebut didapatkan dari nilai

default yang ditentukan saat proses *training*, dimana *feature* tersebut memiliki bobot 0 untuk setiap kategori.

Berikut adalah contoh perhitungan probabilitas dokumen *testing* yang telah melewati tahap *preprocessing* pada Tabel 3.11 berdasarkan model yang telah dibentuk dari proses *training* menggunakan contoh artikel *training* pada Tabel 3.3:

$$P(ekonomi\ \ gempa \mid testing) = \log P(ekonomi\ gempa) + \sum_{1 \le k \le n} \log P(t_k \mid ekonomi\ gempa)$$

$$\begin{aligned} P(ekonomi\ gempa \mid testing) = {} & \log P(ekonomi\ gempa) + \\ & \log P(\text{"perintah"} \mid ekonomi\ gempa) + \\ & \log P(\text{"tengah"} \mid ekonomi\ gempa) + \\ & \log P(\text{"upaya"} \mid ekonomi\ gempa) + \\ & \ldots \\ & \log P(\text{"tengah upaya"} \mid ekonomi\ gempa) + \\ & \log P(\text{"upaya cepat"} \mid ekonomi\ gempa) + \\ & \log P(\text{"cepat cair"} \mid ekonomi\ gempa) \end{aligned}$$

$$\begin{aligned} P(ekonomi\ gempa \mid testing) = {} & (-0{,}77815) + (-2{,}364502405) + \\ & (-2{,}283949853) + (-2{,}533918553) + \\ & \ldots \\ & (-2{,}533918553) + (-2{,}533918553) + \\ & (-2{,}533918553) \end{aligned}$$

$$P(ekonomi\ gempa \mid testing) = -72{,}40887754$$

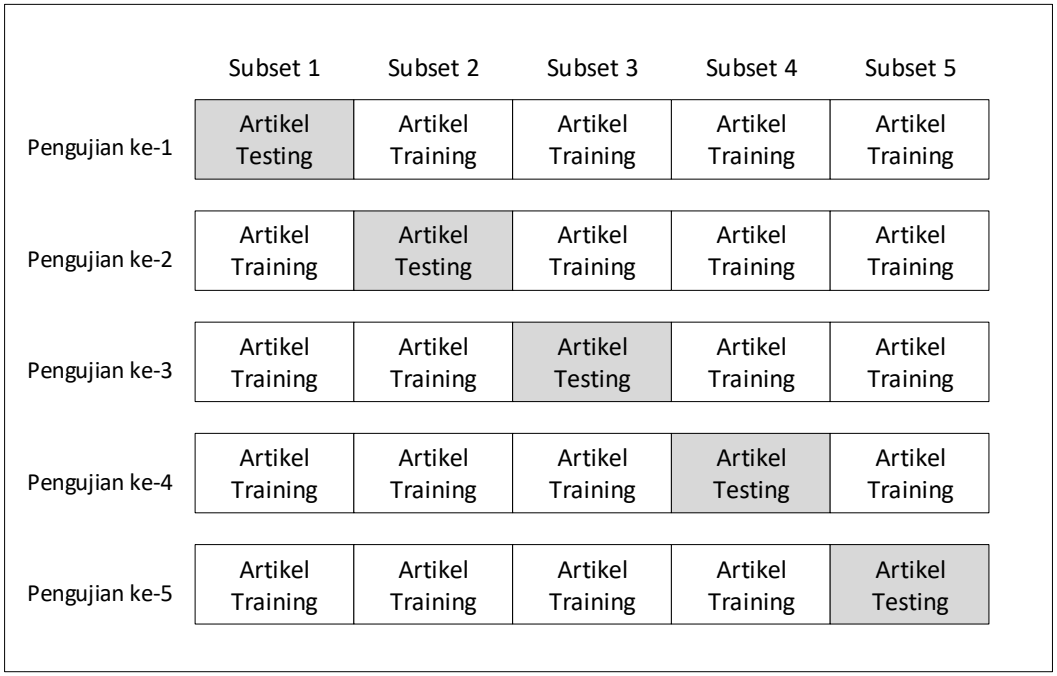
Perhitungan serupa juga dilakukan untuk kategori kesehatan dan pariwisata. Hasil yang didapatkan untuk $P(kesehatan\ gempa \mid testing)$ adalah -74,133, untuk $P(pariwisata\ gempa \mid testing)$ adalah -73,552, $P(ekonomi\ non\ gempa \mid testing)$ adalah -73,611, $P(kesehatan\ non\ gempa \mid testing)$ adalah -74,156, $P(pariwisata\ non\ gempa \mid testing)$ adalah -72,512. Dari probabilitas dari 6 kategori yang telah dihitung, didapatkan c_{map} -72,4089 yaitu untuk kategori Ekonomi Gempa. Hasil klasifikasi adalah benar karena artikel *testing* sebenarnya berada di kategori Ekonomi Gempa.

3.6 Pengumpulan Data

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 bidang ilmu komunikasi Shinta Desiyana Fajarica, S.IP., M.Si. 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*.

3.7 Rencana Pengujian

Pengujian yang dilakukan dalam penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes* adalah dengan teknik *k-fold cross validation*. Teknik ini membagi *dataset* menjadi *k* bagian, dimana tiap bagian memiliki ukuran yang sama dan berisi data yang acak. Nilai *k* yang digunakan adalah 5 karena jumlah *dataset* yang cukup terbatas. Selain itu, jumlah *dataset* adalah 1000 data, yang merupakan kelipatan 5. Ilustrasi dari pemisahan data untuk setiap percobaan dapat dilihat pada Gambar 3.3.



Gambar 3.3 Ilustrasi 5-*fold cross validation*.

Pada setiap percobaan, evaluasi dilakukan dengan menghitung *recall*, *precision*, dan *f-measure* dari model. Untuk menghitung nilai-nilai tersebut, diperlukan *confusion matrix* untuk menyajikan hasil klasifikasi dalam bentuk tabel. Tabel *confusion matrix* dalam penelitian ini dapat dilihat pada Tabel 3.13.

Tabel 3.13 *Confusion matrix* yang digunakan dalam penelitian.

Realita	Sistem						Total
	Kelas Ekonomi Gempa	Kelas Kesehatan Gempa	Kelas Pariwisata Gempa	Kelas Ekonomi Non-gempa	Kelas Kesehatan Non-gempa	Kelas Pariwisata Non-gempa	
Kelas Ekonomi Gempa	*True Positive*	*Error*	*Error*	*Error*	*Error*	*Error*	Total Kelas Ekonomi Gempa
Kelas Kesehatan Gempa	*Error*	*True Positive*	*Error*	*Error*	*Error*	*Error*	Total Kelas Kesehatan Gempa
Kelas Pariwisata Gempa	*Error*	*Error*	*True Positive*	*Error*	*Error*	*Error*	Total Kelas Pariwisata Gempa
Kelas Ekonomi Non-gempa	*Error*	*Error*	*Error*	*True Positive*	*Error*	*Error*	Total Kelas Ekonomi Non-gempa
Kelas Kesehatan Non-gempa	*Error*	*Error*	*Error*	*Error*	*True Positive*	*Error*	Total Kelas Kesehatan Non-gempa
Kelas Pariwisata Non-gempa	*Error*	*Error*	*Error*	*Error*	*Error*	*True Positive*	Total Kelas Pariwisata Non-gempa
	Prediksi Kelas Ekonomi Gempa	Prediksi Kelas Kesehatan Gempa	Prediksi Kelas Pariwisata Gempa	Prediksi Kelas Ekonomi Non-gempa	Prediksi Kelas Kesehatan Non-gempa	Prediksi Kelas Pariwisata Non-gempa	

Recall dan *precision* untuk tiap kategori dihitung dengan menggunakan Persamaan (2-7) dan Persamaan (2-8). Sedangkan untuk *f-measure* dihitung menggunakan Persamaan (2-9). 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.

3.8 Jadwal Kegiatan

Waktu yang digunakan dalam proses pengembangan sistem klasifikasi artikel online gempa di Indonesia yaitu selama enam bulan. Jadwal kegiatan dapat dilihat pada Tabel 3.14.

Tabel 3.14 Jadwal kegiatan.

No.	Kegiatan	Waktu (Bulan)					Keterangan
		I	II	III	IV	V	
1	Analisa						Analisa kebutuhan
2	Pengumpulan Data						Pengumpulan artikel *online*
3	Pembangunan Sistem						Pengkodean sistem
4	*Testing*						Pengujian sistem
5	Implementasi						Penerapan sistem
6	Dokumentasi						Dokumentasi sistem

BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data yang digunakan dalam penelitian tentang klasifikasi artikel online gempa di Indonesia menggunakan *multinomial naïve Bayes* ini adalah artikel online tentang gempa di Indonesia serta artikel *online* yang tidak berhubungan dengan gempa yang dikumpulkan dari situs *www.kompas.com*, *www.detik.com*, *www.liputan6.com*, dan *www.okezone.com*. 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.

4.2 Pengujian

Pengujian dilakukan dengan menggunakan teknik 5-*fold cross validation*. Penggunaan teknik ini bertujuan agar setiap data pada *dataset* digunakan sebagai data *training* dan data *testing*. Setiap data dibagi menjadi 5 *subset*, kemudian dilakukan 5 perulangan 5-*fold cross validation* dimana *subset* yang digunakan sebagai data uji berbeda-beda untuk setiap pengujian. Pada *dataset*, 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, dan 240 artikel untuk label Pariwisata Non-gempa. Setiap *subset* terdiri dari 20% dari jumlah data untuk tiap kategori, sehingga setiap subset terdiri dari 20 artikel untuk label Kesehatan Gempa, 20 artikel untuk label Ekonomi Gempa, 20 artikel untuk label Pariwisata Gempa, 46 artikel untuk label Kesehatan Non-gempa, 46 artikel untuk label Ekonomi Non-gempa, dan 48 artikel untuk label Pariwisata Non-

gempa. Ilustrasi proses pengujian dapat dilihat pada Gambar 4.3. *Subset* yang digunakan sebagai data *testing* ditandai dengan latar berwarna abu-abu.

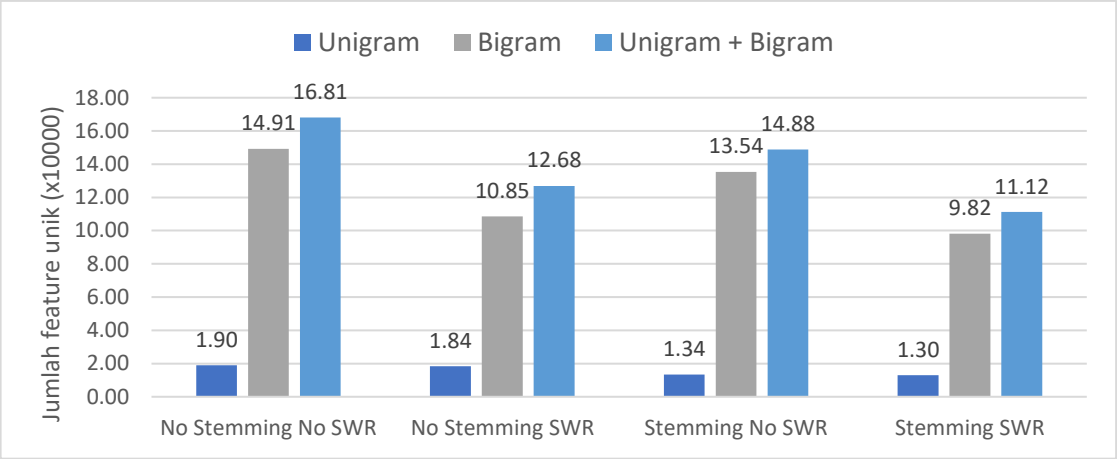
Iterasi 1	Ekonomi Gempa	**20**	20	20	20	20
	Kesehatan Gempa	**20**	20	20	20	20
	Pariwisata Gempa	**20**	20	20	20	20
	Ekonomi Non-gempa	**46**	46	46	46	46
	Kesehatan Non-gempa	**46**	46	46	46	46
	Pariwisata Non-gempa	**48**	48	48	48	48
Iterasi 2	Ekonomi Gempa	20	**20**	20	20	20
	Kesehatan Gempa	20	**20**	20	20	20
	Pariwisata Gempa	20	**20**	20	20	20
	Ekonomi Non-gempa	46	**46**	46	46	46
	Kesehatan Non-gempa	46	**46**	46	46	46
	Pariwisata Non-gempa	48	**48**	48	48	48
Iterasi 3	Ekonomi Gempa	20	20	**20**	20	20
	Kesehatan Gempa	20	20	**20**	20	20
	Pariwisata Gempa	20	20	**20**	20	20
	Ekonomi Non-gempa	46	46	**46**	46	46
	Kesehatan Non-gempa	46	46	**46**	46	46
	Pariwisata Non-gempa	48	48	**48**	48	48
Iterasi 4	Ekonomi Gempa	20	20	20	**20**	20
	Kesehatan Gempa	20	20	20	**20**	20
	Pariwisata Gempa	20	20	20	**20**	20
	Ekonomi Non-gempa	46	46	46	**46**	46
	Kesehatan Non-gempa	46	46	46	**46**	46
	Pariwisata Non-gempa	48	48	48	**48**	48
Iterasi 5	Ekonomi Gempa	20	20	20	20	**20**
	Kesehatan Gempa	20	20	20	20	**20**
	Pariwisata Gempa	20	20	20	20	**20**
	Ekonomi Non-gempa	46	46	46	46	**46**
	Kesehatan Non-gempa	46	46	46	46	**46**
	Pariwisata Non-gempa	48	48	48	48	**48**

Gambar 4.1 Ilustrasi pengujian dengan 5-*fold cross validation*.

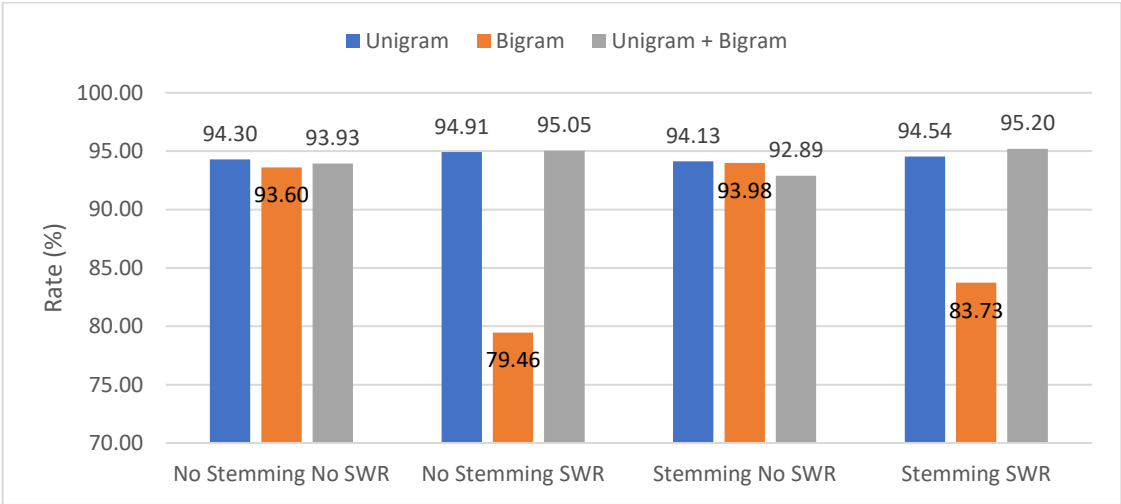
Terdapat beberapa skema pengujian yang dilakukan pada penelitian ini, antara lain adalah klasifikasi hanya dengan menggunakan *feature unigram*, klasifikasi hanya dengan menggunakan *feature bigram*, dan klasifikasi dengan *feature unigram* dan *bigram*. Selain itu, dilakukan juga pengujian dengan *stemming*,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

4.3 Hasil Pengujian

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 4.2 dan Gambar 4.3.



Gambar 4.2 Pengaruh jenis pengujian terhadap ukuran *vocabulary*.



Gambar 4.3 Pengaruh jenis pengujian terhadap nilai *f-measure*.

Pada Gambar 4.2, 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 4.3, 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 4.3 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.

4.3.1 Pengujian dengan *feature unigram*

Pada skema pengujian ini, *feature* yang digunakan adalah *feature unigram* atau 1 *token* per *feature*. Pengujian juga dilakukan dengan melewati tahap *stemming*, tanpa melewati tahap *stemming*, dengan menyertakan *stopwords*, dan tanpa menyertakan *stopwords*. Hasil yang didapatkan untuk pengujian dengan menggunakan *stemming* dapat dilihat pada Tabel 4.1. Sedangkan hasil dari pengujian tanpa menggunakan *stemming* dapat dilihat pada Tabel 4.2.

Tabel 4.1 Hasil pengujian dengan *feature unigram* tanpa *stopwords removal*.

Iterasi	Tanpa Stemming			Dengan Stemming		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	94.76%	92.72%	93.59%	94.87%	93.19%	93.86%
2	94.38%	93.20%	93.62%	93.66%	92.84%	93.03%
3	93.36%	92.77%	93.00%	94.49%	94.79%	94.58%
4	96.69%	94.98%	95.67%	95.90%	93.67%	94.55%
5	96.19%	94.29%	95.16%	95.63%	93.60%	94.48%
6	94.88%	93.34%	94.04%	94.14%	93.48%	93.74%
7	94.19%	92.26%	93.12%	93.82%	91.91%	92.77%
8	95.58%	94.05%	94.73%	95.32%	93.72%	94.40%
9	95.81%	93.67%	94.57%	95.53%	93.80%	94.46%
10	94.44%	94.16%	94.13%	94.44%	94.16%	94.13%
11	96.78%	94.99%	95.68%	97.65%	97.02%	97.27%
12	94.49%	92.42%	93.28%	94.07%	91.58%	92.65%
13	94.85%	94.65%	94.64%	95.24%	95.48%	95.25%
14	95.99%	94.98%	95.45%	95.97%	94.51%	95.20%
15	93.83%	91.90%	92.76%	92.27%	91.30%	91.62%
16	96.07%	95.34%	95.56%	97.12%	96.90%	96.92%
17	93.43%	91.19%	92.10%	93.43%	91.19%	92.10%
18	96.31%	95.61%	95.82%	94.53%	94.07%	94.02%
19	94.72%	91.68%	92.95%	95.44%	92.86%	93.90%
20	96.71%	94.04%	95.18%	95.81%	93.69%	94.58%
21	96.31%	95.36%	95.81%	94.30%	93.34%	93.77%
22	96.34%	94.41%	95.30%	96.73%	95.25%	95.92%
23	95.92%	93.19%	94.22%	95.92%	93.19%	94.22%
24	94.83%	92.86%	93.74%	94.11%	92.15%	93.04%
25	93.42%	93.36%	93.26%	92.85%	93.01%	92.75%
Rata-rata	**95.21%**	**93.66%**	**94.30%**	**94.93%**	**93.63%**	**94.13%**
Standar deviasi	**1.08%**	**1.21%**	**1.09%**	**1.27%**	**1.47%**	**1.32%**

Tabel 4.2 Hasil pengujian dengan *feature unigram* dan melewati *stopwords removal*.

Iterasi	Tanpa Stemming			Dengan Stemming		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
1	95.48%	95.23%	95.30%	93.18%	93.34%	93.21%
2	93.50%	95.10%	94.09%	93.22%	94.63%	93.80%
3	94.30%	95.63%	94.68%	93.44%	94.94%	93.79%
4	95.76%	96.20%	95.97%	95.78%	97.02%	96.34%
5	93.45%	94.99%	94.15%	94.09%	94.52%	94.29%
6	94.50%	95.50%	94.90%	94.13%	95.15%	94.53%
7	93.87%	95.34%	94.53%	93.19%	94.98%	93.96%
8	95.55%	96.10%	95.73%	95.81%	96.10%	95.90%
9	97.00%	97.40%	97.19%	95.49%	95.84%	95.63%
10	92.24%	93.91%	92.79%	92.05%	93.56%	92.60%
11	97.00%	97.39%	97.16%	97.36%	98.22%	97.75%
12	94.86%	93.60%	94.11%	94.50%	93.25%	93.74%
13	93.30%	95.57%	94.24%	92.97%	95.24%	93.92%
14	95.48%	96.31%	95.86%	95.85%	96.68%	96.22%
15	91.73%	93.45%	92.26%	90.13%	92.03%	90.67%
16	97.47%	98.11%	97.70%	96.55%	97.26%	96.86%
17	92.38%	94.05%	93.12%	92.80%	94.88%	93.73%
18	93.49%	94.08%	93.49%	93.49%	94.08%	93.49%
19	94.13%	94.28%	94.05%	93.71%	93.94%	93.69%
20	97.02%	97.86%	97.38%	95.86%	96.68%	96.20%
21	92.88%	94.28%	93.45%	93.22%	94.65%	93.80%
22	96.17%	96.57%	96.32%	97.07%	96.91%	96.91%
23	96.24%	96.16%	96.15%	96.32%	96.17%	96.14%
24	95.10%	96.66%	95.78%	94.09%	95.48%	94.64%
25	91.60%	93.95%	92.45%	91.16%	93.24%	91.81%
Rata-rata	**94.58%**	**95.51%**	**94.91%**	**94.22%**	**95.15%**	**94.54%**
Standar deviasi	**1.70%**	**1.33%**	**1.54%**	**1.79%**	**1.47%**	**1.66%**

Pada Tabel 4.1 dan Tabel 4.2, dapat dilihat nilai dari *precision*, *recall*, dan *f-measure* untuk tiap iterasi beserta nilai rata-ratanya untuk seluruh iterasi. Tabel 4.1 menyajikan hasil dari klasifikasi tanpa melewati tahap *stemming*, baik dengan menyertakan *stopwords* maupun tidak menyertakan *stopwords*. Nilai *precision*, *recall*, dan *f-measure* untuk pengujian tanpa melewati tahap *stemming* dan tidak menyertakan *stopwords* berturut-turut adalah 95.21% 93.66%, dan 94.30%. Sedangkan hasil pengujian tanpa melewati tahap *stemming* dan menyertakan *stopwords* adalah *precision* sebesar 94.58%, *recall* sebesar 95.51%, dan *f-measure* sebesar 94.91%. Tabel 4.2 juga menyajikan pengujian hampir yang sama, yaitu dengan *stopwords* dan tanpa menyertakan

stopwords, akan tetapi dilakukan *stemming* juga pada pengujian ini. Nilai *precision*, *recall* dan *f-measure* dengan menyertakan *stopwords* adalah 94.93%. 93.63%, dan 94.13%. Sedangkan nilai *precision*, *recall* dan *f-measure* tanpa menyertakan *stopwords* adalah 94.22%. 95.15%, dan 94.54%. Setiap percobaan memiliki hasil yang cukup konsisten, dimana nilai dari standar deviasi paling tinggi yang didapatkan adalah 1.66%.

4.3.2 Pengujian dengan *feature bigram*

Pada skema pengujian ini, *feature* yang digunakan adalah *feature bigram* atau 2 *token* bersebelahan per *feature*. Pengujian juga dilakukan dengan melewati tahap *stemming*, tanpa melewati tahap *stemming*, dengan menyertakan *stopwords*, dan tanpa menyertakan *stopwords*. Hasil yang didapatkan untuk pengujian dengan menggunakan *stemming* dapat dilihat pada Tabel 4.3. Sedangkan hasil dari pengujian tanpa menggunakan *stemming* dapat dilihat pada Tabel 4.4.

Tabel 4.3 Hasil pengujian dengan *feature bigram* tanpa *stemming*.

Iterasi	Tanpa Stopwords Removal			Dengan Stopwords Removal		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	92.47%	93.09%	92.76%	79.48%	83.77%	79.10%
2	93.73%	95.50%	94.35%	76.83%	82.03%	75.23%
3	92.59%	94.09%	93.18%	82.68%	87.78%	83.05%
4	91.62%	93.35%	92.24%	82.14%	87.24%	81.79%
5	92.56%	94.76%	93.49%	79.63%	84.88%	78.86%
6	94.48%	95.51%	94.89%	81.92%	87.13%	82.19%
7	92.18%	94.28%	93.04%	78.72%	83.99%	78.42%
8	91.69%	93.74%	92.45%	77.86%	82.98%	77.13%
9	93.57%	94.63%	93.95%	81.63%	86.69%	81.52%
10	92.53%	93.61%	92.65%	77.65%	82.85%	77.50%
11	95.09%	95.01%	94.99%	81.45%	86.97%	82.03%
12	92.77%	93.97%	93.18%	78.40%	83.58%	78.15%
13	91.87%	93.70%	92.57%	79.73%	85.38%	79.24%
14	97.08%	97.40%	97.17%	82.58%	87.69%	82.99%
15	87.78%	89.66%	88.35%	78.42%	82.78%	76.50%
16	94.08%	95.59%	94.69%	80.31%	85.91%	80.41%
17	94.33%	95.81%	94.93%	81.07%	87.06%	81.74%
18	95.83%	95.75%	95.52%	82.65%	87.57%	83.18%
19	92.66%	94.12%	93.17%	76.50%	81.94%	75.72%
20	93.12%	93.81%	93.41%	80.04%	84.82%	79.16%
21	93.83%	95.46%	94.58%	79.57%	84.87%	78.97%
22	95.23%	96.08%	95.62%	83.03%	87.49%	83.85%
23	91.30%	93.68%	92.21%	76.69%	80.49%	74.21%
24	94.80%	95.50%	95.05%	82.44%	87.86%	82.73%
25	90.74%	93.27%	91.57%	76.68%	80.18%	72.87%
Rata-rata	**93.12%**	**94.46%**	**93.60%**	**79.93%**	**84.96%**	**79.46%**
Standar deviasi	**1.85%**	**1.44%**	**1.69%**	**2.12%**	**2.34%**	**2.99%**

Tabel 4.4 Hasil pengujian dengan *feature bigram* dan melewati *stemming*.

Iterasi	Tanpa Stopwords Removal			Dengan Stopwords Removal		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
1	92.49%	92.37%	92.36%	81.23%	85.55%	81.46%
2	93.44%	94.65%	93.83%	81.53%	86.74%	81.64%
3	92.81%	94.44%	93.51%	84.75%	89.20%	85.15%
4	92.31%	93.71%	92.82%	84.00%	88.95%	84.49%
5	94.63%	95.83%	95.17%	82.24%	87.69%	82.85%
6	93.78%	94.79%	94.18%	87.55%	91.75%	88.70%
7	91.73%	93.45%	92.43%	83.57%	88.73%	84.42%
8	91.04%	93.38%	91.89%	81.88%	86.93%	82.33%
9	94.92%	94.16%	94.41%	83.19%	87.29%	83.07%
10	93.14%	94.77%	93.72%	80.05%	85.31%	80.87%
11	96.70%	96.54%	96.53%	81.72%	86.84%	82.36%
12	94.77%	95.02%	94.83%	85.31%	88.59%	85.41%
13	92.19%	93.71%	92.78%	83.89%	89.26%	85.01%
14	97.78%	97.73%	97.73%	85.26%	89.97%	85.77%
15	87.92%	89.64%	88.45%	79.73%	84.06%	78.80%
16	94.42%	95.95%	95.07%	83.84%	88.64%	84.83%
17	96.25%	96.54%	96.30%	82.94%	88.38%	84.02%
18	96.16%	96.10%	95.88%	85.07%	89.02%	85.56%
19	92.91%	94.11%	93.37%	81.11%	86.66%	81.85%
20	93.79%	94.18%	93.96%	85.00%	89.55%	85.35%
21	94.23%	94.98%	94.54%	83.09%	88.43%	83.82%
22	96.43%	96.43%	96.40%	86.85%	90.33%	87.65%
23	90.82%	92.36%	91.36%	80.21%	85.53%	80.27%
24	96.26%	96.20%	96.19%	87.70%	92.13%	88.83%
25	90.99%	93.26%	91.75%	79.98%	84.44%	78.81%
Rata-rata	**93.68%**	**94.57%**	**93.98%**	**83.27%**	**88.00%**	**83.73%**
Standar deviasi	**2.23%**	**1.68%**	**1.99%**	**2.27%**	**2.03%**	**2.62%**

Pada Tabel 4.3 dan Tabel 4.4, dapat dilihat nilai dari *precision*, *recall*, dan *f-measure* untuk tiap iterasi beserta nilai rata-ratanya untuk seluruh iterasi. Tabel 4.3 menyajikan hasil dari klasifikasi tanpa melewati tahap *stemming*, baik dengan menyertakan *stopwords* maupun tidak menyertakan *stopwords*. Nilai *precision*, *recall*, dan *f-measure* untuk pengujian tanpa melewati tahap *stemming* dan tidak menyertakan *stopwords* berturut-turut adalah 93.12%, 94.46%, dan 93.60%. Sedangkan hasil pengujian tanpa melewati tahap *stemming* dan menyertakan *stopwords* adalah *precision* sebesar 79.93%, *recall* sebesar 84.96%, dan *f-measure* sebesar 79.46%. Tabel 4.4 juga menyajikan pengujian hampir yang sama, yaitu dengan *stopwords* dan tanpa menyertakan *stopwords*, akan tetapi dilakukan *stemming* juga pada pengujian ini. Nilai *precision*,

recall dan *f-measure* dengan menyertakan *stopwords* adalah 93.68%, 94.57%, dan 93.98%. Sedangkan nilai *precision, recall* dan *f-measure* tanpa menyertakan *stopwords* adalah 83.27%, 88.00%, dan 83.73%. Setiap percobaan memiliki hasil yang cukup konsisten, akan tetapi konsistensi dari hasil percobaan tidak lebih baik dari *feature unigram*. Hal ini dapat dilihat dari nilai standar deviasi paling tinggi yang didapatkan adalah 2.99%.

4.3.3 Pengujian dengan *feature unigram* dan *bigram*

Pada skema pengujian ini, *feature* yang digunakan adalah *feature unigram* dan *bigram*. Pengujian juga dilakukan dengan melewati tahap *stemming*, tanpa melewati tahap *stemming*, dengan menyertakan *stopwords*, dan tanpa menyertakan *stopwords*. Hasil yang didapatkan untuk pengujian dengan menggunakan *stemming* dapat dilihat pada Tabel 4.5. Sedangkan hasil dari pengujian tanpa menggunakan *stemming* dapat dilihat pada Tabel 4.6.

Tabel 4.5 Hasil pengujian dengan *feature unigram* serta *bigram* tanpa *stemming*.

Iterasi	Tanpa Stopwords Removal			Dengan Stopwords Removal		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	93.13%	90.22%	91.35%	95.38%	94.75%	94.99%
2	94.86%	93.19%	93.80%	95.15%	95.83%	95.39%
3	93.22%	92.30%	92.69%	93.93%	95.18%	94.34%
4	95.45%	93.80%	94.47%	95.44%	95.84%	95.61%
5	95.98%	92.62%	94.06%	94.17%	95.36%	94.71%
6	95.25%	93.69%	94.40%	95.45%	96.33%	95.87%
7	95.09%	91.79%	93.22%	94.23%	95.36%	94.74%
8	92.77%	88.70%	90.22%	95.46%	95.73%	95.55%
9	95.17%	92.49%	93.57%	96.15%	96.57%	96.35%
10	96.06%	94.99%	95.45%	92.45%	93.91%	92.95%
11	95.72%	93.33%	94.09%	97.32%	97.75%	97.52%
12	94.78%	91.44%	92.90%	94.86%	93.60%	94.11%
13	94.72%	94.18%	94.38%	93.56%	95.59%	94.44%
14	97.32%	95.23%	96.13%	96.57%	97.39%	96.94%
15	93.46%	91.54%	92.39%	90.77%	92.76%	91.52%
16	96.88%	96.17%	96.41%	98.14%	98.46%	98.25%
17	95.34%	91.41%	92.85%	95.01%	95.83%	95.36%
18	95.86%	94.78%	95.20%	93.49%	94.08%	93.49%
19	95.02%	93.00%	93.87%	92.66%	93.23%	92.77%
20	95.87%	93.20%	94.27%	95.43%	95.36%	95.35%
21	97.04%	96.07%	96.52%	94.67%	95.83%	95.19%
22	96.44%	93.93%	95.02%	94.49%	94.07%	94.22%
23	95.44%	92.36%	93.66%	97.12%	96.54%	96.75%
24	94.86%	92.37%	93.40%	97.71%	97.62%	97.64%
25	94.62%	93.23%	93.88%	91.28%	93.47%	92.16%
Rata-rata	**95.21%**	**93.04%**	**93.93%**	**94.84%**	**95.46%**	**95.05%**
Standar deviasi	**1.15%**	**1.69%**	**1.43%**	**1.81%**	**1.46%**	**1.65%**

Tabel 4.6 Hasil pengujian dengan *feature unigram* serta *bigram* dan melewati *stemming*

Iterasi	Tanpa Stopwords Removal			Dengan Stopwords Removal		
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
1	92.02%	88.19%	89.64%	94.67%	94.04%	94.28%
2	94.10%	91.52%	92.47%	95.05%	95.34%	95.11%
3	93.22%	91.81%	92.45%	93.53%	94.82%	93.99%
4	95.45%	93.80%	94.47%	96.51%	97.39%	96.92%
5	95.29%	90.96%	92.76%	96.46%	96.43%	96.42%
6	94.87%	92.86%	93.77%	95.86%	96.19%	95.98%
7	94.07%	89.76%	91.54%	96.13%	95.59%	95.84%
8	91.78%	86.69%	88.62%	95.46%	95.75%	95.54%
9	94.62%	90.81%	92.24%	96.65%	97.02%	96.80%
10	94.99%	92.84%	93.72%	92.91%	93.92%	93.20%
11	95.83%	92.84%	93.74%	97.32%	97.75%	97.52%
12	93.62%	89.43%	91.17%	95.53%	93.94%	94.64%
13	93.63%	91.68%	92.48%	93.59%	95.12%	94.23%
14	97.35%	94.75%	95.88%	96.57%	97.39%	96.94%
15	92.42%	89.98%	91.03%	90.68%	92.87%	91.51%
16	96.22%	94.51%	95.22%	97.23%	97.62%	97.42%
17	95.89%	91.07%	92.54%	95.19%	95.46%	95.23%
18	95.52%	94.41%	94.85%	94.03%	94.43%	94.00%
19	95.03%	92.51%	93.62%	94.13%	93.94%	93.93%
20	95.56%	92.37%	93.65%	95.04%	95.01%	94.99%
21	96.02%	94.04%	94.94%	93.90%	94.99%	94.36%
22	95.81%	92.26%	93.71%	95.85%	95.25%	95.42%
23	94.55%	90.33%	91.94%	97.12%	96.54%	96.75%
24	95.11%	92.25%	93.47%	97.35%	96.79%	97.04%
25	93.60%	91.20%	92.26%	91.31%	93.12%	92.02%
Rata-rata	**94.66%**	**91.71%**	**92.89%**	**95.12%**	**95.47%**	**95.20%**
Standar deviasi	**1.34%**	**1.93%**	**1.66%**	**1.75%**	**1.38%**	**1.58%**

Pada Tabel 4.5 dan Tabel 4.6, dapat dilihat nilai dari *precision*, *recall*, dan *f-measure* untuk tiap iterasi beserta nilai rata-ratanya untuk seluruh iterasi. Tabel 4.5 menyajikan hasil dari klasifikasi tanpa melewati tahap *stemming*, baik dengan menyertakan *stopwords* maupun tidak menyertakan *stopwords*. Nilai *precision*, *recall*, dan *f-measure* untuk pengujian tanpa melewati tahap *stemming* dan tidak menyertakan *stopwords* berturut-turut adalah 95.21%, 93.04%, dan 93.93%. Sedangkan hasil pengujian tanpa melewati tahap *stemming* dan menyertakan *stopwords* adalah *precision* sebesar 94.84%, *recall* sebesar 95.46%, dan *f-measure* sebesar 95.05%. Tabel 4.6 juga menyajikan pengujian hampir yang sama, yaitu dengan *stopwords* dan tanpa menyertakan *stopwords*, akan tetapi dilakukan *stemming* juga pada pengujian ini. Nilai *precision*,

recall dan *f-measure* dengan menyertakan *stopwords* adalah 94.66%, 91.71%, dan 92.89%. Sedangkan nilai *precision, recall* dan *f-measure* tanpa menyertakan *stopwords* adalah 95.12%, 95.47%, dan 95.20%. Setiap percobaan memiliki hasil yang cukup konsisten, dimana nilai dari standar deviasi paling tinggi yang didapatkan adalah 1.66%.

4.3.4 Analisis hasil pengujian

Berdasarkan berbagai skema pengujian yang telah dilakukan, dapat dilakukan analisa terhadap hasil dari pengujian. Analisa dilakukan pada bobot *feature*, akurasi dan ukuran *vocabulary* berdasarkan teknik *preprocessing* serta pemilihan jenis *feature*.

a. Bobot *feature* pada 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*. 10 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel 4.7, Tabel 4.8, Tabel 4.9, dan Tabel 4.10.

Tabel 4.7 *Feature unigram* tanpa *stemming* dan *stopwords removal* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.1270825	kesehatan	241.0430748	gili	290.8000024
rp	191.1953699	korban	232.627195	lombok	264.1514978
bank	190.7032159	gempa	168.469832	gempa	220.7168052
bantuan	181.4396991	lombok	150.3147809	pariwisata	207.1011543
lombok	176.1009986	palu	149.3517939	wisatawan	203.8486565
bencana	175.2007744	rsud	144.2374253	arief	158.3730062
gempa	166.8704348	pasien	123.5745097	hotel	144.8582907
debitur	155.1529411	rs	121.4884034	trawangan	122.6445649
miliar	145.8427944	pmi	120.3294733	rinjani	111.6392347
kredit	136.1750993	tim	119.6683376	pendakian	105.2862693

Tabel 4.8 *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	304.7509219	jantung	369.9599978	pantai	243.6843136
rp	291.1955798	anda	343.1345668	gili	227.2638674
harga	265.4417459	kanker	266.2844775	wisata	223.991637
banjir	233.9265923	serangan	254.4250816	pengunjung	218.8732451
tol	227.7198308	seks	245.6974378	gunung	214.8632987
rokok	223.1448338	pneumonia	236.0462055	kamu	196.421652
jiwasraya	196.0795295	penyakit	223.7589572	pulau	194.9913499
inflasi	161.0291482	tubuh	217.479655	candi	190
investasi	152.9760165	pria	205.8637888	wisatawan	178.90515
as	148.3174195	wanita	205.5237109	air	174.3351265

Tabel 4.9 *Feature bigram* tanpa *stemming* dan *stopwords removal* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.1781573	korban gempa	122.8708326	gili trawangan	116.0507711
rp triliun	88.87016019	rumah sakit	99.16081029	gunung rinjani	92.89560346
sri mulyani	88.32696723	lombok utara	85.09940946	gili air	90.98090168
di palu	86.8704646	kementerian kesehatan	82.44171456	di lombok	83.45195403
di lombok	82.63379762	rsud tanjung	76.60416712	di gili	76.51596746
rp juta	79.62400924	gempa lombok	64.39938389	jalur pendakian	65.40736316
kantor cabang	74.9119314	para korban	62.41140068	gempa bumi	64.2120098
palu dan	70.91513323	gempa dan	59.31687961	lombok barat	60.74092847
sulawesi tengah	70.50874369	dan tsunami	59.01599157	arief yahya	60.56083678
gempa di	58.83231308	bpjs kesehatan	49.5474302	lombok utara	58.83415963

Tabel 4.10 *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	170.6086016	serangan jantung	221.9110599	kompas com	108.6543891
rp triliun	146.2420358	dikutip dari	90.65783148	gili trawangan	106.8194598
menjadi rp	115.7668687	bunuh diri	88.87394998	gunung rinjani	105.0786334
harga rokok	109.3912856	nyeri dada	88	gili air	90.98090168
rp ribu	92.94730555	getah bening	81.83029962	di bali	76.51596746
rp per	90.16904931	hal ini	80.56847873	kembang api	75.68636236
dollar as	82.61982715	kelenjar getah	79.68258049	di sini	72.82302159
dari rp	76.13007164	orang yang	73.54269953	di sana	72.39320343
rp menjadi	69.19803043	paru paru	69.30853216	berada di	71.87425881
pasar modal	65.66071467	kanker kelenjar	68.95686272	tempat wisata	70.90600039

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.

b. Bobot *feature* pada pengujian tanpa *stemming* dan 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*. 10 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel 4.11, Tabel 4.12, Tabel 4.13, dan Tabel 4.14.

Tabel 4.11 *Feature unigram* tanpa *stemming* dan *stopwords* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.1270825	kesehatan	241.0430748	gili	290.8000024
rp	191.1953699	korban	232.627195	lombok	264.1514978
bank	190.7032159	gempa	168.469832	gempa	220.7168052
bantuan	181.4396991	lombok	150.3147809	pariwisata	207.1011543
lombok	176.1009986	palu	149.3517939	wisatawan	203.8486565
bencana	175.2007744	rsud	144.2374253	arief	158.3730062
gempa	166.8704348	pasien	123.5745097	hotel	144.8582907
debitur	155.1529411	rs	121.4884034	trawangan	122.6445649
miliar	145.8427944	pmi	120.3294733	rinjani	111.6392347
kredit	136.1750993	tim	119.6683376	pendakian	105.2862693

Tabel 4.12 *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	304.7509219	jantung	369.9599978	pantai	243.6843136
rp	291.1955798	kanker	266.2844775	gili	227.2638674
harga	265.4417459	serangan	254.4250816	wisata	223.991637
banjir	233.9265923	seks	245.6974378	pengunjung	218.8732451
tol	227.7198308	pneumonia	236.0462055	gunung	214.8632987
rokok	223.1448338	penyakit	223.7589572	pulau	194.9913499
jiwasraya	196.0795295	tubuh	217.479655	candi	190
inflasi	161.0291482	pria	205.8637888	wisatawan	178.90515
investasi	152.9760165	wanita	205.5237109	air	174.3351265
as	148.3174195	kesehatan	187.5551302	turis	172.080392

Tabel 4.13 *Feature bigram* tanpa *stemming* dan *stopwords* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.1781573	korban gempa	123.8537992	gili trawangan	116.0507711
rp triliun	88.87016019	rumah sakit	99.16081029	gunung rinjani	92.89560346
sri mulyani	88.32696723	lombok utara	84.70635674	gili air	90.98090168
rp juta	79.62400924	kementerian kesehatan	82.44171456	jalur pendakian	65.40736316
kantor cabang	74.9119314	rsud tanjung	76.60416712	gempa bumi	64.2120098
sulawesi tengah	70.50874369	gempa lombok	64.32981289	lombok barat	60.74092847
gempa lombok	64.32981289	gempa tsunami	57.63165363	arief yahya	60.56083678
rumah rusak	61.17777482	bpjs kesehatan	49.5474302	pasca gempa	59.38692185
bank btn	55.2247199	krisis kesehatan	49.14816492	kunjungan wisatawan	58.72232468
korban bencana	54.26189294	tanjung lombok	49.03747284	lombok utara	58.56241947

Tabel 4.14 *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	170.6086016	serangan jantung	221.9110599	kompas com	108.6543891
rp triliun	146.2420358	nyeri dada	90	gili trawangan	108.1382185
harga rokok	111.2773422	getah bening	81.83029962	gunung rinjani	105.0786334
rp ribu	92.94730555	kelenjar getah	79.68258049	gili air	90.98090168
dollar as	82.61982715	paru paru	71.1324409	kembang api	75.68636236
ribu rp	69.54026025	kanker kelenjar	68.95686272	kaum nudis	62.12780988
pasar modal	65.66071467	rumah sakit	66.44796566	jalur pendakian	56.80113117
sungai ciliwung	64.88551875	berat badan	62.86189016	bunga bangkai	56.67837009
dki jakarta	64.79155033	serangan panik	62.34644023	pendakian gunung	55.83127984
rp rp	63.17176157	ria irawan	60.81039038	air terjun	55.67228054

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.

c. Bobot *feature* pada pengujian dengan *stemming* dan 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*. 10 *feature* dengan bobot tertinggi untuk masing-masing jenis *feature* dan kategori dapat dilihat pada Tabel 4.15, Tabel 4.16, Tabel 4.17, dan Tabel 4.18.

Tabel 4.15 *Feature unigram* dengan *stemming* dan *stopwords* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.1270825	korban	231.9128795	gili	290.8000024
rp	191.1953699	sehat	227.37213	lombok	264.1514978
bank	190.7032159	gempa	168.469832	gempa	221.7830699
lombok	176.1009986	lombok	150.3147809	pariwisata	211.1681634
bencana	175.9129727	palu	149.3517939	wisatawan	203.8486565
gempa	166.8704348	rsud	144.2374253	arief	158.3730062
debitur	155.1529411	pasien	125.21284	hotel	150.8617848
bantu	147.2767851	rs	121.4884034	daki	135.1039396
miliar	146.8216051	pmi	120.3294733	trawangan	122.6445649
kredit	139.608625	tim	119.6683376	evakuasi	113.5151812

Tabel 4.16 *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	310.3426819	jantung	373.9438409	pantai	248.1281967
rp	291.1955798	anda	338.9979172	wisata	227.4941982
harga	256.8221993	sakit	298.4332136	gili	227.2638674
banjir	233.905361	kanker	271.0395575	gunung	220.4745817
tol	225.1506374	serang	249.5083854	unjung	216.5683441
rokok	210.6514743	seks	247.4353815	pulau	208.7633095
naik	201.6387916	tubuh	240.4613529	kamu	196.421652
jiwasraya	196.0795295	pneumonia	239.091963	candi	190
normalisasi	175.5671327	sehat	231.6854224	air	188.5398993
kerja	163.5525773	pria	205.8637888	libur	187.8870337

Tabel 4.17 *Feature bigram* dengan *stemming* dan *stopwords* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
rp miliar	157.1781573	korban gempa	122.8708326	gili trawangan	116.0507711
rp triliun	88.87016019	rumah sakit	100.1830867	gunung rinjani	92.89560346
sri mulyani	88.32696723	menteri sehat	86.38746263	gili air	90.98090168
di palu	86.8704646	lombok utara	85.09940946	di lombok	83.45195403
di lombok	82.63379762	rsud tanjung	76.60416712	di gili	76.51596746
rp juta	77.73889056	gempa lombok	64.39938389	menteri pariwisata	73.52245559
kantor cabang	74.9119314	para korban	62.41140068	jalur daki	65.40736316
menteri uang	73.18780941	dan tsunami	59.01599157	gempa bumi	64.2120098
palu dan	70.91513323	gempa dan	59.01599157	lombok barat	60.74092847
sulawesi tengah	70.50874369	layan sehat	54.3494689	arief yahya	60.56083678

Tabel 4.18 *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	170.6086016	serang jantung	221.9110599	kompas com	108.6543891
rp triliun	146.2420358	orang yang	101.2455661	gili trawangan	106.8194598
jadi rp	118.2445794	kutip dari	90.37777806	gunung rinjani	105.0786334
harga rokok	109.3912856	bunuh diri	88.87394998	gili air	90.98090168
rp ribu	92.32795922	nyeri dada	88	ada di	88.11440266
rp per	90.16904931	getah bening	81.83029962	kembang api	78.2092411
dollar as	82.61982715	hal ini	80.77959669	di bal	76.51596746
dari rp	76.13007164	kelenjar getah	79.68258049	di sini	73.95728743
rp jadi	68	paru paru	73.63108071	di sana	72.39320343
pasar modal	65.66071467	kanker kelenjar	68.95686272	salah satu	72.03866828

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* yang mengalami perubahan bobot akibat *stemming* dapat dilihat pada Tabel 4.19.

Tabel 4.19 *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 4.20.

Tabel 4.20 *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

d. Bobot *feature* pada 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 4.21, Tabel 4.22, Tabel 4.23, dan Tabel 4.24.

Tabel 4.21 *Feature unigram* dengan *stemming* dan *stopwords removal* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	*Feature*	Bobot
palu	201.1270825	korban	231.9128795	gili	290.8000024
rp	191.1953699	sehat	227.37213	lombok	264.1514978
bank	190.7032159	gempa	168.469832	gempa	221.7830699
lombok	176.1009986	lombok	150.3147809	pariwisata	211.1681634
bencana	175.9129727	palu	149.3517939	wisatawan	203.8486565
gempa	166.8704348	rsud	144.2374253	arief	158.3730062
debitur	155.1529411	pasien	125.21284	hotel	150.8617848
bantu	147.2767851	rs	121.4884034	daki	135.1039396
miliar	146.8216051	pmi	120.3294733	trawangan	122.6445649
kredit	139.608625	tim	119.6683376	evakuasi	113.5151812

Tabel 4.22 *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	310.3426819	jantung	373.9438409	pantai	248.1281967
rp	291.1955798	sakit	298.4332136	wisata	227.4941982
harga	256.8221993	kanker	271.0395575	gili	227.2638674
banjir	233.905361	serang	249.5083854	gunung	220.4745817
tol	225.1506374	seks	247.4353815	unjung	216.5683441
rokok	210.6514743	tubuh	240.4613529	pulau	208.7633095
jiwasraya	196.0795295	pneumonia	239.091963	candi	190
normalisasi	175.5671327	sehat	231.6854224	air	188.5398993
investasi	163.521205	pria	205.8637888	libur	187.8870337
inflasi	161.0291482	wanita	205.5725489	daki	182.0966143

Tabel 4.23 *Feature bigram* dengan *stemming* dan *stopwords removal* pada artikel gempa.

Ekonomi Gempa		**Kesehatan Gempa**		**Pariwisata Gempa**	
Feature	Bobot	*Feature*	Bobot	Bobot	*Feature*
rp miliar	157.178157	korban gempa	123.853799	gili trawangan	116.050771
rp triliun	88.8701602	rumah sakit	100.183087	gunung rinjani	92.8956035
sri mulyani	88.3269672	menteri sehat	86.3874626	gili air	90.9809017
rp juta	77.7388906	lombok utara	84.7063567	menteri pariwisata	73.5224556
kantor cabang	74.9119314	rsud tanjung	76.6041671	jalur daki	65.4073632
menteri uang	73.18780941	gempa lombok	64.32981289	gempa bumi	64.2120098
sulawesi tengah	70.50874369	gempa tsunami	57.63165363	lombok barat	60.74092847
dampak gempa	64.83561033	layan sehat	54.3494689	arief yahya	60.56083678
gempa lombok	64.32981289	tenaga sehat	50.19635955	kunjung wisatawan	59.96647808
rumah rusak	63.7038388	bpjs sehat	49.5474302	pasca gempa	59.3869219

Tabel 4.24 *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	170.6086016	serang jantung	221.9110599	kompas com	108.6543891
rp triliun	146.2420358	nyeri dada	90	gili trawangan	106.8194598
harga rokok	111.2773422	getah bening	81.83029962	gunung rinjani	105.0786334
rp ribu	92.32795922	kelenjar getah	79.68258049	gili air	90.98090168
dollar as	82.61982715	paru paru	75.42696073	daki gunung	78.43181179
ribu rp	66.72986987	kanker kelenjar	71.11176468	kembang api	78.2092411
pasar modal	65.66071467	hubung seks	68.24344066	kaum nudis	62.12780988
sungai ciliwung	64.88551875	rumah sakit	66.44796566	jalur daki	60.24362397
dki jakarta	64.79155033	berat badan	64.56086016	air terjun	57.46816056
rp rp	63.17176157	serang panik	62.34644023	bunga bangkai	56.67837009

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 tetapi dianggap bersebelahan dikarenakan penghilangan *stopwords* memiliki bobot yang lebih tinggi dibanding *feature* yang sama namun berbeda imbuhan. Proses ini tidak menghilangkan pelanggaran terhadap konsep *bigram*.

BAB V
PENUTUP

5.1 Kesimpulan

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 terdapat *feature* yang sebenarnya tidak bersebelahan tapi dianggap bersebelahan akibat dari penghilangan *stopwords*.
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*.

5.2 Saran

Berdasarkan hasil penelitian yang sudah didapatkan kemudian terdapat beberapa catatan saran untuk dapat diperbaiki serta dikembangkan pada penelitian serupa selanjutnya yaitu :

1. Melakukan *feature selection* pada model agar *feature* yang *noise* dapat dihilangkan sehingga dapat meningkatkan performa serta akurasi.
2. Menambah *feature trigram* untuk mendapatkan hasil yang lebih bervariasi sehingga dapat menjadi pertimbangan dalam menentukan jenis *feature* yang digunakan.

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] A. A. N. Hidayat, “Ini Data Lengkap Kerusakan Gempa Lombok Versi BNPB,” Tempo.co, 10 September 2018. [Online]. Available: https://bisnis.tempo.co/read/1125319/ini-data-lengkap-kerusakan-gempa-lombok-versi-bnpb. [Diakses 19 November 2019].

[3] A. Prasetia, “Update Data BNPB: 2.113 Orang Meninggal Akibat Gempa Sulteng,” detikNews, 21 Oktober 2018. [Online]. Available: https://news.detik.com/berita/d-4265923/update-data-bnpb-2113-orang-meninggal-akibat-gempa-sulteng. [Diakses 19 November 2019].

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.