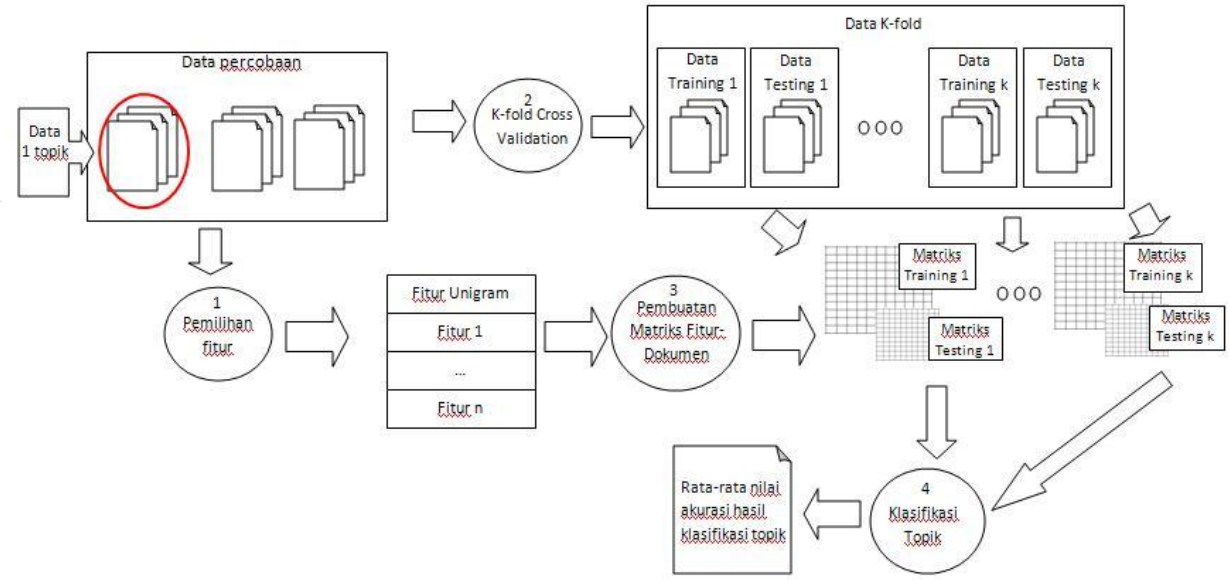
BAB 3 PERANCANGAN

Pada bab ini dijelaskan perancangan untuk melakukan klasifikasi topik pada artikel media massa dan abstrak tulisan ilmiah. Klasifikasi topik dilakukan dengan mengelompokkan dokumen ke dalam salah satu topik yang ada. Perancangan klasifikasi topik ini meliputi persiapan data, penentuan variabel percobaan, dan perancangan klasifikasi topik menggunakan *machine learning*.

3.1 Gambaran umum proses klasifikasi topik

Pada tugas akhir ini klasifikasi topik akan dilakukan dengan pendekatan *machine learning*. Pada percobaan yang sudah dilakukan (Franky, 2008), percobaan tersebut melakukan suatu analisis sentimen pada bahasa Indonesia. Percobaan tersebut pun memakai pendekatan *machine learning*. Dengan melihat hasil yang telah dilakukan, percobaan ini kemudian melakukan pendekatan yang sama untuk klasifikasi topik. Percobaan ini juga dilakukan untuk melihat pengaruh pendekatan *machine learning* ke data berbahasa Indonesia.

Pendekatan *machine learning* ini memakai beberapa metode, yaitu Naïve Bayes, Naïve Bayes Multinomial, dan Maximum Entropy. Semua metode yang dipakai ini merupakan pendekatan *supervised learning* yaitu pendekatan dengan terlebih dahulu melakukan pembelajaran untuk kemudian melakukan *testing*. Data yang dipakai adalah artikel media massa berbahasa Indonesia dan abstrak tulisan ilmiah. Artikel media massa akan didapat dari internet sedangkan abstrak tulisan ilmiah akan diambil dari basis data sistem Lontar yang dimiliki oleh perpustakaan fasilkom. Artikel media massa yang akan digunakan sebagai *input* sebelumnya akan dilakukan proses randomisasi. Hal ini perlu dilakukan untuk menghindari pemusatan data karena hal yang dibahas pada beberapa artikel sama. Kedua data yang akan dipakai ini kemudian akan ditentukan terlebih dahulu data tersebut masuk ke dalam suatu topik. Fitur yang digunakan untuk *machine learning* ini adalah *unigram language model* berdasarkan kata-kata yang dipilih dari data yang ada, dengan variasi pemilihan top-n fitur berdasarkan frekuensinya.



Gambar 3. 1 Alur klasifikasi topik dengan *machine learning*

Klasifikasi topik ini dilakukan dengan menggunakan *k-fold cross validation* untuk tiap metode pada *machine learning* yang dipakai. *K-fold cross validation* ini membagi data menjadi k bagian dan masing-masing bagian tersebut akan secara bergantian digunakan sebagai data *training* ataupun sebagai data *testing*. Nilai tiap fitur dari suatu dokumen dari data akan disimpan dalam sebuah matriks pasangan fitur-dokumen. Pembuatan matriks pasangan ini akan dilakukan untuk setiap variasi data *training* dan data *testing*, kemudian matriks ini akan menjadi *input* untuk diolah oleh berbagai metode *machine learning* yang dipakai. Nilai akurasi dari percobaan ini akan dilihat dari rata-rata nilai akurasi yang dihasilkan pada tiap k bagian data yang diujikan. Gambaran umum mengenai alur proses dari percobaan dapat dilihat pada gambar 3.1.

3.2 Data

Dalam melakukan percobaan untuk klasifikasi topik ini akan dipakai dua data yang berbeda. Data pertama adalah artikel dari internet. Artikel tersebut mengandung sebuah topik yang sudah diketahui sebelumnya. Data kedua adalah abstrak tulisan ilmiah yang didapatkan dari perpustakaan. Data ini pun dari awal sudah diketahui jenis ataupun topik yang dikandung dari data tersebut. Percobaan ini memakai dua data yang cukup berbeda untuk melihat keterkaitan topik yang terkandung dalam data. Data pertama akan dipisahkan dengan topik yang hanya

memiliki sedikit keterkaitan, sedangkan untuk data yang kedua terkandung keterkaitan yang cukup tinggi di bidang komputer. Kedua data ini akan diperlakukan secara sama yaitu keduanya nanti akan diolah dengan menggunakan berbagai variasi metode *machine learning*.

3.2.1. Persiapan Data

Data yang digunakan untuk percobaan ini berupa artikel dan abstrak tulisan ilmiah. Artikel media massa didapat dari www.kompas.com. Adapun dari *website* ini diambil beberapa jenis artikel, topik yang dikandung dari artikel tersebut sudah diketahui sebelumnya. Topik dari artikel tersebut sudah ditentukan oleh pihak *website*. Data ini dipilih karena kemudahan akses yang didapat dan juga pemilihan bahasa yang ada dalam artikel tersebut merupakan bahasa yang resmi. Beberapa topik yang dipakai dalam penggunaan artikel ini adalah ekonomi, kesehatan, olahraga, properti, dan travel. Artikel-artikel yang didapat dianggap akan memiliki kata-kata khusus yang menandakan bahwa artikel tersebut termasuk ke dalam suatu topik. Selain itu, lima topik ini dipilih karena dianggap tingkat keterkaitan dari topik tersebut rendah sehingga timbul asumsi bahwa tingkat akurasi juga akan lebih tinggi dalam menentukan klasifikasi topik.

Data yang kedua merupakan abstrak tulisan ilmiah dari basis data pada sistem Lontar pada fakultas ilmu Komputer Universitas Indonesia. Data ini dipilih dikarenakan data ini termasuk ke dalam dunia ataupun topik computer. Akan tetapi, sebenarnya dalam dunia komputer pun, kita masih bisa memilah-milah topik tersebut ke dalam jenis-jenis yang lebih spesifik. Dalam hal ini, abstrak tulisan ilmiah yang akan dipakai dibagi ke dalam tiga topik, yaitu *Information Retrieval* (IR), citra, dan Rekayasa Perangkat Lunak (RPL).

3.2.2 Analisis data

Sebelum kedua data dapat dipakai menjadi *input* dalam percobaan ini, kedua data tersebut haruslah diolah terlebih dahulu. Untuk artikel media massa, data tersebut diambil secara otomasi oleh sebuah program, karena itu haruslah dilihat terlebih dahulu apakah artikel media massa telah tersimpan dengan baik atau tidak. Setelah membuang data kosong, kemudian data tersebut akan dilakukan proses randomisasi. Proses randomisasi ini dirasa perlu untuk menghindari pemusatan

data atau topik yang sama untuk beberapa artikel. Pemusatan data ataupun topik dapat terjadi karena pengambilan data yang secara otomatis menyimpan data tersebut ke dalam sebuah *file* dengan menggunakan nama artikel tersebut dan tanggat artikel tersebut diterbitkan. Abstrak tulisan ilmiah tidak dilakukan proses randomisasi dikarenakan data tersebut diambil dari basis data sistem Lontar perpustakaan fasilkom dan disimpan ke dalam data teks dengan nama kode dari abstrak tulisan ilmiah tersebut.

Salah satu permasalahan yang didapat dalam melakukan percobaan ini adalah tidak adanya data dalam bahasa Indonesia yang sudah siap untuk dipakai. Ada pun beberapa data belum dilakukan pengelompokkan data, sehingga diperlukan pengolahan data terlebih dahulu. Penggolongan abstrak tulisan ilmiah dilakukan sendiri tanpa ada referensi. Secara subyektif, dengan pengetahuan yang dimiliki penulis abstrak tulisan ilmiah itu satu persatu dibaca dan digolongkan ke dalam beberapa topik untuk kemudian akan dipakai dalam percobaan.

3.3. Pemilihan Fitur

salah satu acuan yang dipakai dalam topik klasifikasi adalah pembuatan dalam sentimen analisis (Pang, Lee & Vaithyanathan, 2002). Pada paper tersebut, sentimen analisis yang dipakai adalah kata-kata dan tanda baca dalam *movie review*. Pemilihan fitur ini menggunakan konsep *n-gram language model*, dengan n menyatakan jumlah *token* berurutan pada dokumen. Percobaan ini memakai fitur yang sama dikarenakan apabila pada sentimen analisis, dokumen hanya dibagi ke dalam dua bagian maka pada klasifikasi topik ini, dokumen dibagi ke dalam beberapa bagian yang sama atau lebih banyak. Keduanya, baik sentimen analisis maupun klasifikasi topik, dianggap memiliki tujuan yang sama yaitu mengelompokkan dokumen menjadi beberapa bagian. Oleh karena itu, pada percobaan ini akan dilihat juga pengelompokkan dokumen dengan menggunakan fitur yang sama.

Manusia dapat menyimpulkan suatu dokumen termasuk ke dalam suatu kategori dengan melihat beberapa kata yang ada pada dokumen tersebut. Terdapat beberapa kata khusus yang hanya digunakan untuk beberapa topik, karena itu

dipikirkan bahwa dengan memakai fitur ini pada *machine learning* dapat membuat *machine learning* bekerja dengan lebih baik.

Pemilihan fitur ini dilakukan dengan menghitung terlebih dahulu kata-kata yang muncul pada data dokumen. Setelah itu, data tersebut akan diurutkan sesuai dengan frekuensi kemunculan data. Kata yang diurutkan itu hanyalah kata yang unik. Selain itu, nilai minimum frekuensi kata juga akan ditetapkan untuk menangani masalah waktu dan komputasi. Dengan adanya nilai minimum frekuensi kata ini juga akan membuat kita menghiraukan kata-kata yang jarang muncul dan tidak memiliki makna yang berarti. Implementasi untuk pemilihan fitur ini akan dijelaskan lebih lanjut pada bab 4.2.

Pada percobaan ini, fitur yang digunakan adalah fitur 1-*gram* atau lebih sering disebut sebagai *unigram*. Fitur ini membuat data yang diambil hanya satu *token* per-fiturnya. Pemilihan fitur ini menganggap tiap kata memiliki makna yang berarti yang dapat mempengaruhi penilaian dalam pengelompokkan dokumen ke dalam suatu topik. Selain itu, berdasarkan paper ini (Pang, Lee & Vaithyanathan, 2002), *unigram* memberikan hasil yang terbaik daripada n-*gram* lainnya. Percobaan yang ada pada tugas akhir ini akan menggunakan *unigram* dengan beberapa variasi. Hal ini dilakukan untuk melihat variasi percobaan mana yang akan memberikan nilai akurasi yang lebih baik. Variasi tidak hanya dilakukan variasi pada saat pemilihan semua fitur tetapi juga saat penggunaan fitur itu sendiri apakah akan mengambil semua fitur atau hanya top-n fitur.

3.4 K-*fold Cross Validation*

Pada percobaan ini , dalam melakukan percobaan dipakai teknik k-*fold cross validation*. K-*fold cross validation* ini merupakan teknik yang membagi data ke dalamn k bagian untuk kemudian masing-masng bagian data tersebut akan dilakukan proses klasifikasi topik. Dengan menggunakan k-*fold cross validation* akan dilakukan percobaan sebanyak k buah. Tiap percobaan itu akan menggunakan satu buah data *testing* dan k-1 bagian menjadi data *training*, dan kemudian data *testing* tersebut akan ditukar dengan satu buah data *training* sehingga untuk tiap percobaan akan didapatkan data *testing* yang berbeda-beda.

Data *training* adalah data yang akan dipakai dalam melakukan pembelajaran untuk melakukan klasifikasi topik, sedangkan data *testing* adalah data yang belum pernah dipakai sebagai pembelajaran dan akan berfungsi sebagai data yang akan digunakan untuk pengujian kebenaran ataupun akurasi dari hasil pembelajaran tersebut. Pada percobaan ini, akan dilakukan dua cara yang berbeda yaitu untuk percobaan satu akan dipakai data dengan jumlah yang sama banyaknya untuk tiap topik, sedangkan untuk data dua akan dipakai data yang jumlahnya berbeda-beda dalam tiap topik. Kedua hal ini kemudian akan dilihat apakah perbedaan jumlah dalam tiap topik akan mempengaruhi penilaian dalam klasifikasi topik.

3.5. Matriks Pasangan Fitur Dokumen

Dalam pembuatan data *training* dan data *testing*, *input* yang dipakai untuk *machine learning* adalah matriks pasangan fitur dan dokumen. Matriks ini merepresentasikan kemunculan fitur-fitur yang dipakai dalam keseluruhan dokumen. Baris dari matriks tersebut merupakan data dokumen, sedangkan kolom dari matriks ini merupakan fitur yang dipakai. Dengan matriks ini, dokumen dianggap sebagai vektor kumpulan fitur-fitur yang ada, atau dengan kata lain :

$$Di= [f_{i1}\ f_{i2}\ f_{i3}\ \dots\ f_{ij}]$$

Dengan D_i menyatakan dokumen ke-i (i merupakan nomor dokumen) dan f_{ij} adalah kemunculan fitur ke j untuk dokumen ke i (j adalah nomor fitur yang dipakai). Representasi nilai dari fitur yang ada pada dokumen terdapat dua macam, yaitu sebagai ada atau tidaknya fitur dan yang kedua adalah sebagai frekuensi kemunculan. Apabila kita memakai nilai fitur sebagai ada atau tidaknya fitur, maka nilai dari fitur adalah berupa binaer,yaitu nilainya akan 1 apabila fitur tersebut terdapat pada dokumen dan nilainya akan 0 apabila fitur tersebut tidak ada pada dokumen. Selain itu, fitur pada suatu dokumen dapat bernilai n dengan n adalah nilai frekuensi kemunculan fitur pada dokumen. Representasi dari kemunculan fitur pada dokumen dapat dilihat pada gambar 3.2.

	f_1	f_2	.	.	.	f_j
d_1	f_{11}	f_{12}	.	.	.	f_{1j}
d_2	f_{21}	f_{22}				f_{2j}
.	.	.				.
.	.	.				.
.	.	.				.
d_i	f_{i1}	f_{i2}	.	.	.	f_{ij}

Gambar 3. 2 Matriks Pasangan Fitur-Dokumen

Normalisasi dapat diberikan pada nilai fitur tersebut apabila nilai yang dipakai pada percobaan adalah frekuensi kemunculan fitur pada dokumen dan bukan ada atau tidaknya fitur pada dokumen. Pemberian normalisasi dapat dilakukan dengan membagi nilai dari fitur tersebut dengan jumlah *token* yang ada pada dokumen tersebut, atau dengan kata lain

$$F_{ij} = f_{ij}/\sum_k f_{ik}$$

, dengan f_{ij} adalah nilai frekuensi fitur j pada dokumen ke i dan f_{ik} adalah jumlah seluruh frekuensi fitur pada dokumen.

Pembuatan matriks fitur dokumen ini akan dilakukan dengan variasi dari pemilihan fitur dan dokumen untuk data yang akan diklasifikasi. Matriks ini dibentuk dengan menghitung nilai dari masing-masing fitur pada dokumen yang akan diklasifikasi, bergantung pada informasi nilai fitur yang digunakan. Hasil dari pembuatan matriks ini kemudian akan berbentuk menjadi data *training* dan data *testing* yang akan dipakai sebagi *input* untuk menjalankan beberapa metode *machine learning*. Untuk tiap percobaan pada metode *machine learning* akan dibentuk data *training* dari k-1 bagian data dan satu bagian sisanya sebagai data *testing* untuk validasi akurasi klasifikasi topik.

3.6 Metode Klasifikasi Topik

Pada tugas akhir ini akan dipakai dua metode *machine learning*, yaitu Naïve Bayes dan Maximum Entropy. Klasifikasi dokumen ke dalam beberapa topik bergantung dengan jenis data yang dipakai. Pada subbab ini dijelaskan perancangan dari penerapan kedua metode yang digunakan untuk melakukan klasifikasi tersebut. Matriks yang telah dihasilkan pada subbab 3.5 akan

digunakan sebagai *input* dari metode untuk menghasilkan model sebagai dasar untuk melakukan klasifikasi topik.

3.6.1 Naïve Bayes

Klasifikasi topik ini dilakukan dengan terlebih dahulu menentukan topik yang terkandung t Є T {IR, citra, RPL} dari dokumen yang dimiliki d Є D = $\{d_1,d_2,d_3,..,d_n\}$ berdasarkan fitur-fitur yang dimiliki pada dokumen. Dokumen d merupakan kumpulan vektor yang menyatakan nilai fitur sesuai dengan subbab 3.5. Topik dari dokumen dihitung dengan menggunakan persamaan yang ada pada subbab 2.3.1 dengan dokumen d sebagai konteks dan t sebagai kelas, sehingga menjadi :

$$T^* = \underset{t\in\{IR,citra,RPL\}}{\operatorname{argmax}} p(t|d) = \underset{t\in\{IR,citra,RPL\}}{\operatorname{argmax}} \prod_j p(f_j|t)\, x\, p(t)$$

Dimana f_j merupakan fitur-fitur dari dokumen d yang ingin diketahui topiknya, berupa fitur-fitur yang telah disebutkan pada subbab 3.3 dan fitur-fitur tersebut diasumsikan saling independen satu dengan yang lain. Nilai probabilitas p(fj|t) dipelajari dari data *training* dan menggunakan fij Є $\mathbb{R}^+$ U {0} dengan variasi informasi fitur yang ada pada subbab 3.5. Dengan menggunakan Naïve Bayes, nilai dari T* didapat dengan menggunakan persamaan:

$$T^* = \underset{t\in\{IR,citra,RPL\}}{\operatorname{argmax}} \prod_j p(f_j|t)^{f_{ij}} x\, p(t)$$

3.6.2 Maximum Entropy

Dengan menggunakan metode Maximum Entropy, proses klasifikasi dilakukan dengan hanya memakai informasi kemunculan(*presence*) dari suatu fitur dalam suatu dokumen, berbeda dengan Naïve Bayes yang dapat menggunakan tiga variasi fitur, yaitu *presence, frequency,* dan *frequency normalized*. Hal ini berhubungan dengan fungsi untuk fitur yang digunakan yaitu dengan menggunakan f_{ij}Є {0,1}. Secara garis besar, metode Maximum Entropy mencari distribusi probabilitas yang paling seragam dengan memakai asumsi minimal. Pada tugas akhir ini, pencarian probabiltas menggunakan model parametrik dengan GIS.

$$p^*(a,b) = \pi \prod_{j=1}^{n} \alpha_j{}^{f_j(a,b)}$$

Probabilitas p(a,b) menyatakan probabilitas kemunculan suatu dokumen b sebagai bagian dari topik a sebagai kelasnya pada distribusi dengan *entropy* maksimum. Kelas aЄ A = {IR, citra, RPL} dari dokumen b Є B= {b1, b2,…,bq} dihitung pada distribusi probabilitas $p^*(a,b)$. Himpunan kelas yang dimiliki oleh A tidak terbatas pada tiga topik tersebut tetapi dapat lebih ataupun berkurang, bergantung topik yang akan dimiliki pada data yang sedang dipakai. Dokumen b direpresentasikan sebagai vektor dari kemunculan (*presence*) fitur-fitur $f_j(a,b)$ yang didapat dari pemilihan fitur. Fungsi $f_j(a,b)$ untuk fitur ke-j dengan topik a pada dokumen b, dapat dinyatakan sebagai:

$$f_j(a,b) = \begin{cases} 1, \text{jika } f_j \text{ muncul di dokumen b pada kelas a} \\ 0, \text{jika } f_j \text{tidak muncul di dokumen b pada kelas a} \end{cases}$$

Penentuan topik dari dokumen n didapatkan dengan :

$$a^* = \underset{a \in \{IR, citra, RPL\}}{\operatorname{argmax}} \; p(a,b)$$

Bab 4 IMPLEMENTASI

Pada bab ini dijelaskan secara rinci penerapan dari perancangan yang telah dilakukan untuk klasifikasi topik. Implementasi yang dijelaskan berupa persiapan data, proses pemilihan fitur, dan implementasi klasifikasi topik menggunakan metode *machine learning*. Implementasi persiapan data dan pemilihan fitur dilakukan dengan menggunakan PERL dan Java. Sementara, klasifikasi topik dengan *machine learning* dilakukan dengan menggunakan *tools* dan *library* yang sudah tersedia, yang dibuat dalam bahasa Java. Hasil yang didapat setelah melakukan implementasi persiapan data dan pemilihan fitur adalah *input* data berupa data *training* dan data *testing* yang sesuai dengan masing-masing *tools*.

4.1 Persiapan data

Persiapan data dilakukan dengan cara mencari data yang akan dipakai untuk percobaan. Data ini sebelumnya akan diolah terlebih dahulu sebelum nantinya siap digunakan. Persiapan data ini dilakukan dengan menggunakan perl dan java dalam prosesnya.

4.1.1. Pengambilan data

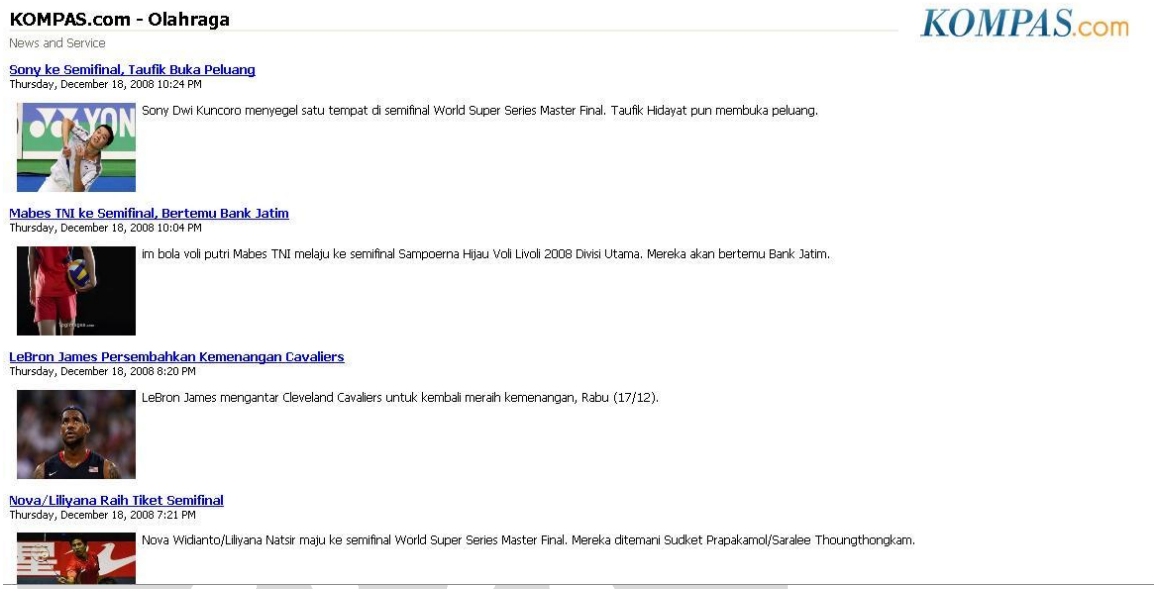
Data yang akan dipakai pada tugas ini terbagi menjadi dua macam, yaitu artikel media massa dan data basis data sistem Lontar. Pada artikel media massa, data ini diambil secara otomatis oleh crawler. Kerangka umum dari program ini dapat dilihat pada *pseudocode* yang ditampilkan pada gambar 4.1. Data ini disimpan dengan nama *file* yang terbentuk dari kombinasi tanggal dari artikel dan judul dari artikel tersebut. Sebagai contoh, penamaan kompas-2008-09-21-20350937-tarif-referensi-baru-premi-kendaraan-bermotor ini memberikan arti bahwa artikel tersebut diambil dari sumber Kompas dengan judul “tarif referensi baru premi kendaraan bermotor” dan artikel tersebut diterbitkan pada tanggal 21 September 2008. Angka 20350937 merupakan jam diterbitkannya artikel tersebut. Penamaan ini dilakukan untuk menghindari disimpannya artikel yang sama lebih dari satu kali. Data yang akan dipakai sudah terbagi-bagi menjadi beberapa topik, yaitu ekonomi, kesehatan, olahraga, travel, dan properti. Artikel media massa ini sudah

dilakukan proses klasifikasi oleh sumber dari diambilnya artikel. Akan tetapi, perlu dilakukan tahap pengolahan untuk dapat digunakan pada percobaan. Hal ini akan dijelaskan pada subbab 4.1.2.

```
function Main
        setPath(place to save the article);
        get All LinkArticle(link);

function get All LinkArticle(link)
        setConnection;
        if connect
            linksAtPages <- get All link with regex
            for each link in linksAtPages
                getArticle(link)
            end
        else give notice
        end

function getArticle(link)
        setConnection
        if connect
                title <- article_title
                date <- article_published_date
                context <- article_content
                print outFile <- title , date, context
        else
                give notice
        end
```

Gambar 4. 1 pseudocode pengambilan artikel secara otomatis

Alur pengambilan artikel media massa ini dimulai dengan memasukkan *link* suatu kumpulan artikel media massa, contoh http://www.kompas.com/getrss/olahraga. *Link* tersebut akan memunculkan kumpulan link lain seperti pada gambar 4.2.



Gambar 4. 2 Tampilan yang ditunjukkan oleh *link* http://www.kompas.com/getrss/olahraga.

Setelah itu, crawler akan menuju halaman yang dirujuk oleh semua *link* yang ada pada halaman seperti gambar 4.2. Halaman yang ditunjukkan oleh salah satu *link* yang ada pada gambar 4.2 dapat dilihat pada gambar 4.3. Pada gambar tersebut dapat dilihat judul dan tanggal publikasi yang nantinya akan dipakai pada penamaan *file*. Data tersebut kemudian akan disimpan dengan penamaan yang sudah ditentukan sebelumnya. Satu *link* seperti http://www.kompas.com/getrss/olahraga menunjukkan kumpulan *link* untuk satu topik. Oleh karena itu, semua halaman yang berhubungan dengan *link* seperti ini akan disimpan dalam satu folder sesuai dengan topiknya.

Gambar 4. 3 Tampilan halaman dari salah satu link yang dirujuk pada halaman http://www.kompas.com/getrss/olahraga.

Selain artikel media massa, data yang akan dipakai adalah data yang didapat dari basis data Lontar. Lontar adalah sistem yang dipakai oleh Fakultas Ilmu Komputer yang menyimpan semua data mengenai buku dari perpustakaan. Data yang akan dipakai merupakan abstrak tulisan ilmiah dari buku-buku ataupun karya ilmiah yang ada di perpustakaan. Abstrak tulisan ilmiah ini belum mengalami proses klasifikasi. Oleh karena itu, penulis harus melakukan proses klasifikasi secara manual.

Pada proses klasifikasi data ini, abstrak tulisan ilmiah terbagi menjadi beberapa topik, seperti basis data, *information retrieval* (IR), citra, rekayasa perangkat lunak(RPL), dan sebagainya. Akan tetapi, topik yang akan digunakan hanyalah tiga yaitu IR, citra, dan RPL. Pemilihan topik ini berdasarkan pemikiran mengenai keterbatasan dari segi waktu dan komputasi. Selain itu, abstrak tulisan ilmiah yang dimiliki tidak secara merata terbagi sehingga hanya beberapa topik itulah yang dipakai.

4.1.2. Prapemrosesan data

Data-data yang sudah diperoleh akan melalui proses pengolahan sebelum nantinya akan dipakai dalam percobaan. Secara garis besar baik artikel media massa maupun data yang diperoleh dari basis data lontar akan melalui proses pengolahan yang sama. Akan tetapi, artikel media massa akan dilakukan proses randomisasi terlebih dahulu sebelum nantinya data tersebut akan melalui proses yang sama dengan data dari basis data lontar.

```
function Random(first_data) returns new_documents
        id ← 0
        for each document in first_data
                random_number←do-random-from-range-0-to-1000
                new_document ←random_number-id.txt
                copy content of document to new_document
                id++
        end
return(All new_document)
```

Gambar 4. 4 *pseudocode* untuk melakukan proses randomisasi artikel

Proses randomisasi pada artikel media massa dilakukan untuk menghindari pemusatan data pada beberapa artikel. Hal ini mungkin terjadi apabila pada waktu tertentu suatu peristiwa sedang terjadi dan peristiwa ini ramai dibicarakan. Oleh karena itu,proses randomisasi perlu dilakukan. Kemudian data hasil randomisasi akan dibagi menjadi beberapa bagian. Mungkin dengan adanya randomisasi ini kemungkinan pemusatan data masih tetap ada, tetapi setidaknya kemungkinan terjadinya pemusatan itu sudah menurun.

Selain proses randomisasi, data yang akan digunakan juga dilihat terlebih dahulu apakah isi dari artikel sudah benar atau tidak. Kebenaran yang akan dilihat pada data ini adalah apakah isi *file* yang akan dilibatkan dalam proses percobaan tidaklah kosong ataupun hanya terdiri dari judul. Hal ini dikarenakan penulis

memakai *regular expression* (regex) dalam pengambilan data dan terkadang dari pihak Kompas mengganti pola untuk suatu artikel yang dimunculkan. Apabila *file* seperti ini masih ada dan dilakukan percobaan, hal ini dapat mempengaruhi hasil dari percobaan ataupun dapat memperburuk hasil dari klasifikasi. Kemudian, setelah dilakukan proses randomisasi data tersebut akan melalui proses pengolahan seperti pembagian *fold* data yang akan dibahas pada subbab 4.1.3.

Artikel media massa yang digunakan dapat dibagi menjadi dua yaitu data kecil dan data besar. Data kecil yang dimaksud adalah data dengan jumlah yang sedikit dan data besar adalah data dengan jumlah besar. Untuk data kecil, jumlah data yang dipakai untuk percobaan memiliki kisaran 92-185 artikel, sedangkan jumlah data yang dipakai untuk data besar berkisar 132-364 artikel.

Abstrak tulisan ilmiah yang diambil dari basis data sistem Lontar harus dilakukan proses klasifikasi secara manual sebelum dapat digunakan. Proses klasifikasi ini dilakukan secara manual berdasarkan pengetahuan yang dimiliki oleh penulis. Pada proses klasifikasi ini, abstrak tulisan ilmiah tersebut dapat terbagi menjadi beberapa topik, seperti Rekayasa Perangkat lunak (RPL), *Information Retrieval* (IR), citra, jarkom, basis data, dan *Knowledge Management* (KM). Akan tetapi, tidak semua data dari berbagai topik tersebut akan digunakan. Hal ini diakibatkan kurangnya jumlah data dari suatu topik untuk dapat digunakan pada percobaan. Oleh karena itu, akhirnya data yang terpakai adalah data dengan topik RPL, IR, dan citra.

Abstrak tulisan ilmiah yang dipakai pada percobaan ini memiliki jumlah yang berbeda-beda untuk tiap topiknya. Jumlah data yang dipakai untuk topik citra adalah 108 data, untuk topik IR adalah 93, dan untuk RPL adalah 150. Hal ini diakibatkan karena jumlah abstrak tulisan ilmiah yang tidak seimbang untuk tiap topik. Selain itu, jumlah data untuk tiga topik yang akan dibandingkan dengan data pada topik citra, IR dan RPL juga akan disesuaikan dengan jumlah yang ada.

Jumlah data yang dipakai, baik artikel media massa maupun abstrak tulisan ilmiah, merupakan kelipatan tiga. Hal ini dilakukan untuk mempermudah proses selanjutnya, yaitu pembagian *fold* data. *Fold* data yang akan dipakai pada

percobaan ini adalah tiga, sehingga dengan memiliki jumlah data dengan kelipatan tiga, data akan terbagi merata ke *fold-fold* yang ada. Pembahasan lebih lanjut mengenai cara membagi data ke beberapa *fold* akan dijelaskan pada subbab selanjutnya.

4.1.3 Pembagian *Fold* Data

Dalam proses klasifikasi pada tugas akhir ini, digunakan 3-*fold cross validation*. Penggunaan *cross validation* ini pun terdiri dari dua macam, yaitu pembagian data yang sama untuk tiap topik yang ada dan pembagian data yang berbeda-beda untuk tiap topik. Hal ini dilakukan untuk melihat perbedaan jumlah data tiap topik dengan nilai akurasi dari percobaan yang dihasilkan. Seperti yang telah dijelaskan sebelumnya, percobaan ini dapat dibagi menjadi dua yaitu data kecil dan data besar. Proses pembagian data yang seragam dan tidak seragam untuk tiap topik juga akan dilakukan ke dua bagian data tersebut, baik data kecil maupun data besar.

Berikut merupakan pembagian data yang dilakukan pada data kecil dan data untuk tiap topik tidak seragam :

1. Jumlah data untuk topik ekonomi : 186. Pembagian data untuk tiap *fold*-nya adalah 62.
2. Jumlah data untuk topik kesehatan : 162. Pembagian data untuk tiap *fold*-nya adalah 54.
3. Jumlah data untuk topik olahraga : 147. Pembagian data untuk tiap *fold*-nya adalah 49.
4. Jumlah data untuk topik properti : 108. Pembagian data untuk tiap *fold*-nya adalah 36.
5. Jumlah data untuk topik travel : 93. Pembagian data untuk tiap *fold*-nya adalah 31.

Berikut merupakan pembagian data yang dilakukan pada data besar dan data untuk tiap topik tidak seragam :

1. Jumlah data untuk topik ekonomi : 363. Pembagian data untuk tiap *fold*-nya adalah 121.

2. Jumlah data untuk topik kesehatan : 312. Pembagian data untuk tiap *fold*-nya adalah 104
3. Jumlah data untuk topik olahraga : 285. Pembagian data untuk tiap *fold*-nya adalah 95.
4. Jumlah data untuk topik properti : 141. Pembagian data untuk tiap *fold*-nya adalah 47.
5. Jumlah data untuk topik travel : 132. Pembagian data untuk tiap *fold*-nya adalah 44.

Jumlah artikel media massa untuk topik travel selalu merupakan nilai terkecil yang ada untuk jumlah pada semua topik. Hal ini dapat dilihat bahwa pada data kecil, jumlah data travel hanya 93 dan pada data besar , jumlah data yang dimilikinya hanya 132. Oleh karena itu, pada pembagian data yang seragam untuk tiap topik, jumlah data yang akan dipakai adalah jumlah data untuk topik travel.

Jumlah abstrak tulisan ilmiah yang dipakai tidak banyak. Jumlah data yang dipakai untuk topik citra adalah 108, sehingga data yang ada untuk tiap *fold* adalah 36. Untuk topik IR, data yang dipakai adalah 93 sehingga data akan dibagi menjadi 31 untuk tiap *fold*. Terakhir, untuk data RPL, data yang dipakai adalah 150 sehingga data akan dibagi menjadi 50 untuk tiap *fold*.

Abstrak tulisan ilmiah ini akan dipakai untuk melihat pengaruhan kemiripan data dengan nilai akurasi klasifikasi topik yang dihasilkan. Nilai yang dihasilkan dengan abstrak tulisan ilmiah ini akan dibandingkan dengan nilai yang dihasilkan oleh percobaan yang menggunakan tiga topik. Tiga topik yang akan digunakan adalah ekonomi, kesehatan, dan travel. Jumlah data yang akan digunakan tiga topik ini juga akan disesuaikan dengan abstrak tulisan ilmiah, yaitu 108 data untuk topik kesehatan, 93 data untuk olahraga, dan 150 data untuk data ekonomi. Pembagian data untuk *fold* pun disesuaikan sehingga dihasilkan 36 data untuk tiap *fold* pada topik kesehatan, 31 data untuk tiap *fold* pada topik olahraga, dan 50 data untuk tiap *fold* pada data ekonomi. Proses penyesuaian jumlah artikel media massa dengan jumlah abstrak tulisan ilmiah dilakukan agar proses perbandingan nilai akurasi dari kedua data tersebut dapat dibandingkan dengan adil.

4.2 Implementasi Pemilihan Fitur

Fitur yang digunakan untuk melakukan klasifikasi topik adalah fitur *unigram* berupa fitur yang terdiri dari satu *token* unik. percobaan yang dilakukan akan memiliki tiga variasi fitur, yaitu top-2000, top 5000, dan *All* features. Penjelasan lebih lanjut mengenai tiga variasi yang dipakai dapat dilihat pada tabel 4.1.

Tabel 4. 1 Variasi Pemilihan Fitur

Label fitur	Keterangan
Top-2000	Pemilihan 2000 fitur dengan nilai frekuensi tinggi.
Top-5000	Pemiihan 5000 fitur dengan nilai frekuensi tinggi.
All features	Pemilihan semua fitur yang ada pada semua dokumen dengan minimal frekuensi 4.

Tiga variasi fitur ini digunakan untuk melihat hubungan penggunaan fitur dengan nilai akurasi yang dihasilkan setelah dilakukan klasifikasi topik. Penggunaan nilai 2000 dan 5000 dianggap sebagai asumsi bahwa jumlah *token* tersebut sudah cukup signifikan untuk melakukan klasifikasi. *Pseudocode* untuk melakukan pemilihan fitur dapat dilihat pada gambar 4.5.

```
function featureSelect (All_topik_data, feature-used) returns feature-list
for each document in All_topik_data
        for each token in document
                token-hash(token)++
for each token in token-hash
        if token-hash(token) < 4
                remove token from token-hash
sort-descending-by-value(token-hash)
if feature-used == top-2000
        feature-list ← GETTOPN(token-hash, 2000)
else if feature-used == top-5000
        feature-list ← GETTOPN(token-hash, 5000)
```

```
else
        feature-list ← get-All-keys(token-hash)


return (feature-list)


function geTTopN(sorted-hash, n) return top-n-list
        for I ← 1 to n
                top-n-list[i] ← next-key(sorted-hash)
        return (top-n-list)
```

Gambar 4. 5 Pseudocode Pemilihan Fitur

Pemilihan fitur ini dilakukan untuk masing-masing variasi data yang akan digunakan. Apabila pada suatu percobaan akan dipakai artikel media massa dengan 5 topik dan jumlah data untuk tiap topik seragam, maka pemilihan *feature-list* akan dilakukan pada semua dokumen tersebut. Oleh karena itu, semua jenis data yang akan dipakai pada percobaan akan dilakukan pemilihan *feature-list*. pemilihan fitur ini dilakukan dengan cara menghitung dan mengurutkan semua *token* unik berdasarkan frekuensi kemunculan *token* pada dokumen yang digunakan. Proses perhitungan frekuensi kemunculan *token* pada seluruh dokumen dilakukan dengan menggunakan *token*-hash. Pada *token-hash* ini, kata *token* tersebut akan berperan sebagi *key* dan frekuensi kemunculan sebagai *value*. Frekuensi kemunculan *token* akan selalu bertambah setiap kali *token* tersebut ditemui pada dokumen. Batas minimum dari kemunculan *token* yang akan dipakai adalah 4, hal ini mengikuti nilai yang digunakan pada (Pang, Lee, & Vaithyanathan, 2002), dengan jumlah *token* yang tidak dibatasi. *Token* yang memiliki frekuensi kemunculan kurang dari 4 akan dihapus dari *token-hash*. Kemudian, nilai dari *feature-used* akan menentukan variasi fitur mana yang akan digunakan. setelah variasi fitur telah dilakukan, *token* yang dihasilkan akan disimpan dalam *feature-list*. *Token* yang ada pada *feature-list* diurutkan berdasarkan nilai dari frekuensi kemunculan *token*, dari yang tertinggi hingga terendah.

4.3 Pembuatan Matriks Pasangan Fitur-Dokumen

Dalam melakukan klasifikasi topik, *input* dari *machine learning* yang digunakan berupa matriks pasangan fitur-dokumen. Matriks ini digunakan sebagai dasar untuk membuat *input* data dengan format yang sesuai dengan *library/tools* dari *machine learning* yang digunakan. Matriks ini menyimpan data mengenai fitur yang telah dipilih sebelumnya pada subbab 4.2. Pembuatan matriks pasangan fitur-dokumen ini dilakukan untuk setiap variasi percobaan dengan variasi yang berbeda. Penjelasan secara mendalam mengenai matriks fitur-dokumen ini dan variasi dari informasi yang akan disimpan dapat dilihat pada subbab 3.5. Penjelasan singkat mengenai variasi informasi dari fitur yang akan disimpan dapat dilihat pada tabel 4.2.

Tabel 4. 2 Variasi informasi nilai fItur

Label	Keterangan
presence	Menyimpan informasi mengenai ada atau tidaknya suatu fitur dalam suatu dokumen
Frequency	Menyimpan informasi mengenai jumlah kemunculan suatu fitur dalam suatu dokumen.
Frequency-normalized	Menyimpan informasi mengenai nilai dari jumlah kemunculan suatu fitur dalam suatu dokumen dibagi dengan jumlah seluruh fitur yang ada pada dokumen tersebut.

Pada satu kali percobaan, dari data yang dipakai akan dibuat tiga matriks pasangan fitur-dokumen untuk tiap masing-masing variasi informasi nilai fitur. Hal ini dilakukan dengan mengikuti pembuatan *fold* data sehingga untuk masing-masing *fold* data akan terdapat matriks pasangan fitur dokumen berdasarkan variasi fitur yang digunakan. *Pseudocode* pembuatan matriks pasangan fitur-dokumen untuk satu *fold* dapat dilihat pada gambar 4.6.

```
function createFeatureDocumentMatrix (array-data, feature-list, feature-info, normalize) returns feature-document-matrix
for each data-based-on-topik in array-data
```

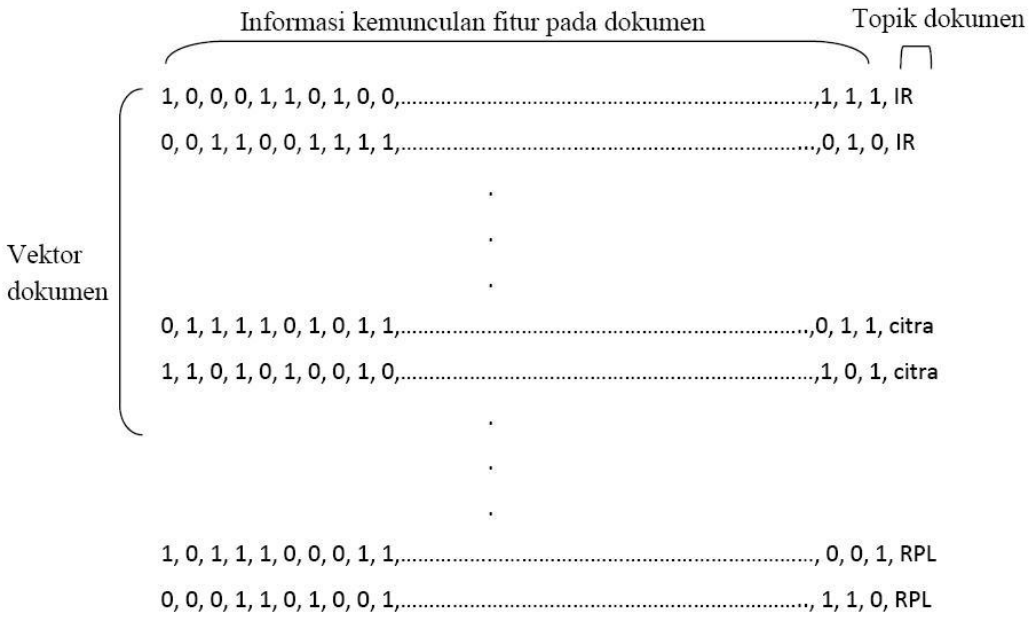
```
        Total-document ← total-document + count-documents(data-based-on-topik)
        Feature-sized ← count(feature-list)
          Document-matrix ← create matrix of size (total-document) x (feature-size + 1) with default
        value 0
        i ← 1
        for each data-based-on-topik in array-data
            for each document in data-based-on-topik
                j ← 1
                for each token in feature-list
                        if feature-list == frequency
                                document-matrix [i][j++] ← count-frequency (token, document)
                        else if feature-info == presence and exists(token, document)
                                document-matrix[i][j++] ← 1
                        document-matrix[i++][j] ← topik-name-i
                if normalize==true
                for I ← 1 to total-document
                        total-freq-in-a-document ← sum-column(document-matrix[i])
                        for j ← 1 to feature-size -1
                                document-matrix [i][j] ← document – matrix [i][j] / total-freq-in a-document

feature-document-matrix ← document-matrix
return (feature-document-matrix)
```

Gambar 4. 6 *Pseudocode* pembuatan matriks pasangan fitur-dokumen untuk satu *fold*

Pembuatan matriks dokumen ini didahului dengan menghitung seluruh dokumen yang ada pada data yang digunakan, kemudian dihitung juga jumlah dari fitur yang akan digunakan. Kedua hal ini diperlukan untuk membuat dasar matriks yang kosong terlebih dahulu, sehingga nilai fitur yang akan dihasilkan dapat diletakkan pada matriks tersebut. Informasi nilai fitur yang akan disimpan disesuaikan dengan *feature-info*. Nilai dari fitur tersebut akan dihitung untuk tiap *token* dan tiap dokumen yang digunakan. Informasi mengenai topik yang terkandung pada dokumen akan disimpan pada akhir dari vektor dokumen. Jika menggunakan variasi informasi normalisasi, maka nilai dari tiap sel yang menyatakan nilai fitur dari dokumen akan dibagi dengan jumlah *token* yang dimiliki oleh dokumen tersebut.

Matriks pasangan fitur dokumen ini disimpan dalam sebuah *file* .csv (*comma separated value*) dengan tiap barisnya menyatakan satu dokumen. Kolom terakhir dari *file* tersebut berisi informasi topik yang dimiliki oleh dokumen yang terlibat. Percobaan yang dilakukan melibatkan 8 topik dan pada kolom terakhir penamaan yang ada mengikuti nama topik yang terlibat, yaitu ekonomi, kesehatan, olahraga, kesehatan properti, travel, IR, citra, dan RPL. Penamaan ini diperlukan untuk data *training* dan data *testing*. Contoh format dari matriks pasangan fitur-dokumen hasil perhitungan denagn *presence* dapat dilihat pada gambar 4.7.

Dalam melakukan klasifikasi topik, metode Naïve Bayes dan Naïve Bayes Multinomial dapat dilakukan dengan menggunakan semua variasi fitur, tetapi untuk metode Maximum Entropy hanya dapat dilakukan dengan menggunakan fitur *presence*. Oleh karena itu, untuk suatu percobaan, metode ini hanya menghasilkan matriks yang memberikan informasi *presence* untuk fiturnya.



Gambar 4. 7 Format Penyimpanan matriks pasangan fitur-dokumen dengan informasi *presence*

4.4. implementasi klasifikasi topik dengan *machine learning*

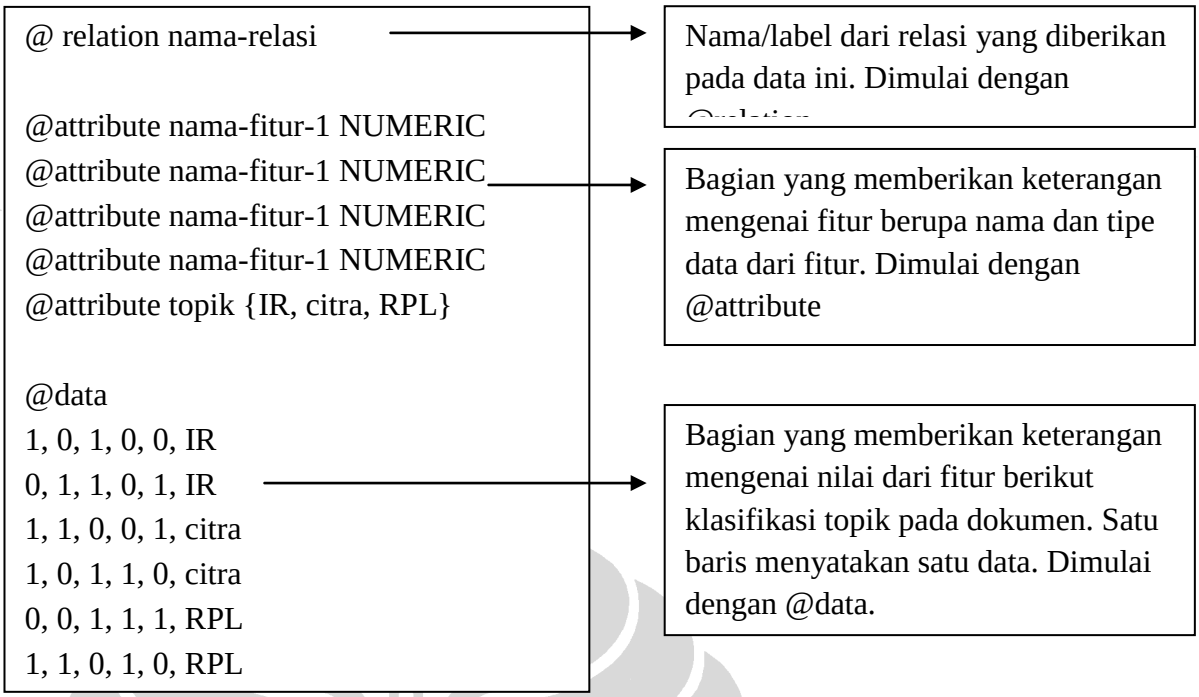
Implementasi klasifikasi topik dengan *machine learning* menggunakan dua *library/tools* yang berbeda untuk setiap metode. Masing-masing *library/tools* tersebut memiliki format *input* yang berbeda-beda. Oleh karena itu, matriks hasil dari perhitungan pada subbab 4.3 akan menjadi *input* untuk membuat pembuatan data *input* tersebut dilakukan untuk setiap data *training* dan data *testing* yang akan dipakai. Dengan memakai 3-*fold cross validation*, maka untuk masing-masing metode dengan variasi fitur yang digunakan akan dibuat tiga data jenis data *training* dan tiga data *testing*, dengan ketentuan sebagai berikut:

- Data *training* 1 yang dipakai adalah gabungan matriks *fold* 1 dan *fold* 2, sedangkan untuk data *testing*nya adalah matriks *fold* 3
- Data *training* 2 yang dipakai adalah gabungan matriks *fold* 1 dan *fold* 3, sedangkan untuk data *testing*nya adalah matriks *fold* 2
- Data *training* 3 yang dipakai adalah gabungan matriks *fold* 2 dan *fold* 3, sedangkan untuk data *testing*nya adalah matriks *fold* 1

Pada proses pembuatan data *training* dan data *testing* untuk setiap metode *machine learning* dan variasi fitur yang digunakan, format dari *input* yang digunakan akan disesuaikan dengan format *input* dari masing-masing library. Oleh karena itu, format *input* yang ada pada gambar 4.7 akan diubah mengikuti format *input* dari masing-masing library. Variasi fitur yang digunakan akan mengikuti tabel 4.1 dan variasi informasi dari fitur itu sendiri akan mengikuti tabel 4.2. Klasifikasi topik akan dilakukan pada setiap variasi percobaan dengan menggunakan variasi data, fitur, berikut info yang terkandung pada fitur yang dipakai, kecuali pada metode Maximum Entropy yang hanya menggunakan informasi *presence* dari fitur yang ada. Pada satu percobaan, nilai akurasi akan didapatkan dengan menghitung rata-rata dari nilai akurasi tiga percobaan yang telah dilakukan pada masing-masing data *training* dan data *testing*.

4.4.1 Analisis Sentiment dengan Naïve Bayes

Pada klasifikasi topik dengan menggunakan *machine learning*, implementasi untuk metode Naïve Bayes menggunakan *library* WEKA 3.5.7 yang didapat dari http://www.cs.waikato.ac.nz/~ml/weka/. WEKA adalah *library* yang dituliskan dalam bahasa java. Data yang akan menjadi *input* merupakan data ARFF yang merupakan format *input* dari *library* WEKA. Pada gambar 4.8 ditampilkan contoh format *input* dari data ARFF.



Gambar 4. 8 Format data ARFF

Proses klasifikasi ini dilakukan dengan terlebih dahulu mempersiapkan data *training* dan data *testing*. Matrix pasangan fitur-dokumen diproses lebih lanjut dengan menggabungkan matriks kombinasi tiga *fold* dan menyesuaikan matriks tersebut dengan format *input* ARFF. Penyesuaian format *input* ARFF cukup sederhana dengan hanya menambahkan informasi @relation, @attribute, dan @data. Keterangan nama-relasi diganti dengan topik_classification, dan nama-fitur diganti dengan nilai kemunculan fitur {1, 2, 3, ..., N} dengan tipe data numerik dimana nilai fitur tersebut dianggap sebagai *input* yang berkelanjutan. Atribut untuk topik diganti dengan topik dan nilainya terdiri atas tiga kelas yaitu iR, citra, dan RPL. Banyaknya kelas yang dipakai pada klasifikasi topik dapat lebih dari tiga. Hal ini bergantung pada pemakaian topik yang ingin diujikan.

Data dari matriks pasangan fitur-dokumen dimasukkan langsung di bawah @data, dengan satu baris menyatakan satu data atau satu dokumen. Proses pembuatan data ARFF ini dilakukan untuk semua data *training* dan data *testing*. Selain WEKA dapat menerima *input* data ARFF, WEKA juga dapat menerima *input* data CSV. Akan tetapi, proses klasifikasi dengan menggunakan data ARFF

menghabiskan waktu yang lebih sedikit dibanding dengan menggunakan data CSV. Oleh karena itu, data pada file .csv akan diubah bentuknya ke dalam format data ARFF, walau proses perubahan format ini juga cukup menghabiskan waktu dan sumber daya.

Klasifikasi topik dengan menggunakan naïve bayes dilakukan dengan membuat program dalam bahasa Java dan menggunakan library WEKA. Pseudocode program untuk melakukan klasifikasi dengan menggunakan data *training* dan data *testing* dapat dilihat pada gambar 4.9.

```
function naiveBayesTopikClassification (data-training, data-testing) returns accuracy
    Training-instances ← create instances of document from training data
    Test-instances ← create instances of document from testing data
    Naïve-bayes-classifier ← build-classifier (training-instances)
    Right-classification ← 0
    Test-instances-size ← count-instances(test-instances)
    For each instance in test-instances
        Instance-topik ← get-instance-topik(instace)
        Prob-dist ← get-prob-dist-for-instance(naïve-bayes-classifier, instance)
        Many_topiks ← count_many_number_in(prob_dist)
        Max ← prob_dist(topik_1)
        Max_number ← topik_1
        For I ← 2 to many_topiks
            If prob_dist(topik_i) > max
            Max_number ← topik_i
            Max ← prob_dist(topik_i)
            For I ←1 to many_topiks
                    If max_number == topik_i and instance_topik == topik_i
                    Right-classification++;
    Accuracy ← right-clasification / test-instance-size
```

```
Return(accuracy)
```

Gambar 4. 9 *Pseudocode* klasifikasi topik menggunakan metode Naïve Bayes

Klasifikasi topik dilakukan dengan membaca data ARFF untuk mendapat data *training* (dianggap sebagai data-*training*) dan data *testing* (dianggap sebagai data-*testing*). Metode Naive Bayes yang digunakan pada percobaan memakai dua variasi metode Naive Bayes, yaitu Naive Bayes dan Naive Bayes Multinomial. Kedua variasi metode Naive Bayes ini dinyatakan sebagai Naive Bayes *Classifier*, dengan mengimplementasikan Naive Bayes yang ada pada *library* WEKA, dengan menggunakan nilai *default* untuk semua parameter. Data *training* dan data *testing* yang dipakai dibaca sebagai sebuah kelas *instance* yaitu *training-instances* dan *test-instances*. Proses pembelajaran yang dilakukan oleh metode Naive Bayes ini dengan membangun terlebih dahulu model klasifikasi topik pada *build-classifier* (*training-instances)* berupa nilai $p(f_j|s)$ untuk semua fitur. Proses klasifikasi topik dilakukan dengan mencari nilai distribusi probabilitas p(a|b) untuk a Є {IR, citra, RPL}. Nilai dari a merupakan topik yang akan dipakai dalam sebuah percobaan dan tidak terbatas pada penggunaan tiga topik yang telah disebutkan. Berdasarkan model yang telah dibangun sebelumnya, distribusi probabilitas didapatkan dari suatu *instance* pada *test-instances* berupa prob-dist. Setelah itu, dicari nilai prob-dist untuk suatu kelas dari suatu *instance* yang menghasilkan nilai maksimum dan apabila kelas yang diprediksi tersebut sesuai dengan kelas yang telah ditetapkan sebelumnya, maka nilai kebenarannya, pada *pseudocode* tersebut dianggap sebagai *right-classification*, akan bertambah. Nilai akurasi akan didapatkan secara sederhana dengan membagi nilai prediksi yang benar dengan total data *testing*.

4.4.2 Klasifikasi Topik dengan Maximum Entropy

Pada klasifikasi topik dengan menggunakan *machine learning*, implementasi untuk metode Maximum Entropy menggunakan *library* OpenNLP Maxent v2.4.0 yang didapat dari http://maxent.sf.net. OpenNLP Maxent ini merupakan *library* yang dituliskan dalam bahasa java. Data yang akan menjadi *input* merupakan data

.txt yang merupakan format *input* dari *library* WEKA seperti yang ditunjukkan pada gambar 4.10.

```
1_0 1_2 IR
1_0 1_1 1_2 IR
1_1 1_3 1_4 citra
1_2 1_3 1_4 citra
1_0 1_4 1_5 RPL
1_1 1_4 1_5 RPL
```

Gambar 4. 10 Contoh format data untuk OpenNLP Maxent

Data yang ada pada matriks pasangan fitur-dokumen akan diproses lebih lanjut sehingga akn dihasilkan suatu file .txt yang tiap barisnya memiliki nilai fitur dengan format nilaiFitur_indeksFitur. Nilai fitur merupakan nilai yang menyimpan informasi kemunculan fitur dan indeks fitur merupakan nilai yang menyimpan nomor kolom dimana fitur itu muncul, dimulai dari kolom 0. Pada data *maxent*, fitur dengan nilai 0 tidak dimasukkan artinya apabila fitur tersebut tidak muncul, informasi tersebut tidak akan disimpan. Nilai fitur yang disimpan mengandung arti *presence*, yaitu muncul atau tidaknya suatu fitur pada dokumen. Kolom yang terakhirpada suatu baris akan menyimpan informasi mengenai topik dari dokumen tersebut.

Klasifikasi topik dilakukan dengan membuat program Java dengan memanfaatkan fungsi yang telah disediakan *library* OpenNLP Maxent. Gambar 4.11 merupakan *pseudocode* dari proses klasifikasi dokumen dengan menggunakan metode maximum entropy.

function maxentTopikCLassification (data-*training*, data-*testing*) returns accuracy

Maxent-model ← create maxaimum entropy model from data-trining using GIS

Right-classification ← 0

Data-*testing*-size ← count(data-*testing*)

for each document in test data

```
        Doc-topik ← get topik from current document in data-testing
        Best-outcome ← evaluate (model, document)
          if best-outcome == doc-topik
            Right-classification++
    Accuracy ← right-classification/data-testing-size
    Return (accuracy)
```

Gambar 4. 11 *Pseudocode* klasifikasi topik dengan menggunakan metode Maxent Entropy

Proses klasifikasi dilakukan dengan sebelumnya membuat suatu model dari data-*training* untuk pembelajarannya. Model ini dibuat dengan mencari nilai α_j untuk tiap fitur f_j ada data-*training* dengan menggunakan algoritma GIS. Model ini kemudian akan digunakan untuk mencari distribusi p(a|b) untuk a Є {IR, citra, RPL} dengan menggunakan fungsi evaluate(model, document). Nilai dari a merupakan topik yang akan dipakai dalam sebuah percobaan dan tidak terbatas pada penggunaan tiga topik yang telah disebutkan. Variabel *best-outcome* akan menyimpan kelas a yang memberikan nilai maksimum pada nilai probabilitas p(a|b). Apabila ternyata prediksi yang diberikan sama dengan topik yang seharusnya dimiliki oleh dokumen tersebut, maka nilai kebenarannya, *right-classification*, akan bertambah. Cara mendapatkan akurasi pada metode ini sama dengan cara mendapatkan akurasi pada metode Naïve Bayes, yaitu membagi nilai kebenarannya dengan total dokumen pada data *testing*.

BAB 5 HASIL DAN PEMBAHASAN

Pada bab ini diberikan hasil dari percobaan yang dilakukan dalam melakukan klasifikasi topik dengan menggunakan Naïve Bayes dan Maximum Entropy. Pembahasan dari hasil mencakup perbandingan antara metode *machine learning* yang digunakan serta pembahasan klasifikasi topik dengan variasi penggunaan fitur, nilai fitur, dan jenis data yang digunakan

5.1 Hasil Klasifikasi Topik

Hasil dari klasifikasi topik ini dibahas pada beberapa subbab. Pembagian penjelasan ini dilakukan berdasarkan aspek yang ingin dilihat dari percobaan tersebut. Selain itu, dengan pembagian ini diharapkan penjelasan juga akan lebih terarah dan mendalam. Beberapa aspek yang ingin dibahas pada bab ini adalah perbedaaan metode, jumlah *token*, informasi yang terkandung dalam fitur, dan jumlah data yang dipakai. Beberapa aspek ini telah dibahas sedikit pada bab-bab yang ada sebelumnya. Selain aspek tersebut, ditambahkan juga beberapa aspek lain seperti banyak topik pada data yang digunakan. Variasi *input* yang digunakan pada percobaan, dapat dilihat pada tabel 5.1.

Tabel 5. 1 Keterangan mengenai variabel *input* yang digunakan pada percobaan.

Variabel	Nilai
Metode	- Naïve Bayes - Naïve Bayes Multinomial - Maximum Entropy
Jumlah *token*	- *All tokens* - 2000 *tokens* - 5000 *tokens*
Informasi yang terkandung dalam fitur	- *Presence* - *Frequency* - *Frequency normalized* * metode Maximum Entropy hanya memakai nilai

	presence dari fitur
Banyak topik yang digunakan	2, 3, 4, dan 5
Keseragaman data untuk tiap topik	- Seragam (semua topik memiliki jumlah data yang sama) - Tidak seragam (tiap topik memiliki jumlahnya masing-masing)
Jumlah data	Data kecil dan data besar
Aspek kemiripan data	- Tingkat kemiripan data kecil (memakai topik ekonomi, kesehatan, dan olahraga) - Tingkat kemiripan data tinggi (memakai topik IR, citra, dan RPL)

Pada variabel informasi yang terkandung dalam fitur, metode Maximum Entropy hanya memakai nilai *presence* dari fitur yang ada. Akan tetapi, untuk metode Naïve Bayes dan Naïve Bayes Multinomial, kedua metode ini akan memakai seluruh informasi yang terkandung dalam fitur dalam melakukan percobaan.

Hasil yang dilihat pada semua percobaan adalah nilai akurasi setelah melakukan klasifikasi topik. Nilai akurasi ini didapat dengan menghitung rata-rata dari klasifikasi untuk tiap *fold*. Untuk tiap *fold*, nilai akurasi merupakan hasil pembagian jumlah dokumen yang terklasifikasi dengan benar dibagi dengan jumlah dokumen yang dipakai pada data *testing*.

5.2 Hasil Klasifikasi Topik dari Aspek Metode dan Fitur yang Digunakan

Percobaan yang dilakukan pada subbab ini bertujuan untuk membandingkan tiga metode yang ada. Metode-metode tersebut adalah Naïve Bayes, Naïve Bayes, dan Maximum Entropy. Ketiga metode ini akan dilihat dengan menggunakan variasi fitur dan informasi dari fitur. Percobaan ini memakai artikel media massa dengan tiga topik, yaitu ekonomi, kesehatan, dan olahraga. Data yang dipakai pun adalah data yang kecil dan seragam. Arti dari data yang kecil dan seragam tersebut

adalah data yang dipakai adalah data dengan jumlahnya berkisar antara 93-185 dan karena data ini adalah data seragam, maka artikel media massa untuk tiap topik adalah 93. Hasil percobaan dari aspek metode dan fitur yang digunakan dapat dilihat pada tabel 5.2.

Tabel 5. 2 Hasil nilai akurasi dari percobaan dengan menggunakan berbagai metode dan variasi pada jumlah *token* dan informasi yang terkandung dalam *token*.

Variasi Metode dan Fitur			*fold 1*	*fold 2*	*fold 3*	*average*
All tokens	Naïve Bayes	p	82.80	89.25	90.32	87.46
		f	90.32	91.40	82.80	88.17
		f-n	97.85	93.55	88.17	93.19
	Naïve Bayes Multinomial	p	68.82	79.57	76.34	74.91
		f	74.19	81.72	74.19	76.70
		f-n	43.01	49.46	40.86	44.44
	Maximum Entropy	p	88.17	89.25	87.1	88.17
5000 *tokens*	Naïve Bayes	p	81.72	86.02	90.32	86.02
		f	89.25	90.32	81.72	87.10
		f-n	92.47	95.70	89.25	92.47
	Naïve Bayes Multinomial	p	66.67	77.42	75.27	73.12
		f	68.82	72.04	72.04	70.97
		f-n	43.01	49.46	41.94	44.80
	Maximum Entropy	p	90.32	87.097	78.49	85.30
2000 *tokens*	Naïve Bayes	p	80.65	84.95	91.4	85.66
		f	88.17	89.25	80.65	86.02
		f-n	94.62	94.62	90.32	93.19
	Naïve Bayes Multinomial	p	69.89	79.57	76.34	75.27
		f	67.74	70.97	69.89	69.53

		f-n	43.01	49.46	43.01	45.16
	Maximum Entropy	p	77.42	81.72	61.29	73.48

Pada tabel 5.2, kolom paling kiri pada tabel menjelaskan jumlah *token* yang dipakai untuk melakukan percobaan dan nilai pada kolom selanjutnya merupakan jenis metode yang dipakai. Kolom ketiga merupakan variasi dari informasi yang terkandung pada *token* yang dipakai, nilai p menyatakan *presence*, f menyatakan *frequency*, dan f-n menyatakan *frequency-normalized*. Empat kolom paling kanan merupakan nilai akurasi dari proses klasifikasi topik. Apabila nilai yang ditampilkan adalah 90,32, artinya adalah 90,32% dari data *testing* telah diprediksi topiknya dengan benar. Tiap baris menyatakan satu kali melakukan percobaan, kotak dengan warna merah muda menyatakan nilai terbesar dari nilai akurasi yang dihasilkan untuk tiga *fold* dalam satu kali percobaan. Kotak dengan warna biru menyatakan bahwa nilai yang ada pada kotak tersebut merupakan nilai terbesar yang dihasilkan dari percobaan dengan menggunakan berbagai metode, variasi informasi pada fitur dan satu jenis untuk jumlah *token*.

5.2.1 Hasil Klasifikasi dari Aspek Metode

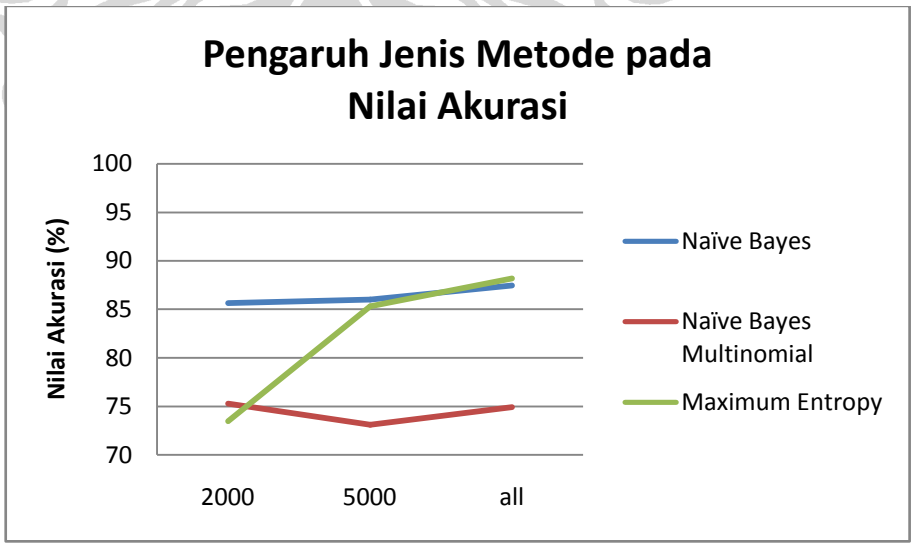
Pada subbab ini, analisis hasil klasifikasi dilakukan dengan melihat jenis metode yang digunakan. Analisis ini dilakukan dengan melihat hasil percobaan yang ditampilkan pada tabel 5.2 dan akan lebih melihat rata-rata hasil akurasi, bukan nilai akurasi tiap *fold*.

Tabel 5. 3 Hasil akurasi berbagai metode dengan variasi jumlah *tokens* dan memakai informasi *presence* untuk nilai fiturnya

Metode	*Tokens*		
	All	5000	2000
Naïve Bayes	87.46	86.02	85.66
Naïve Bayes Multinomial	74.91	73.12	75.27

Maximum Entropy	88.17	85.3	73.48

Tabel 5.3 menggambarkan nilai akurasi untuk tiga metode dengan berbagai variasi jumlah *token* tetapi hanya memakai satu jenis informasi pada nilai fiturnya, yaitu nilai *presence*. Dari tabel tersebut, tidak dapat ditentukan satu metode yang lebih baik daripada metode yang lain. Hal ini dikarenakan saat memakai semua *token*, metode Maximum Entropy menghasilkan nilai akurasi tertinggi tetapi untuk jumlah *token* adalah 2000 dan 5000, metode Naïve Bayes lebih baik daripada metode yang lain. Akan tetapi, apabila perbandingan tersebut dengan jenis informasi nilai fitur yang lain, dimana metode Maximum Entropy tidak termasuk dalam perbandingan, dapat disimpulkan bahwa metode Naïve Bayes menghasilkan nilai akurasi yang lebih tinggi daripada metode Naïve Bayes Multinomial. Oleh karena itu, apabila hanya membandingkan metode Naïve Bayes dan Naïve Bayes Multinomial, maka dapat disimpulkan metode Naïve Bayes lebih baik dalam melakukan klasifikasi topik dibandingkan dengan menggunakan metode Naïve Bayes Multinomial. Pengaruh jenis metode yang dipilih pada nilai akurasi yang dihasilkan dapat dilihat lebih jelas pada gambar 5.1.



Gambar 5. 1 Grafik pengaruh pemilihan metode pada nilai akurasi yang dihasilkan.

Setelah membandingkan ketiga metode yang ada berdasarkan hasil yang didapat, maka masih belum dapat disimpulkan apakah metode Naïve Bayes merupakan yang terbaik atau metode Maximum Entropy yang merupakan metode terbaik. Akan tetapi, tentunya metode yang terbaik bukanlah metode Naïve Bayes Multinomial karena nilai akurasi yang dihasikan oleh metode Naïve Bayes lebih tinggi daripada nilai akurasi yang dihasilkan oleh metode Naïve Bayes Multinomial. Oleh karena itu, pada pembahasan selanjutnya akan lebih dibahas mengenai metode Naïve Bayes dan Maximum Entropy.

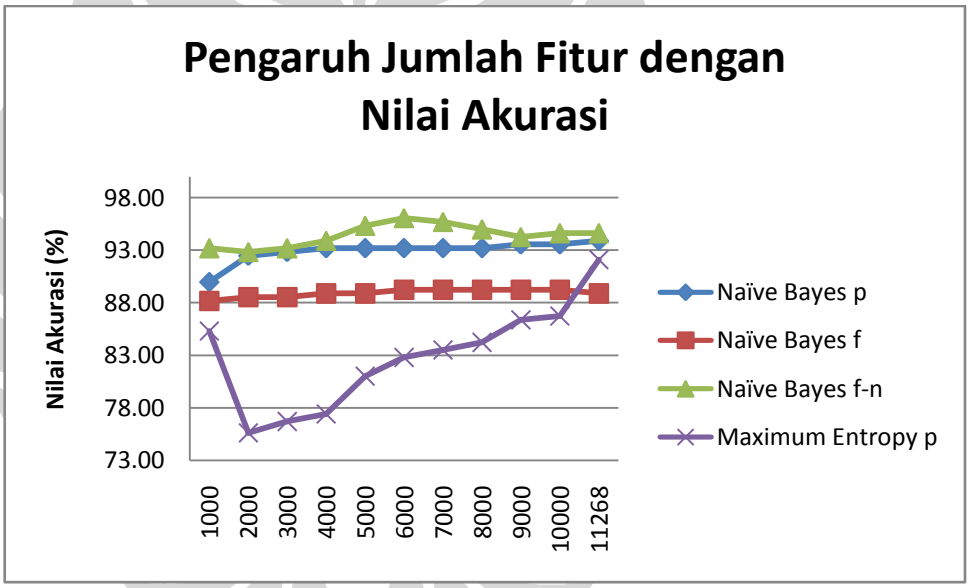
5.2.2 Hasil Klasifikasi dari Aspek Jumlah *token*

Pada subbab ini akan dilihat pengaruh pemakaian jumlah *token* pada nilai akurasi yang dihasilkan. Pemakaian fitur dibutuhkan dalam membangun pengetahuan tentang suatu topik. Oleh karena itu, berdasarkan kalimat tersebut, akan timbul suatu prediksi bahwa dengan memakai fitur yang lebih banyak, pengetahuan menjadi lebih banyak terkumpul, dan hasil dari akurasi juga akan menjadi lebih baik. Dengan memakai hipotesis seperti ini, hasil klasifikasi akan dibahas dan dianalisis dilihat dari aspek jumlah *token*.

Tabel 5. 4 Hasil nilai akurasi dari metode Naïve Bayes dan Maximum Entropy dengan menggunakan variasi jumlah *token* dan nilai informasi yang dimiliki fitur.

Jumlah *Token*	Naïve Bayes			Maximum Entropy
	p	f	f-n	p
1000	89.96	88.17	93.19	85.30
2000	92.47	88.53	92.83	75.63
3000	92.83	88.53	93.19	76.70
4000	93.19	88.89	93.91	77.42
5000	93.19	88.89	95.34	81.00
6000	93.19	89.25	96.06	82.80
7000	93.19	89.25	95.70	83.51
8000	93.19	89.25	94.98	84.23
9000	93.55	89.25	94.27	86.38
10000	93.55	89.25	94.62	86.74
All	93.91	88.89	94.62	92.11

Hipotesis yang telah diprediksi sebelumnya ternyata benar untuk metode Naïve Bayes dengan menggunakan informasi fitur *presence*. Pada percobaan ini, seiring dengan bertambahnya jumlah *token* yang digunakan akan memberikan nilai akurasi yang lebih baik. Sementara, pada percobaan Naïve Bayes dengan menggunakan informasi fitur *frequency* dan *frequency-normalized*, nilai akurasi yang diberikan akan naik hingga pada pertengahan jumlah *token* yang digunakan dan selanjutnya akan menurun. Lain halnya dengan menggunakan Maximum Entropy, apabila nilai akurasi yang diberikan pada saat menggunakan 1000 *token* diabaikan, maka perubahan nilai akurasi akan sesuai dengan hipotesis. Hal ini dapat dilihat bahwa nilai akurasi yang diberikan pada saat menggunakan 1000 *token* cukup tinggi dan selanjutnya akan turun saat menggunakan 2000 *token*.



Gambar 5. 2 Grafik pengaruh jumlah *token* yang digunakan dengan nilai akurasi yang dihasilkan pada beberapa percobaan.

Gambar 5.2 merupakan gambaran laju pengaruh jumlah *token* pada nilai akurasi yang dihasilkan. Pada gambar tersebut, dapat dilihat bahwa untuk metode Naïve Bayes, nilai akurasi yang diberikan cenderung akan naik seiring dengan bertambahnya jumlah *token*. Selain itu, nilai akurasi untuk metode Naïve Bayes dengan menggunakan informasi fitur *frequency* dan *frequency-normalized* akan naik hingga pada pertengahan keseluruhan jumlah *token*, yaitu 6000 *token* dari

11268 *token*. Selanjutnya nilai akurasi tersebut akan turun seiring dengan pertambahan jumlah *token*. Sementara, pada Maximum Entropy terlihat perubahan nilai akurasi yang janggal pada penggunaan 1000 *token* dan 2000 *token*. Nilai akurasi yang diberikan dengan menggunakan 1000 *token* begitu tinggi dan kemudian jatuh pada saat percobaan menggunakan 2000 *token*. Akan tetapi, apabila perubahan nilai ini diabaikan, maka nilai akurasi yang diberikan dari penggunaan 2000 *token* dan 11268 *token* akan naik. Berdasarkan fakta ini dan gambar 5.2, secara umum, dengan menggunakan jumlah *token* yang lebih banyak nilai akurasi yang dihasilkan juga cenderung lebih baik,.

5.2.3 Hasil Klasifikasi dari Aspek Informasi Fitur

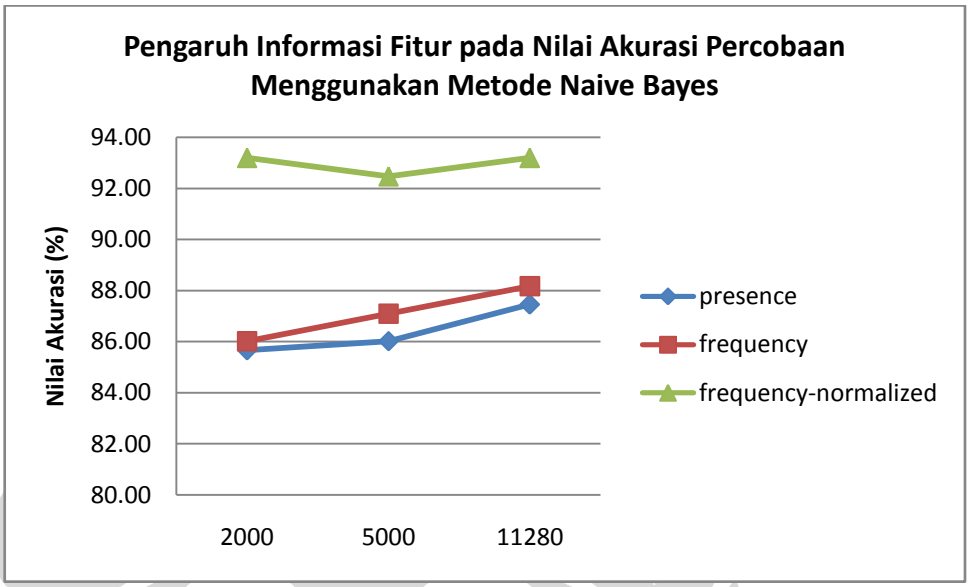
Pada subbab ini akan dibahas mengenai pengaruh jenis informasi yang dipakai dengan nilai akurasi yang dihasilkan. Pada tiga metode yang digunakan, diterapkan tiga variasi informasi fitur, yaitu *presence*, *frequency*, dan *frequency-normalized*. Akan tetapi, pada metode Maximum Entropy, hanya dapat digunakan informasi *presence*. Oleh karena itu, untuk melakukan perbandingan nilai akurasi dengan jenis informasi fitur yang digunakan, hanya akan dilihat pada percobaan dengan metode Naïve Bayes dan Naïve Bayes Multinomial.

Tabel 5. 5 Tabel Pengaruh informasi fitur pada nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes (keterangan: p adalah *presence*, f adalah *frequency*, dan f-n adalah *frequency-normalized*).

Informasi fitur	*Tokens*		
	2000	5000	*All*
p	85.66	86.02	87.46
f	86.02	87.10	88.17
f-n	93.19	92.47	93.19

Pada percobaan yang dilakukan dengan menggunakan metode Naïve Bayes, nilai akurasi yang dihasilkan dengan menggunakan informasi *frequency* lebih baik daripada nilai akurasi dengan menggunakan informasi *presence*. Bahkan nilai akurasi yang dihasilkan dengan menggunakan informasi *frequency-normalized* adalah yang terbaik di antara ketiga informasi fitur yang tersedia. Hal ini dapat

dilihat pada tabel 5.5 bahwa untuk setiap jumlah *token* yang digunakan, metode Naïve Bayes dengan informasi *frequency-normalized* untuk *token*nya menghasilkan nilai akurasi yang terbaik. Pernyataan ini dikuatkan dengan gambaran pengaruh informasi fitur pada metode Naïve Bayes pada gambar 5.3.



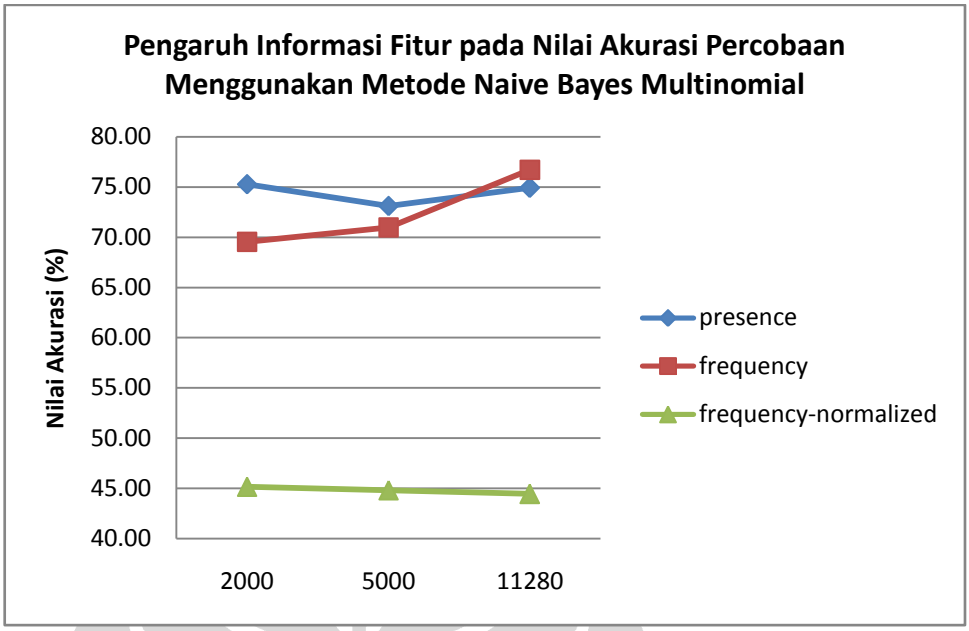
Gambar 5. 3 grafik pengaruh jenis informasi fitur pada nilai akurasi percobaan menggunakan metode Naïve Bayes.

Pada gambar 5.3, diketahui bahwa penggunaan jenis informasi fitur berpengaruh pada nilai akurasi percobaan yang menggunakan metode Naïve Bayes, tetapi belum dapat dipastikan penggunaan jenis informasi fitur ini juga akan berpengaruh pada percobaan dengan menggunakan metode yang lain. Oleh karena itu, selanjutnya akan dilihat pengaruh penggunaan informasi fitur pada percobaan dengan menggunakan metode Naïve Bates Multinomial. Tabel 5.6 merupakan hasil percobaan dengan menggunakan metode Naïve Bayes Multinomial.

Tabel 5. 6 Tabel pengaruh informasi fitur pada nilai akurasi yang dihasikan pada percobaan yang menggunakan metode Naïve Bayes Multinomial (keterangan: p adalah *presence*, f adalah *frequency*, dan f-n adalah *frequency-normalized*).

Informasi fitur	*Tokens*		
	2000	5000	*All*
p	75.27	73.12	74.91
f	69.53	70.97	76.70
f-n	45.16	44.80	44.44

Berbeda dengan nila akurasi yang dihasilkan pada tabel 5.5 yaitu pada metode Naïve Bayes, dengan menggunakan Naïve Bayes Multinomial, penggunaan informasi fitur *frequency-normalized* membuat nilai akurasi menjadi lebih buruk. Dapat dilihat pada tabel 5.6, nilai akurasi dari percobaan yang menggunakan metode Naïve Bayes Multinomial dengan informasi fitur *frequency-normalized* merupakan nilai akurasi yang paling buruk dibanding dengan nilai akurasi yang memakai informasi fitur lain untuk metode yang sama. Akan tetapi, tidak bisa dikatakan bahwa dengan menggunakan informasi fitur *presence* akan dapat memberikan nilai akurasi yang lebih baik dibandingkan menggunakan informasi fitur *frequency*, ataupun sebaliknya. Untuk lebih jelasnya dapat dilihat pada gambar 5.4.



Gambar 5. 4 grafik pengaruh jenis informasi fitur pada nilai akurasi percobaan menggunakan metode Naïve Bayes Multinomial.

Dapat dilihat pada gambar 5.4, informasi fitur yang dipakai tidak akan selalu memberikan pengaruh yang sama untuk tiap kondisi. Kondisi yang dimaksud pada kalimat tersebut adalah penggunaan jumlah data, jumlah *token*, dan metode yang dipakai sama yaitu metode Naïve Bayes Multinomial. Apabila pada metode Naïve Bayes, fitur *frequency-normalized* selalu memberikan nilai akurasi yang terbaik, maka tidak begitu halnya dengan metode Naïve Bayes Multinomial. Pada metode Naïve Bayes Multinomial ini, tidak dapat disimpulkan informasi fitur mana yang akan memberikan nilai akurasi yang lebih baik. Oleh karena itu, pengaruh pemilihan informasi fitur tergantung pada metode yang dipakai.

5.3 Hasil Klasifikasi dari Banyaknya Topik yang Digunakan

Percobaan yang dilakukan pada subbab 5.2 menunjukkan bahwa metode Naïve Bayes dan Maximum Entropy menghasilkan nilai akurasi yang lebih baik dibandingkan metode Naïve Bayes Multinomial. Oleh karena itu, pada percobaan yang akan dijelaskan pada subbab ini, hasil percobaan yang diberikan merupakan hasil percobaan dengan menggunakan metode Naïve Bayes dan Maximum Entropy. Pada metode Naïve Bayes, akan dipakai informasi fitur *frequency*

normalized karena telah terbukti menghasilkan akurasi yang terbaik untuk metode ini dan untuk Maximum Entropy memakai infomasi fitur *presence*. Jumlah *token* yang dipakai adalah 1000 hingga jumlah *token* yang dimiliki.

Pada subbab ini akan dilihat keterkaitan jumlah topik yang digunakan dengan nilai akurasi yang dihasilkan. Untuk artikel media massa, jumlah topik yang akan terlibat dimulai dari 2 hingga 5. Untuk 2 topik, topik yang digunakan adalah ekonomi dan kesehatan. Untuk 3 topik, digunakan topik ekonomi, kesehatan, dan olahraga, sedangkan untuk 4 topik digunakan topik ekonomi, kesehatan, olahraga, dan properti. Terakhir, untuk 5 topik, topik ekonomi, kesehatan, olahraga, properti, dan travel digunakan sebagai data. Selain itu, untuk abstrak tulisan ilmiah hanya akan dilihat pada 2 topik dan 3 topik. Hal ini dikarenakan kurangnya data yang dapat digunakan. Untuk 2 topik, topik yang digunakan adalah RPL dan citra, sedangkan untuk 3 topik, digunakan RPL, citra, dan IR.

Tabel 5. 7 Tabel nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes dengan informasi *frequency* normalized dan Maximum Entropy untuk perubahan jumlah topik dari 2 hingga 5 pada artikel media massa.

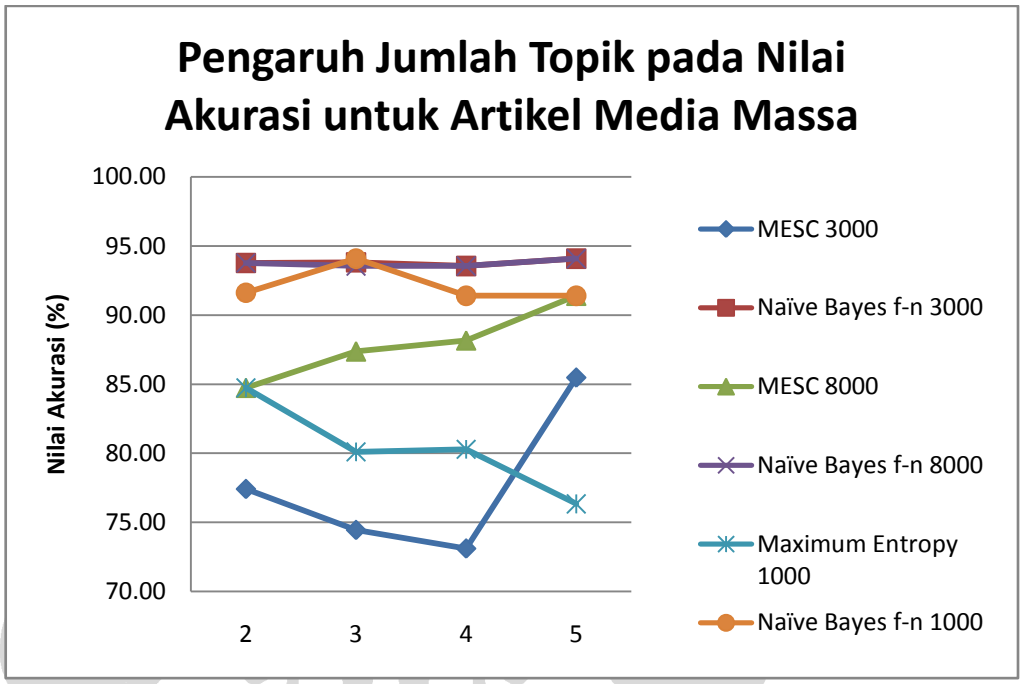
Jenis Percobaan	Jumlah *Token*	Topik			
		2	3	4	5
Maximum Entropy	1000	84.73	80.11	80.29	76.34
	2000	78.49	75.54	73.48	84.95
	3000	77.42	74.46	73.12	85.48
	4000	75.91	76.34	81.72	90.86
	5000	76.99	79.03	85.30	92.47
	6000	80.00	82.80	86.38	91.40
	7000	81.51	84.95	87.10	91.40
	8000	84.73	87.37	88.17	91.40
	9000	86.24	88.98	87.81	91.40
	10000	86.45	89.78	88.17	91.40
	ALL	92.26	92.20	88.17	91.40
Naïve Bayes	1000	91.61	94.09	91.40	91.40
	2000	92.26	93.55	93.19	95.70

f-n	3000	93.76	93.82	93.55	94.09
	4000	93.12	93.82	92.83	94.62
	5000	93.12	93.55	92.47	94.09
	6000	92.69	93.55	92.11	94.09
	7000	92.47	93.82	92.83	94.09
	8000	93.76	93.55	93.55	94.09
	9000	93.12	94.09	92.47	94.09
	10000	93.12	94.35	91.76	94.09
	ALL	92.26	93.82	93.19	94.09

Pada tabel 5.7, diketahui bahwa terdapat beberapa variasi perubahan nilai akurasi dari 3 topik hingga 5 topik untuk artikel media massa ini. Salah satu perubahan nilai akurasi itu adalah nilai akurasi akan naik seiring dengan bertambahnya jumlah topik. Perubahan ini terjadi pada sebagian besar pada percobaan dengan menggunakan metode Maximum Entropy. Sementara, pada percobaan dengan menggunakan metode Naïve Bayes *frequency-normalized*, perubahan yang terjadi adalah nilai akurasi akan naik dari 2 topik hingga 3 topik, lalu akan turun pada penggunaan 4 topik, dan kemudian naik lagi pada 5 topik. Tidak ada kesimpulan yang dapat ditarik dari perubahan nilai akurasi yang tidak menentu ini. Pergerakan naik dan turunnya nilai akurasi dapat dilihat lebih jelas pada gambar 5.5.

Pada gambar 5.5 dapat dilihat beberapa laju pergerakan jumlah topik pada nilai akurasi. MESC pada gambar tersebut berarti menggunakan metode Maximum Entropy dan f-n merupakan informasi fitur *frequency-normalized*. Pada gambar 5.5 tidak menampilkan semua nilai akurasi yang ada pada tabel 5.7, hanya beberapa laju perubahan yang dirasa dapat mewakili laju perubahan yang ada pada tabel 5.7. Pada gambar 5.5, percobaan yang ditampilkan hanya percobaan dengan menggunakan metode Naïve Bayes Multinomial dan Maximum Entropy untuk jumlah *token* 1000, 3000, dan 8000. Perubahan yang telah disebutkan sebelumnya ditampilkan pada metode Naïve Bayes *frequency-normalized* dengan menggunakan 3000 *token* (atau dapat juga pada penggunaan 8000 *token*) dan pada metode Maximum Entropy dengan menggunakan 8000 *token*. Perubahan lain

yang ditampilkan selain dua perubahan yang telah disebutkan tersebut merupakan beberapa perubahan yang terjadi pada jenis percobaan yang dilakukan.



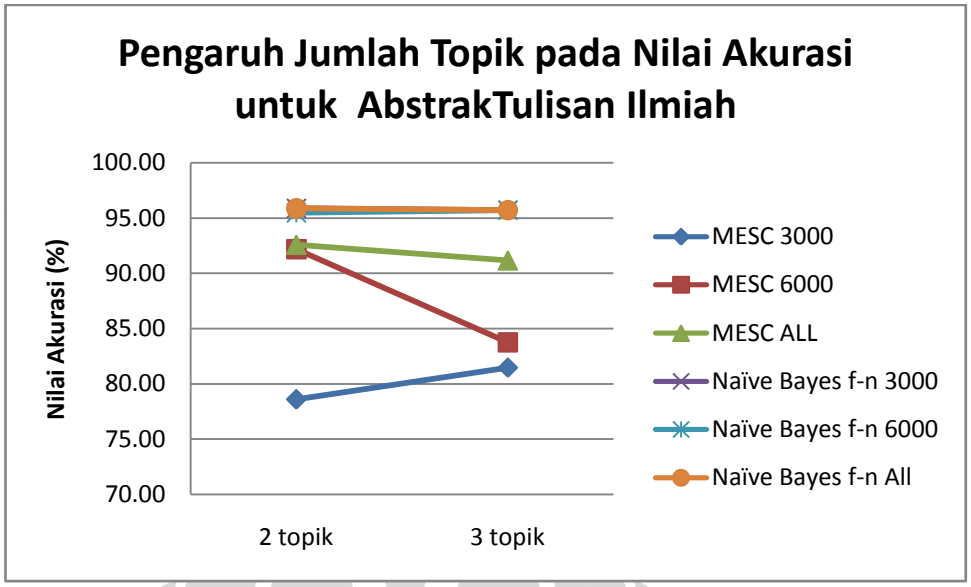
Gambar 5. 5 Grafik pengaruh jumlah topik pada nilai akurasi percobaan menggunakan Naïve Bayes *Frequency-normalized* (f-n) dan Maximum Entropy untuk artikel media massa.

Perubahan lain yang terjadi adalah nilai akurasi yang terus menurun hingga menggunakan 4 topik dan kemudian nilai akurasi akan naik pada penggunaan 5 topik, seperti pada Maximum Entropy dengan 3000 *token*. Selain itu, metode Naïve Bayes *frequency-normalized* pada penggunaan 1000 *token* memberikan nilai akurasi yang naik dari 2 topik ke 3 topik, tetapi setelah itu akan terus menurun hingga 5 topik. Variasi yang terjadi pada perubahan laju pergerakan nilai akurasi terhadap jumlah topik dilihat lebih lanjut pada data abstrak tulisan ilmiah. Penggantian data ini berguna untuk melihat laju pergerakan jumlah topik terhadap data yang berbeda. Akan tetapi, pada abstrak tulisan ilmiah hanya akan dilihat untuk penggunaan 2 topik dan 3 topik.

Tabel 5. 8 Tabel nilai akurasi yang dihasilkan pada percobaan yang menggunakan metode Naïve Bayes dengan informasi *frequency* normalized untuk perubahan jumlah topik 2 dan 3 pada abstrak tulisan ilmiah.

Jenis Metode	Jumlah *token*	2 topik	3 topik
Maximum Entropy *presence*	1000	84.36	84.05
	2000	79.84	82.34
	3000	78.60	81.48
	4000	79.84	82.91
	5000	86.01	82.91
	6000	92.18	83.76
	7000		90.88
	All	92.59	91.17
Naïve Bayes f-n	1000	95.47	94.59
	2000	95.47	94.87
	3000	95.88	95.73
	4000	95.88	95.44
	5000	95.88	95.73
	6000	95.47	95.73
	7000		96.01
	All	95.88	95.73

Tabel 5.8 menunjukkan perubahan nilai akurasi dari abstrak tulisan ilmiah pada penggunaan 2 topik dan 3 topik. Apabila menggunakan metode Naïve Bayes dengan informasi fitur *frequency-normalized*, nilai akurasi yang dihasilkan sebagian besar akan menurun dari 2 topik ke 3 topik. Akan tetapi, pada penggunaan metode Maximum Entropy, nilai akurasi yang dihasilkan terbagi merata antara naik dan turun dari 2 topik ke 3 topik. Naik turunnya nilai akurasi ini akan dapat dilihat lebih jelas pada gambar 5.6.



Gambar 5. 6 Grafik pengaruh jumlah topik pada nilai akurasi percobaan menggunakan Naïve Bayes *Frequency-normalized* (f-n) dan Maximum Entropy untuk abstrak tulisan ilmiah.

Pada gambar 5.6 menampilkan grafik pengaruh jumlah topik pada nilai akurasi yang dihasilkan untuk abstrak tulisan ilmiah. Gambar tersebut tidak menampilkan semua nilai yang ada pada tabel 5.8, hanya penggunaan metode Naïve Bayes dengan informasi *frequency-normalized* dan Maximum Entropy dengan 3000 *token*, 6000 *token*, dan semua *token* (*All tokens*). Pemilihan beberapa jumlah *token* ini dilakukan untuk memperjelas melihat perubahan nilai akurasi dari 2 topik ke 3 topik, sehingga dipilih beberapa variasi perubahan nilai akurasi yang dapat mewakilkan semua perubahan yang terjadi.

Sebagian besar nilai akurasi akan menurun dari 2 topik ke 3 topik pada penggunaan metode Naïve Bayes *frequency-normalized*. Hal ini dapat dilihat pada gambar 5.6 dimana untuk jumlah *token* 3000, 6000, dan all *token*, garis yang dihasilkan agak menumpuk dan menurun. Selain itu, pada penggunaan metode Maximum Entropy, sebagian nilai akurasi yang dihasilkan akan menurun dan juga akan naik dari 2 topik hingga 3 topik. Pada percobaan dengan menggunakan abstrak tulisan ilmiah ini, hanya dilihat pada 2 topik dan 3 topik, sehingga tidak dapat dilihat lebih dalam laju perubahan nilai akurasi seperti pada artikel media massa. Pada metode Maximum Entropy, apabila menggunakan tulisan ilmiah, nilai akurasi akan naik atau turun dari 2 topik dan 3 topik. Apabila menggunakan

artikel media massa data, nilai akurasi yang dihasilkan akan menurun dari penggunaan 2 topik hingga 4 topik dan naik kembali pada penggunaan 5 topik. Lain halnya pada Naïve Bayes dengan informasi *frequency-normalized*, sebagian besar nilai akurasi yang dihasilkan akan menurun baik pada penggunaan artikel media massa maupun pada penggunaan abstrak tulisan ilmiah. Oleh karena itu, baik pada abstrak tulisan ilmiah maupun artikel media massa, seiring dengan bertambahnya topik nilai akurasi yang dihasilkan cenderung akan menurun.

5.4 Hasil Klasifikasi Topik dari Aspek Data yang Digunakan

Pada subbab ini akan dibahas pengaruh data yang digunakan dengan nilai akurasi yang dihasilkan. Percobaan yang dilakukan menggunakan metode Naïve Bayes *frequency-normalized* dan Maximum Entropy, dengan jumlah *token* hanya 2000 *tokens* dan *All tokens*. Selain itu, pada pembahasannya hanya dilakukan dengan menggunakan 3 topik hal ini berkaitan dengan jenis data yang akan dibahas pada subbab 5.4.3 yaitu abstrak tulisan ilmiah. Dikarenakan keterbatasan data yang digunakan dan untuk menyelaraskan topik yang akan dibahas, maka nilai akurasi hanya dilihat pada penggunaan 3 topik. Pengaruh data yang digunakan akan dilihat dari tiga aspek, yaitu keseragaman data untuk tiap topik, jumlah data yang digunakan, dan kemiripan data. Penjelasan lebih lanjut mengenai ketiga aspek ini dibahas untuk tiga subbab mendatang.

5.4.1 Hasil Klasifikasi dari Aspek Keseragaman Jumlah Data untuk Tiap Topik

Percobaan yang dilakukan pada subbab ini menggunakan jumlah data kecil yang tidak seragam untuk tiap topiknya. Pemakaian data tidak seragam ini berlandaskan pemikiran bahwa dengan bertambahnya pengetahuan seseorang mengenai topik yang ada, orang itu akan lebih mudah untuk mengenali suatu dokumen baru. Jumlah data yang digunakan berkisar 93-186 artikel. Percobaan ini melibatkan 3 topik yaitu ekonomi, kesehatan, dan olahraga. Artikel media massa yang digunakan untuk ketiga topik ini tidaklah sama. Artikel media massa yang digunakan berjumlah 186 untuk data ekonomi, 162 untuk data kesehatan, dan 147 untuk data olahraga. Apabila pada percobaan sebelumnya data yang digunakan

memiliki jumlah yang sama untuk tiap topiknya, maka pada subbab ini ingin dilihat apakah penggunaan jumlah data akan memberikan pengaruh pada nilai akurasi yang dihasilkan. Tabel 5.9 merupakan hasil nilai akurasi dengan menggunakan metode Naïve Bayes dengan informasi *frequency-normalized* dan Maximum Entropy pada jumlah data 3 topik yang tidak seragam.

Tabel 5. 9 Hasil nilai akurasi dengan menggunakan metode Naïve Bayes dengan informasi *frequency-normalized* dan Mazimum Entropy pada jumlah data 3 topik yang tidak seragam jumlahnya.

Metode	Jumlah *token*	jumlah data tiap topik	
		seragam	tidak seragam
Naïve Bayes f-n	*All*	93.19	95.15
	2000	93.19	95.35
Maximum Entropy	*All*	88.17	93.13
	2000	73.48	87.47

Untuk jumlah data yang seragam, jumlah seluruh *token* adalah 11280 dan untuk jumlah data yang tidak seragam, percobaannya menggunakan 14369 *token*. Dapat dilihat pada tabel 5.5, dengan memakai data yang tidak seragam, nilai akurasi akan semakin tinggi. Kenaikan akurasi terjadi untuk semua jenis percobaan yang dilakukan. Perbedaan akurasi yang dihasilkan dimulai dari 1,9% hingga 14%. Kenaikan yang paling signifikan terjadi pada percobaan yang menggunakan metode Maximum Entropy dengan jumlah *token* 2000. Dengan melihat nilai akurasi yang dihasilkan pada tabel 5.9, dapat disimpulkan bahwa jumlah data yang dipakai tidaklah harus seragam. Apabila jumlah data yang dimiliki untuk tiap topik berbeda, percobaan masih dapat dilakukan dengan menggunakan data tersebut.

Data percobaan yang dipakai pada data tidak seragam merupakan data penambahan dari data seragam sehingga membuat jumlah data untuk tiap topik

berbeda. Melihat kenaikan nilai akurasi ini mengakibatkan muncul suatu hipotesis bahwa dengan memakai data yang lebih banyak akan membuat nilai akurasi semakin tinggi. Hal inilah yang akan dilihat pada subbab selanjutnya.

5.4.2 Hasil Klasifikasi dari Jumlah Data yang Digunakan

Pada subbab 5.4.1, nilai akurasi yang dihasilkan semakin meningkat pada percobaan yang memakai data tidak seragam. Hasil ini memperkuat pemikiran bahwa dengan bertambahnya pengetahuan, proses pengenalan atau klasifikasi suatu dokumen akan lebih mudah. Untuk lebih mempertegas pemikiran ini, pada subbab ini dilakukan percobaan dengan menggunakan data besar yangmana jumlah datanya lebih besar dibandingkan dengan data kecil. Untuk data besar, jumlah data yang dipakai berkisar 132-364. Percobaan ini akan menggunakan metode Naïve Bayes *frequency-normalized* dan Maximum Entropy dengan jumlah *token* 2000 dan *All tokens*.

Tabel 5. 10 Tabel pengaruh keseragaman data untuk tiap topik pada nilai akurasi yang dihasilkan dengan menggunakan metode Naïve Bayes *frequency-normalized* dan Maximum Entropy.

Metode	Jumlah *token*	Jumlah Data	
		Sedikit	Banyak
Naïve Bayes f-n	*All*	93.19	91,41
	2000	93.19	90.91
Maximum Entropy	*All*	88.17	89.39
	2000	73.48	76.77

Untuk data seragam dengan jumlah sedikit, jumlah *token*nya adalah 11280 dan jumlah *token* adalah 29594 untuk data seragam dengan jumlah banyak. Hasil ini agak bertolak belakang dengan nilai akurasi yang dihasilkan pada subbab 5.4.1. Nilai akurasi yang dihasilkan dengan menggunakan metode Maximum Entropy mengalami kenaikan, tetapi nilai akurasi yang dihasilkan dengan menggunakan metode Naïve Bayes *frequency-normalized* mengalami penurunan. Nilai akurasi yang dihasilkan pada tabel 5.10 memperlemah ataupun membantah pemikiran bahwa dengan bertambahnya pengetahuan, proses klasifikasi topik dapat

dilakukan dengan lebih mudah. Pemikiran ini akan benar apabila digunakan metode Maximum Entropy, sehingga tampaknya pemikiran ini akan kembali pada jenis metode yang digunakan.

5.4.3 Hasil Klasifikasi dari Aspek Kemiripan Data

Pada percobaan yang telah dibahas sebelumnya, data yang dipakai merupakan data dengan tingkat kemiripan yang rendah dimana orang awam mungkin dengan mudah dapat mendeteksi klasifikasi topik dari suatu dokumen. Dengan adanya pemikiran seperti ini, timbul suatu hipotesis bahwa apabila tingkat kemiripan itu tinggi, maka proses klasifikasi akan lebih susah dilakukan. Oleh karena itu, pada subbab ini akan dilihat pengaruh kemiripan data dengan nilai akurasi yang dihasilkan. Percobaan ini memakai dua jenis data, yaitu data dengan tingkat kemiripan kecil (artikel media massa) dan data dengan tingkat kemiripan tinggi (abstrak tulisan ilmiah). Untuk tiap data yang digunakan, jumlah topik yang digunakan hanya 3 dan jumlah data yang dipakai untuk tiap topik adalah 93. Jumlah semua *token* pada artikel media massa adalah 16037 dan pada abstrak tulisan ilmiah adalah 7968.

Tabel 5. 11 Tabel pengaruh tingkat kemiripan data untuk tiap topik pada nilai akurasi yang dihasilkan dengan menggunakan metode Naïve Bayes *frequency-normalized* dan Maximum Entropy.

Metode	Jumlah *token*	Jenis Data	
		Artikel	Abstrak
Naïve Bayes f-n	*All*	93.16	95.73
	2000	94.87	94.87
Maximum Entropy	*All*	93.16	91.17
	2000	80.34	82.34

Nilai akurasi yang dihasilkan pada tabel 5.11 membantahkan hipotesis bahwa semakin tinggi kemiripan data maka akan semakin rendah nilai akurasi yang dihasilkan. Dapat dilihat pada tabel 5.11, sebagian besar nilai akurasi yang dihasilkan dengan menggunakan abstrak tulisan ilmiah lebih tinggi dibandingkan

dengan nilai akurasi yang dihasilkan dengan menggunakan artikel media massa. Perbedaan nilai akurasi ini tidak hanya terjadi pada metode Naïve Bayes *Frequency-normalized* tetapi juga pada metode Maximum Entropy. Nilai akurasi dengan abstrak tulisan ilmiah pun dapat lebih kecil dibanding saat menggunakan artikel media massa seperti yang terjadi pada metode Maximum Entropy dengan memakai semua jumlah *token*. Oleh karena itu, tingkat kemiripan data tidak memiliki pengaruh pada nilai akurasi yang dihasilkan.

5.5 Kesalahan Klasifikasi Pada Beberapa Percobaan

Subbab ini akan melihat kesalahan klasifikasi yang telah dilakukan pada beberapa percobaan baik dengan menggunakan abstrak tulisan ilmiah maupun artikel media massa. Pada artikel media massa, percobaan yang digunakan adalah percobaan dengan menggunakan metode Maximum Entropy dan Naïve Bayes *frequency-normalized* dengan menggunakan semua *token* yang ada dan pada penggunaan 5 topik. Sementara, pada abstrak tulisan ilmiah, dilihat dari percobaan yang menggunakan metode Maximum Entropy dan Naïve Bayes *frequency-normalized* dengan menggunakan semua *token* yang ada pada penggunakan 3 topik.

Pada tabel 5.12 diberikan hasil klasifikasi yang dilakukan pada data artikel dengan menggunakan Maximum Entropy untuk tiap *fold*-nya. Pada baris pertama tersebut untuk fold 1 terdapat nilai 28 untuk [ekonomi,ekonomi] dan nilai 3 untuk [ekonomi, kesehatan], hal ini memberikan arti bahwa artikel ekonomi diklasifikasikan sebagai artikel ekonomi sebanyak 28 artikel dan artikel ekonomi dianggap sebagai artikel kesehatan sebanyak 3 artikel. Dapat dilihat pada tabel tersebut bahwa artikel ekonomi biasanya salah diklasifikasikan sebagai artikel kesehatan, sedangkan artikel kesehatan biasanya disalah klasifikasikan sebagai artikel ekonomi. Hal ini memberikan arti bahwa artikel ekonomi dan kesehatan memiliki tingkat similaritas yang tinggi sehingga dapat terjadi slaah klasifikasi. Selain itu, artikel olahraga dapat diklasifikasikan sebagai ekonomi dan kesehatan. Artikel properti dapat salah diklasifikasikan sebagai artikel kesehatan, dan artikel travel dapat salah diklasifikasikan sebagai artikel ekonomi, kesehatan, olahraga, dan properti. Akan tetapi, dapat dilihat bahwa artikel travel lebih sering salah diklasifikasikan sebagai artikel kesehatan dibandingkan dengan topik yang lain.

Tabel 5. 12 Tabel hasil klasifikasi pada percobaan yang menggunakan metode Maximum Entropy pada penggunaan semua *token untuk artikel media massa*.

Jenis Topik	ekonomi	kesehatan	olahraga	properti	travel
	fold 1				
ekonomi	28	3	0	0	0
kesehatan	1	28	0	0	2
olahraga	2	0	29	0	0
properti	0	1	0	30	0
travel	0	5	1	0	25
	fold 2				
ekonomi	26	4	0	1	0
kesehatan	0	31	0	0	0
olahraga	0	2	28	1	0
properti	0	0	0	31	0
travel	0	0	1	0	30
	fold 3				
ekonomi	30	1	0	0	0
kesehatan	1	30	0	0	0
olahraga	0	4	27	0	0
properti	0	2	0	29	0
travel	1	2	0	1	27

Apabila pada tabel 5.12 hasil klasifikasi dilihat dari penggunaan metode Maximum Entropy pada artikel media massa, maka pada tabel 5.13 hasil klasifikasi dilihat dari penggunaan metode Naïve Bayes *frequency-normalized*. Artikel ekonomi dapat salah diklasifikasikan sebagai artikel kesehatan, olahraga, dan travel. Sementara, artikel kesehatan dapat salah diklasifikasikan menjadi artikel ekonomi, olahraga, properti, dan travel. Artikel olahraga salah diklasifikan sebagai artikel ekonomi dan kesehatan. Selain itu, artikel properti lebih sering salah diklasifikasikan sebagai artikel kesehatan, dan artikel travel dapat salah diklasifikasikan sebagai artikel ekonomi, kesehatan, olahraga dan properti.

Tabel 5. 13 Tabel hasil klasifikasi pada percobaan yang menggunakan metode Naïve Bayes dengan informasi fitur *frequency-normalized* pada penggunaan semua *token untuk artikel media massa*.

Jenis Topik	ekonomi	kesehatan	olahraga	properti	travel
	fold 1				
ekonomi	27	1	2	0	1
kesehatan	0	24	0	2	5
olahraga	0	2	29	0	0
properti	0	1	0	30	0
travel	2	1	2	0	26
	fold 2				
ekonomi	29	0	0	0	2
kesehatan	1	28	1	1	0
olahraga	0	0	31	0	0
properti	0	0	0	30	1
travel	0	0	0	0	31
	fold 3				
ekonomi	29	2	0	0	0
kesehatan	0	28	0	2	1
olahraga	1	1	29	0	0
properti	0	2	0	29	0
travel	0	1	0	1	29

Pada artikel media massa ini, tingkat kemiripan antara ekonomi dan kesehatan cukup tinggi, hal ini dapat dilihat bahwa artikel ekonomi dapat salah diklasifikasikan sebagai artikel kesehatan dan artikel kesehatan dapat salah diklasifikasikan sebagai artikel ekonomi. Selain itu, ketiga topik yang lain juga dapat salah diklasifikasikan sebagai artikel kesehatan tetapi tidak berlaku timbal balik. Hal ini dapat dikatakan bahwa artikel kesehatan memberikan pengaruh yang cukup tinggi dalam melakukan proses klasifikasi.

Apabila pada artikel media massa digunakan 5 topik, maka pada abstrak tulisan ilmiah hanya digunakan 3 topik yaitu Rekayasa Perangkat Lunak (RPL), *Information Retrieval* (IR), dan citra. Jumlah data yang digunakan untuk tiap topik tersebut berbeda-beda, hal ini dilakukan karena telah mengetahui sebelumnya bahwa perbedaan jumlah data untuk tiap topik tidak membuat proses klasifikasi tidak dapat dilakukan. Pada tabel 5.14 ditunjukkan hasil klasifikasi dengan menggunakan metode Maximum Entropy untuk abstrak tulisan ilmiah.

Dapat dilihat pada tabel tersebut bahwa data RPL dapat salah diklasifikasikan sebagai data IR maupun citra, sedangkan data IR cukup sering salah diklasifikasikan sebagai RPL. Selain itu, data citra dapat salah diklasifikasikan sebagai RPL dan IR.

Tabel 5. 14 Tabel hasil klasifikasi pada percobaan yang menggunakan metode Maximum Entropy pada penggunaan semua *token* untuk abstrak tulisan ilmiah.

Jenis Topik	RPL	IR	Citra
	fold 1		
RPL	48	1	1
IR	2	33	1
Citra	3	2	26
	fold 2		
RPL	44	5	1
IR	1	35	0
Citra	6	23	2
	fold 3		
RPL	47	0	3
IR	1	35	0
Citra	0	2	29

Apabila pada data tabel 5.14 ditunjukkan hasil klasifikasi dengan menggunakan metode Maximum Entropy, maka pada tabel 5.15 ditunjukkan hasil klasifikasi dengan menggunakan metode Naïve Bayes dengan informasi fitur *frequency-normalized*. Pada tabel ini juga dapat dilihat bahwa data RPL cukup sering salah diklasifikasikan sebagai IR dan data IR salah diklasifikasikan sebagai data citra. Selain itu, data citra dapat salah diklasifikasikan sebagai data RPL dan data IR.

Tabel 5. 15 Tabel hasil klasifikasi pada percobaan yang menggunakan metode Naïve Bayes dengan informasi fitur *frequency-normalized* pada penggunaan semua *token* untuk abstrak tulisan ilmiah.

Jenis Topik	RPL	IR	Citra
	fold 1		
RPL	48	1	1
IR	0	35	1
Citra	1	2	28

	fold 2		
RPL	48	2	0
IR	0	34	2
Citra	4	0	27
	fold 3		
RPL	50	0	0
IR	0	35	1
Citra	0	0	31

Pada abstrak tulisan ilmiah, data RPL cukup sering salah diklasifikasikan sebagai data IR dan data IR cukup sering salah diklasifikasikan sebagai data citra. Selain itu, data citra dapat salah diklasifikasikan baik sebagai data RPL maupun data IR. Hal ini dapat terjadi mungkin karena perbedaan jumlah data untuk tiap topik. Apabila jumlah data disamakan untuk tiap topik, akan ada kemungkinan bahwa kecenderungan kesalahan dalam klasifikasi ini juga akan berubah.

5.6 Rangkuman Hasil

Pembahasan hasil untuk klasifikasi topik menggunakan *machine learning* telah dilakukan pada subbab 5.1 hingga 5.4. Subbab ini akan memberikan rangkuman singkat dari hasil-hasil yang telah didapatkan pada keempat subbab tersebut. Beberapa hal yang dapat dirangkum, antara lain:

- Melihat dari aspek metode yang digunakan, metode Naïve Bayes memberikan nilai akurasi yang lebih tinggi daripada metode Naïve Bayes Multinomial, sedangkan metode Maximum Entropy cenderung memberikan nilai akurasi yang lebih tinggi daripada metode Naïve Bayes Multinomial. Akan tetapi, metode Naïve Bayes dengan Maximum Entropy terkadang memberikan nilai akurasi yang hamper mirip.
- Melihat dari aspek jumlah *token* yang digunakan, timbul pemikiran bahwa semakin banyak jumlah *token* yang digunakan, semakan tinggi nilai akurasi yang dihasilkan. Berdasarkan hasil percobaan, pemikiran ini cenderung benar.

- Melihat dari aspek informasi fitur yang digunakan, hanya dilakukan percobaan dengan Naïve Bayes dan Naïve Bayes Multinomial. Pada Naïve Bayes, nilai akurasi akan mencapai nilai tertinggi saat menggunakan *frequency-normalized*. Akan tetapi, fitur *frequency-normalized* pada Naïve Bayes Multinomial membuat nilai akurasi mencapai nilai terendah dan tidak dapat disimpulkan fitur mana yang memberikan nilai maksimum pada Naïve Bayes Multinomial
- Banyak topik yang digunakan akan cenderung menurunkan nilai akurasi. Hal ini terjadi baik dengan menggunakan metode Naïve Bayes *frequency-normalized* maupun dengan Maximum Entropy pada abstrak tulisan ilmiah dan artikel media massa. Sama halnya dengan penelitian pada (Sebastiani,2002), penelitian ini menunjukkan semakin banyak jumlah topik yang digunakan semakin rendah pula nilai akurasi yang diberikan.
- Penggunaan banyaknya data yang tidak sama untuk tiap topik memberikan nilai yang lebih tinggi dibandingkan saat menggunakan jumlah data yang seragam untuk tiap topik. Akan tetapi, hal ini tidak membenarkan pemikiran semakin banyak data akan memberikan nilai akurasi yang semakin tinggi. Hal ini terbukti nilai akurasi yang dihasilkan dari percobaan yang telah dilakukan dengan menambah jumlah data untuk tiap topik. Nilai akurasi tersebut tidak bertambah tinggi seiring dengan bertambahnya data.
- Pemikiran bahwa semakin tinggi tingkat kemiripan suatu data, nilai akurasi yang dihasilkan semakin rendah, telah dibuktikan salah. Hal ini dikarenakan nilai akurasi yang dihasilkan dengan menggunakan abstrak tulisan ilmiah cenderung memberikan nilai yang lebih tinggi dibandingkan dengan menggunakan artikel media massa.
- Hasil yang dilakukan untuk melakukan klasifikasi dengan menggunakan *machine learning* menunjukkan bahwa *machine learning* dapat dilakukan pada bahasa Indonesia dan memberikan nilai akurasi yang cukup bagus. Selain itu, kata-kata tunggal dalam suatu dokumen dapat digunakan untuk mengenali topik yang terkandung dalam artikel tersebut.