Online News Classification Using Multinomial Naive Bayes

Amelia Rahman
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
amelia.rahman@student.uns.ac.id

Wiranto
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
wiranto@staff.uns.ac.id

Afrizal Doewes
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
afrizal.doewes@staff.uns.ac.id

ABSTRACT

The huge availability of text in numerous forms is the valuable information resource that can be used for various purposes. One of the text mining methods to analyze text document is classification. Text classification is a process of grouping and categorizing a document based on the training models. This study aimed to categorize Indonesian news automatically using Multinomial Naive Bayes. To get more optimal result, feature selection process using Document Frequency Thresholding method and term weighting using Term Frequency-Inverse Document Frequency (TF-IDF) were applied. The experiment showed that Multinomial Naive Bayes with TF-IDF produced the highest average accuracy to 86,62 % *while Multinomial Naive Bayes reached* 86,28%, *Multinomial Naive Bayes with DF-Thresholding-TFIDF to* 86,15% *and Multinomial Naive Bayes with DF-Thresholding to* 85,98%. *Feature selection with Document Frequency Thresholding is quite efficient to reduce the number of data dimension shown with the result of insignificant final accuracy from Multinomial Naive Bayes method.*

Keywords: text categorization, classification, multinomial, df-thresholding, tf-idf

1 PENDAHULUAN

Seiring dengan semakin berkembangnya teknologi, penyebaran informasi di internet berkembang dengan semakin pesat dan terus mengalami peningkatan. Salah satu bentuk informasi yang jumlahnya terus bertambah adalah berita [1].

Berita merupakan laporan tercepat mengenai fakta atau opini yang penting atau menarik minat bagi sejumlah besar orang [2]. Aliran berita dalam jumlah besar tentu menjadi sumber informasi yang sangat berharga dan dapat dimanfaatkan untuk berbagai kepentingan. Salah satu pemanfaatannya adalah untuk melakukan monitoring isu. Monitoring isu merupakan kegiatan untuk mengamati isu-isu terhangat yang sedang terjadi di masyarakat. Dengan dilakukannya monitoring isu yang diambil dari berita di berbagai media, maka dapat dipetakan kebijakan apa yang akan diambil dalam tatanan pimpinan. Isu-isu yang terjadi di masyarakat sangat beragam, sehingga sangat tepat apabila isu-isu tersebut dikelola dan dianalisis untuk kemudian menjadi pijakan dalam membuat keputusan [3].

Analisis berita yang umumnya berbentuk teks dapat dilakukan dengan *text mining*. *Text mining* dapat didefinisikan sebagai suatu proses untuk mengekstrak informasi yang berguna dari suatu sumber data melalui identifikasi dan eksplorasi pola tertentu [4]. Terdapat beberapa metode dalam *text mining*, salah satunya adalah klasifikasi. Klasifikasi bermanfaat untuk mengelompokan data yang jumlahnya sangat banyak dan sulit dilakukan apabila diproses secara manual.

Penelitian mengenai pengelompokan teks pernah dilakukan oleh Amir Hamzah [5] yang mengelompokan dokumen berita dan abstrak akademis dengan menggunakan *Naive Bayes Classifier*. Akurasi untuk dokumen berita mencapai 91% sedangkan pada dokumen akademik mencapai 82%. Penelitian lain dilakukan oleh Yudi Wibisono [6] yang mengelompokan berita dari www.kompas.com dengan meggunakan *Naive Bayes Classifier*. Hasil pengujian menunjukan tingkat akurasi yang tinggi yaitu sebesar 90,23%. Salah satu model *event Naive Bayes Classifier* adalah model multinomial. Dalam penelitian [7] yang mengelompokan teks Bahasa Indonesia ke dalam beberapa kelas emosi, dapat disimpulkan bahwa metode multinomial lebih baik digunakan untuk klasifikasi teks Bahasa Indonesia ditunjukan dengan pengukuran F-Measure yang mencapai 62,15%.

Masalah yang umum ditemukan dalam proses klasifikasi maupun *clustering* dokumen adalah tingginya dimensi data, sehingga perlu dilakukan proses seleksi fitur untuk memilih beberapa fitur yang dapat digunakan untuk mewakili dokumen. [8]

Berdasarkan hal tersebut, maka penelitian ini dilakukan untuk menganalisis berita dengan cara klasifikasi menggunakan *Naive Bayes* model *Multinomial*. Selain itu, dilakukan juga proses seleksi fitur menggunakan metode *Document Frequency Thresholding* dan pembobotan dengan menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)* untuk mendapatkan perbandingan hasil yang terbaik.

2 DASAR TEORI

2.2 TEXT MINING

Text mining dapat didefinisikan secara luas sebagai suatu proses dimana pengguna berinteraksi dengan koleksi dokumen menggunakan suatu *tool* analisis. Dalam kaitannya dengan *data mining*, *text mining* merupakan suatu proses untuk mengekstrak

informasi yang berguna dari suatu sumber data melalui identifikasi dan eksplorasi pola tertentu*Text mining* merupakan suatu proses untuk mengekstrak informasi yang berguna dari suatu sumber data melalui identifikasi dan eksplorasi pola tertentu [4].

Tujuan dari *text mining* yaitu untuk memproses informasi tekstual yang tidak terstruktur, mengekstrak indeks numerik yang bermakna dari teks, dan kemudian membuat informasi yang terkandung di dalam teks dapat diakses menggunakan berbagai algoritma *data mining* [9].

2.3 TEXT PREPROCESSING

Text preprocessing merupakan suatu proses pengubahan bentuk data tekstual yang belum terstruktur menjadi data yang terstruktur [10]. Tahapan-tahapan *text preprocessing* pada penelitian ini yaitu :

- *Case Folding*
  Proses *case folding* merupakan proses untuk menghilangkan semua karakter selain huruf (seperti angka dan tanda baca) dan mengubah semua huruf menjadi huruf kecil.
- *Tokenization*
  Tokenization merupakan proses pemotongan kalimat berdasarkan tiap kata yang menyusunnya.
- *Stemming*
  Stemming merupakan proses pemotongan imbuhan atau pengembalian kata berimbuhan menjadi kata dasar. Algoritma *stemming* yang digunakan dalam penelitian ini adalah Algoritma Nazief & Adriani.
- *Filtering*
  Filtering atau disebut juga *stopword removal* merupakan proses untuk menghilangkan *stopwords* (kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words). Proses ini bertujuan untuk mengurangi jumlah kata.

2.4 DOCUMENT FREQUENCY THRESHOLDING

Salah satu teknik seleksi fitur yang paling sederhana namun memiliki kinerja yang cukup baik adalah *Document Frequency Thresholding* yang bersifat *class independent* [9].

Document Frequency merupakan banyaknya jumlah dokumen yang mengandung *term* tertentu. *Term* yang jarang muncul memiliki kemungkinan besar tidak memberikan informasi spesifik. Begitupun jika *term* tersebut terlalu sering muncul pada banyak dokumen, maka dianggap bahwa *term* tersebut merupakan *term* yang umum dan tidak akan mempengaruhi kinerja pediksi secara keseluruhan [7].

2.5 TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY (TF-IDF)

Term weighting atau pembobotan kata bertujuan untuk memberikan bobot nilai pada setiap kata. Perhitungan bobot ini memerlukan dua hal, yaitu *Term Frequency (TF)* dan *Inverse Document Frequency (IDF). Term Frequency* merupakan banyaknya jumlah kata atau *term* tertentu yang ada dalam suatu dokumen. Sementara *Inverse Document Frequency* adalah frekuensi kemunculan kata atau *term* pada seluruh dokumen. Nilai IDF berbanding terbalik dengan jumlah dokumen yang mengandung *term* tertentu. *Term* yang jarang muncul pada seluruh dokumen memiliki nilai IDF yang lebih besar dari nilai IDF *term* yang sering muncul. Jika pada setiap dokumen mengandung *term* tertentu, maka nilai IDF *term* tersebut bernilai 0. Hal ini menunjukan bahwa *term* yang muncul pada seluruh dokumen merupakan *term* yang tidak berguna untuk membedakan dokumen berdasarkan topik tertentu [11].

Rumus TF-IDF adalah sebagai berikut :

$$W_{dt} = tf_{dt} \; x \; idf_t = tf_{dt} \; x \; log \; (\frac{N}{df_t}) \qquad (2.1)$$

Dimana :

$W_{d,j}$ = bobot *term* ke-t terhadap dokumen d
tf_d = jumlah kemunculan *term* t dalam dokumen - d
N = jumlah dokumen secara keseluruhan
df_t = jumlah dokumen yang mengandung *term* t

2.6 MULTINOMIAL NAIVE BAYES

Model multinomial memperhitungkan frekuensi setiap kata yang muncul pada dokumen. Misal terdapat dokumen d dan himpunan kelas c. Untuk memperhitungkan kelas dari dokumen d, maka dapat dihitung dengan rumus :

$$P(c|term\ dokumen\ d) = P(c) \; x \; P(t_1|c) \; x \; P(t_2|c) \; x \; P(t_3|c) \; x \; ... \; x \; P(t_n|c) \qquad (2.2)$$

Keterangan :

P(c) = Probabilitas *prior* dari kelas c
t_n = Kata dokumen d ke-n
$P(c|term\ dokumen\ d)$ = Probabilitas suatu dokumen termasuk kelas *c*
$P(t_n|c)$ = Probabilitas kata ke-n dengan diketahui kelas *c*

- Probabilitas *prior* kelas c ditentukan dengan rumus:

  $$P(c) = \frac{N_c}{N} \qquad (2.3)$$

  Keterangan :
  Nc = Jumlah kelas c pada seluruh dokumen
  N = Jumlah seluruh dokumen

- Probabilitas kata ke-n ditentukan dengan menggunakan teknik *laplacian smoothing* :

  $$P(t_n \mid c) = \frac{count(tn,c)+1}{count(c)+|V|} \qquad (2.4)$$

  Keterangan :
  $count(t_n,c)$ = Jumlah term t_n yang ditemukan di seluruh data pelatihan dengan kategori c

count(c) = Jumlah term di seluruh data pelatihan dengan kategori c

V = Jumlah seluruh tem pada data pelatihan

Sementara rumus Multinomial yang digunakan dengan pembobotan kata TF-IDF adalah sebagai berikut :

$$P(t_n \mid c) = \frac{Wct+1}{(\sum W' \in V W' ct)+B'} \quad (2.5)$$

Keterangan :

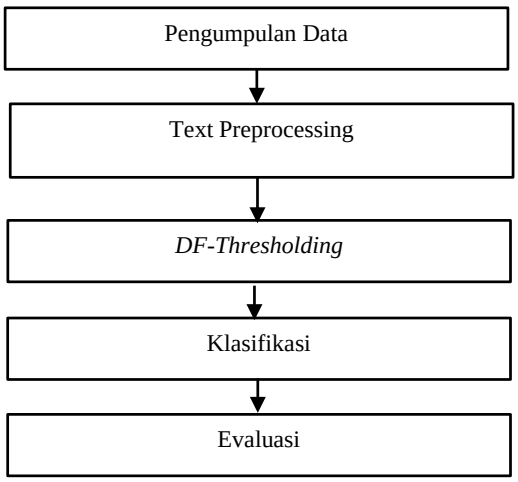
Wct = Nilai pembobotan tfidf atau W dari *term* t di kategori c

$\sum W' \in V W' ct$ = Jumlah total W dari keseluruhan *term* yang berada di kategori c.

B' = Jumlah W kata unik (nilai idf tidak dikali dengan tf) pada seluruh dokumen.

3 METODOLOGI

Tahap-tahap yang dilakukan dalam menyelesaikan penelitian ini dijelaskan pada Gambar 3.1 berikut ini :



Gambar 3.1 Metodologi Penelitian

3.1 Pengumpulan Data

Data yang dibutuhkan dalam penelitian ini merupakan data berita yang dimonitoring oleh Lembaga Pengolahan dan Penyedia Informasi (PPI), Dirjen Informasi dan Komunikasi Publik (IKP), Kementerian Komunikasi dan Informatika.

3.2 *Text Preprocessing*

Seluruh data akan melalui tahapan *text preprocessing* yaitu *case folding*, *tokenization*, *stemming* dan *filtering.* Hasil dari *text preprocessing* ini berupa database kata-kata yang akan digunakan untuk proses klasifikasi.

3.3 *Document Frequency Thresholding*

Sebelum masuk proses klasifikasi, dilakukan terlebih dahulu proses seleksi fitur dengan menggunakan *DF-Thresholding*. Tujuan dari proses ini adalah untuk mengurangi dimensi data. Pada proses ini, dilakukan perhitungan *document frequency* atau jumlah dokumen yang mengandung kata tertentu. Selanjutnya menentukan *threshold*, apabila jumlah data kurang dan lebih dari threshold, maka kata tersebut tidak digunakan pada proses klasifikasi.

3.4 Klasifikasi

Proses klasifikasi dibagi menjadi dua tahap, yaitu *training* dan *testing*. Pada proses *training*, masing-masing data berita diproses dan setiap kata dihitung jumlah kemunculannya. Data-data ini yang kemudian akan digunakan sebagai bahan pembelajaran pada proses *testing* untuk menentukan suatu data berita masuk pada kelas isu tertentu. Proses ini dilakukan dengan menggunakan metode *Multinomial Naive Bayes* dan metode *Multinomial Naive Bayes* dengan *Term Frequency-Inverse Document Frequency (TFIDF).*

3.5 Evaluasi

Proses evaluasi pada penelitian ini menggunakan perhitungan akurasi, *precision* dan *recall* dari hasil klasifikasi yang disajikan dengan tabel *confusion matrix* pada Tabel 3.1 berikut ini.

Tabel 3.1 *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	

Dengan rumus perhitungan sebagai berikut :

Accuracy :

$$\frac{\text{TP(Kelas}-1\text{)} + \text{TP(Kelas}-2\text{)} + .. + \text{TP(Kelas}-n\text{)}}{Total\ (Kelas-1)+Total(Kelas-2)+\cdots+Total(Kelas-n)} \quad (3.1)$$

$$\textit{Precision} : \frac{\text{TP(Kelas}-i\text{)}}{Prediksi\ (Kelas-i)} \quad (3.2)$$

$$\textit{Recall} : \frac{\text{TP(Kelas}-1\text{)}}{Total\ (Kelas-i)} \quad (3.3)$$

4 HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Data dalam penelitian ini diperoleh dari Lembaga Pengolahan dan Penyedia Informasi (PPI), Dirjen Informasi dan Komunikasi Publik (IKP) yang berasal dari 5 media *online* yaitu detik.com, viva.co.id, inilah.com, antaranews.com, dan okezone.com yang diambil dari bulan Februari 2016 sampai dengan Mei 2016 untuk bidang Polhukam, Ekonomi dan Kesra.

Jumlah data berita yang digunakan sebanyak 1011 data dengan total isu yang dikelompokan sebanyak 15 kelas. Pembagian data berdasarkan jumlah kelas isu ditujukkan pada Tabel 4.1 berikut ini.

Tabel 4.1 Data Penelitian

Kategori	Isu	Jumlah
Polhukam	Politik Dalam Negeri	67
	Politik Luar Negeri	87
	Kasus Korupsi	77
	Hukum	42
	Keamanan	42
	Kasus Narkoba	77
	Kekerasan Seksual	68
Perekonomian	Pajak	84
	Perbankan	78
Kesra	Infrastruktur	26
	Kecelakaan	78
	Bencana Alam	83
	Energi	67
	Lingkungan Hidup	79
	Transportasi Publik	52

Data-data ini kemudian dibagi untuk proses *training* dan proses *testing.* Karena data berita dipengaruhi oleh waktu, maka data-data berita yang terbit lebih awal dijadikan sebagai data *training* pada awal proses.

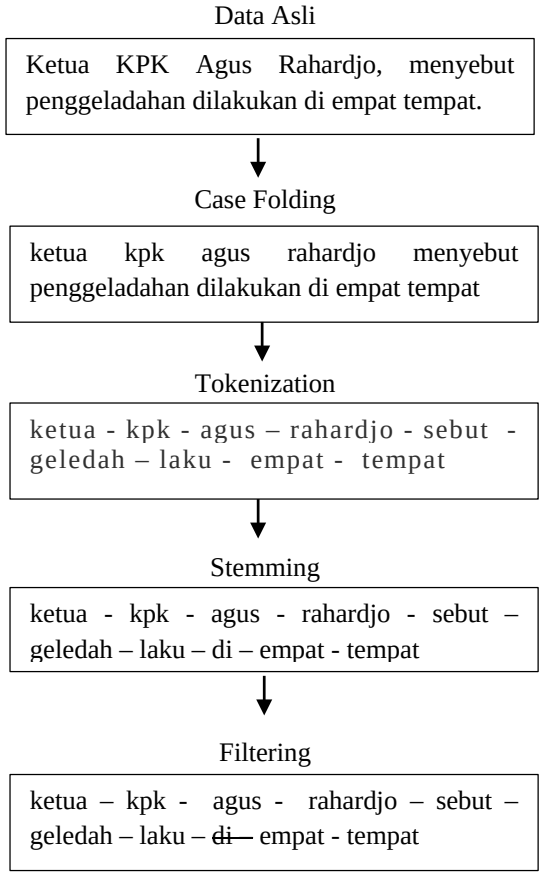
Data yang digunakan pada awal *training* sebanyak 395 data diperoleh dari masing-masing kategori berdasarkan waktu terbit yang lebih awal. Selanjutnya, proses *testing* menggunakan data yang terbit setelah data-data *training* sebelumnya yang dibagi menjadi 6 kali proses *testing.* Setiap data *testing* yang telah dilakukan, akan digunakan sebagai data *training* pada proses selanjutnya.

4.2 *Text Preprocessing*

Proses pertama yang dilakukan adalah *text preprocessing.* Dalam *text preprocessing* terdapat beberapa tahapan. Tahapan pertama adalah *case folding* yang bertujuan untuk menghilangkan semua karakter selain huruf di dalam data dan mengubah semua huruf menjadi huruf kecil (*lowercase).* Tahap kedua yaitu *tokenization* yang bertujuan untuk mengubah bentuk *string* menjadi *token-token.*

Selanjutnya adalah proses f*iltering* yang bertujuan untuk menghilangkan *stopwords.* Dan proses terakhir adalah s*temming*, dimana data berita yang telah melalui proses *case folding* akan diproses untuk menghilangkan imbuhan pada kata sehingga semua kata menjadi kata dasar atau *root word*. Proses *stemming* ini menggunakan Algoritma Nazief & Adriani [11] yang diambil dari *library* Sastrawi. Semua hasil dari *text preprocessing* ini kemudian disimpan di dalam database kata.

Contoh tahapan *text preprocessing* ditunjukkan pada gambar 4.1 berikut ini :



Gambar 4.1 Contoh *Text Preprocessing*

4.3 *Document Frequency Thresholding*

Penentuan *threshold* dilakukan dengan percobaan menentukan batas bawah dan batas atas pada data *training* awal.

Berdasarkan pengamatan pada hasil perhitungan *document frequency*, maka dilakukan pengujian pada data *training* dengan beberapa *threshold* yang ditunjukan pada Tabel 4.2.

Tabel 4.2 Pengujian *DF-Threshold*

Threshold	Jumlah Kata	Akurasi
1 - 55	8.737	99,74
1 - 60	8.865	99,74
2 - 50	7.769	98,97
2 - 55	7.940	99,22
2 - 60	8.068	99,22

Threshold yang dipilih untuk digunakan pada proses selanjutnya adalah 1 – 55.

4.4 TF-IDF

Proses pembobotan TF-IDF ini dimulai dengan menghitung tiap *term* yang ada pada setiap dokumen

(TF). Kemudian proses dilanjutkan dengan menghitung jumlah dokumen yang memiliki *term* tertentu (DF). Setelah itu proses menghitung *Inverse Document Freqeuncy* (IDF) dan yang terakhir nilai TF dikalikan dengan nilai IDF.

Contoh perhitungan TF-IDF ditampilkan pada tabel-tabel berikut ini.

Tabel 4.3 Contoh Berita

No	Berita
1	gubernur bank indonesia bi agus martowardojo kata atur ampun pajak tax amnesty sangat jalan upaya perintah bi sektor uang
2	bahas rancang undang undang ampun pajak ruu tax amnesty tengah gulir dpr kali parlemen minta otoritas hukum menindaklanjuti jelas ruu sebut
3	bahas rancang undang undang ampun pajak ruu tax amnesty laku komisi xi dewan wakil rakyat sore selasa april

Tabel 4.4 Contoh Perhitungan TF-IDF

Kata	TF			DF	IDF	TF-IDF		
	D1	D2	D3			D1	D2	D3
bi	2	0	0	1	0,4771	0,954	0	0

4.5 Klasifikasi

Data *training* yang telah melewati seleksi fitur kemudian digunakan sebagai bahan pembelajaran pada proses *testing* untuk menentukan suatu data berita masuk pada kelas isu tertentu.

Proses *testing* diuji dengan beberapa percobaan data. Sama halnya seperti data pada proses *training*, data *testing* melewati tahap *text preprocessing* dan kemudian dihitung nilai tiap fitur kata yang akan digunakan untuk proses klasifikasi menggunakan metode *Multinomial Naive Bayes.*

Tahapan perhitungan pada proses klasifikasi :

- Menghitung data *prior* masing-masing kelas dengan menggunakan rumus 2.3.
- Menghitung probabilitas kata ke-n data berita dengan menggunakan rumus 2.4 atau 2.5.
- Menghitung probabilitas suatu dokumen masuk ke dalam suatu kelas dengan rumus 2.2.
- Menentukan kelas dokumen dengan memilih nilai probabilitas tertinggi.

4.6 Evaluasi

Hasil evalusi seluruh data *testing* ditunjukan pada Tabel 4.3 berikut ini.

- Multinomial Naive Bayes

Tabel 4.5 Tabel Evaluasi Multinomial Naive Bayes

Data *Training*	Data *Testing*	Akurasi	Precision	Recall
395	98	70,41	69,33	72,06
493	101	86,14	82,41	85,76
594	103	89,32	92,12	91,72
697	101	88,12	91,37	88,87
798	104	89,42	89,17	87,56
902	105	94,29	95	93,05

- Multinomial Naive Bayes dengan DF-Thresholding

Tabel 4.6 Tabel Evaluasi Multinomial Naive Bayes dengan DF-Thresholding

Data *Training*	Data *Testing*	Akurasi	Precision	Recall
395	98	72,45	71,21	74,5
493	101	85,15	79,96	85,24
594	103	89,32	91,57	91,72
697	101	90,09	91,67	90,58
798	104	85,58	84,82	83,27
902	105	93,33	91,55	92,09

- Multinomial Naive Bayes dengan TF-IDF

Tabel 4.7 Tabel Evaluasi Multinomial Naive Bayes dengan TF-IDF

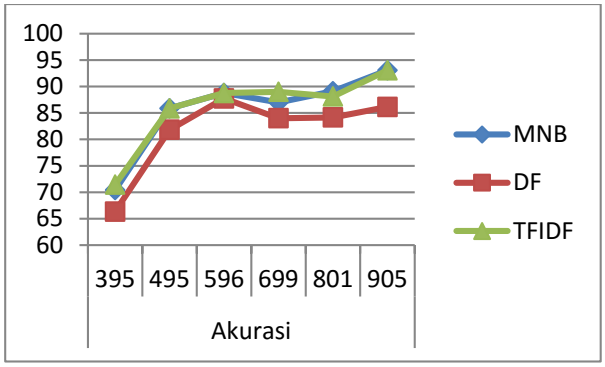
Data *Training*	Data *Testing*	Akurasi	Precision	Recall
395	98	71,43	68,04	72,88
493	101	86,14	83,77	85,76
594	103	89,32	92,12	91,72
697	101	90,09	92,88	90,65
798	104	88,46	88,06	86,44
902	105	94,29	95	93,05

- Multinomial Naive Bayes dengan DF-TFIDF

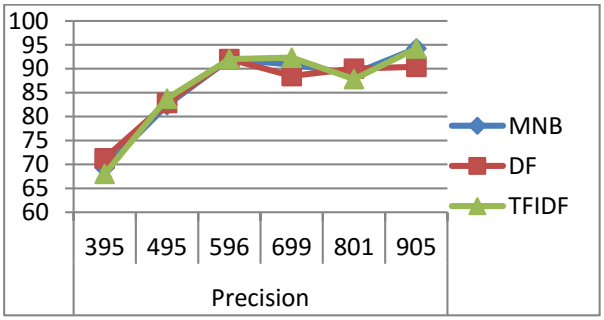
Tabel 4.8 Tabel Evaluasi Multinomial Naive Bayes dengan DF-TFIDF

Data *Training*	Data *Testing*	Akurasi	Precision	Recall
395	98	72,45	69,9	74,5
493	101	86,14	82,18	85,24
594	103	89,32	91,57	91,65
697	101	90,09	92,39	90,58
798	104	86,54	87,64	84,38
902	105	92,38	90,28	91,14

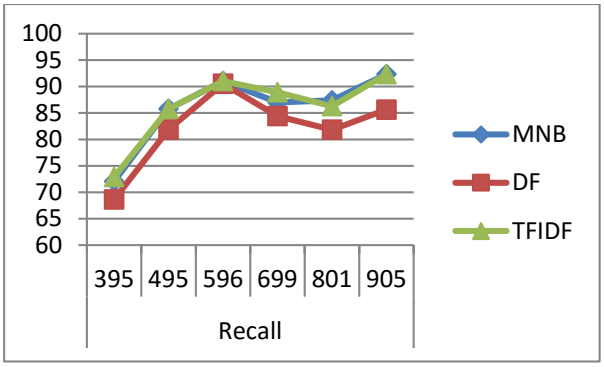
Untuk melihat perbandingan hasil antara ketiga metode tersebut, dapat dilihat dari gambar grafik berikut ini.



Gambar 4.2 Grafik Akurasi



Gambar 4.3 Grafik Precision

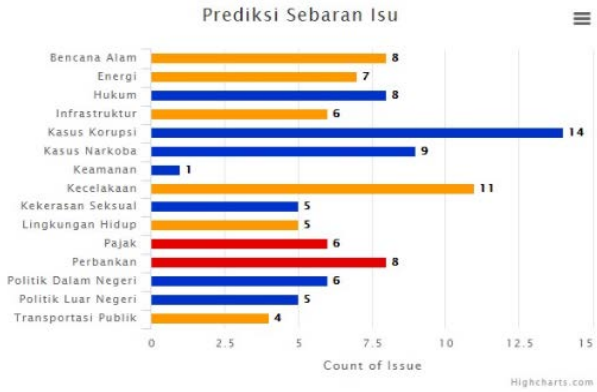


Gambar 4.4 Grafik Recall

Dari hasil proses pengujian, tingkat akurasi di awal proses cenderung kecil. Hal ini disebabkan karena data berita yang diinputkan sebagai data *testing* memiliki fitur-fitur kata yang belum muncul pada data berita sebelumnya yang dijadikan sebagai proses pembelajaran. Namun dari hasil tersebut, dapat dianalisis bahwa dengan semakin bertambahnya data *training*, akurasi cenderung naik dan stabil. Hal ini dikarenakan fitur-fitur kata yang dijadikan sebagai bahan pembelajaran lebih banyak dan beragam.

Metode *Multinomial Naive Bayes* dengan menggunakan pembobotan TF-IDF memiliki hasil yang lebih baik dari metode *Multinomial* maupun *Multinomial* dengan menggunakan fitur seleksi *DF-Thresholding*. Penggunaan *DF-Thresholding* justru mengurangi nilai akurasi disebabkan karena adanya penghilangan beberapa *term* pada pemotongan *threshold* yang mewakili suatu dokumen tertentu dan tidak digunakan dalam proses klasifikasi.

Contoh hasil dari klasifikasi ditampilkan dengan diagram batang yang ditunjukkan pada Gambar 4.1 berikut ini.



Gambar 4.5 Contoh Hasil Klasifikasi

Keterangan :

- Warna biru merupakan diagram untuk persebaran isu bidang Polhukam.
- Warna merah merupakan diagram untuk persebaran isu bidang Perekonomian.
- Warna kuning merupakan diagram untuk persebaran isu bidang Kesra.

5 KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil penelitian yang telah dilakukan dapat ditarik kesimpulan bahwa metode *Multinomial Naive Bayes* dapat digunakan untuk menganalisis berita dalam teks Bahasa Indonesia ditunjukan dengan hasil akurasi akhir sebesar 94,29%. Penggunaan fitur seleksi dengan metode *DF-Thresholding* dapat mengurangi tingginya dimensi data, ditunjukan dengan pengurangan fitur sebanyak 22.215 menjadi 16.766 dengan nilai akurasi akhir yang tidak jauh dari metode *Multinomial Naive Bayes* yaitu sebesar 93,33%. Sementara penggunaan *TFIDF* pada metode *Multinomial Naive Bayes* menunjukan hasil akurasi akhir sebesar 94,29% dan dapat meningkatkan nilai rata-rata akurasi yang lebih besar dari metode *Multinomial Naive Bayes* dari 86,28% menjadi 86,62%. Penggunaan *DF-Threshold* pada *Multinomial Naive Bayes* dengan *TFIDF* pun menunjukan nilai akurasi akhir yang lebih rendah dari *Multinomial Naive Bayes* dengan *TFIDF*, yaitu sebesar 92,38% namun dengan penggunaan jumlah fitur yang lebih sedikit dalam proses klasifikasi.

5.2 Saran

Beberapa saran yang dapat dilakukan untuk pengembangan pada penelitian ini :

1. Menambah fungsi untuk mengatur besar *thresholding* pada seleksi fitur secara otomatis sehingga tahap pengujian dapat dilakukan pada beberapa *threshold.*
2. Uji coba dengan beberapa metode seleksi fitur selain *DF-Thresholding* dan pembobotan TF-IDF, seperti *Mutual Information, Information Gain* dan *Chi Square.*

DAFTAR PUSTAKA

[1] Y. Wibisono dan M. L. Khodra, "Clustering Berita Berbahasa Indonesia," *FMIPA UPI & STEI ITB*, 2005.

[2] S. Nasution. [Online]. Available: http://sumut.kemenag.go.id/file/file/Teknismembuatberitadanpressrelease/kurg1330936704.pdf. [Diakses 15 Mei 2016].

[3] Kominfo, Oktober 2014. [Online]. Available: https://kominfo.go.id/index.php/content/detail/4214/Kemkominfo+Selenggarakan+FGD+Monitoring+Informasi+Publik/0/berita_satker. [Diakses 16 Mei 2016].

[4] R. Feldman dan J. Sanger, The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2006.

[5] A. Hamzah, "Klasifikasi Teks dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis," dalam *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*, Yogyakarta, 2012.

[6] S. Sumpeno dan I. Destuardi, "Klasifikasi Emosi untuk Teks Bahasa Indonesia menggunakan Metode Naive Bayes," *Seminar Nasional Pascasarjana*, 2009.

[7] R. Nallaswamy, "A Study on Analysis of SMS Classification Using Document Frequency Threshold," dalam *I.J. Information Engineering and Electronic Business*, 2012.

[8] StatSoft, "Text Mining Introductory Overview," 2016. [Online]. Available: http://www.statsoft.com/Textbook/Text-Mining. [Diakses 8 Mei 2016].

[9] D. P. Langgeni, Z. A. Baizal dan Y. F. A.W., "Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection," dalam *Seminar Nasional Informatika*, Yogyakarta, 2010.

[10] G. Miner, A. Fast, D. Delen, T. Hill, J. Elder dan B. Nisbet, Practical Text Mining and Statistical Analysis for Non-Structured Text Data Application, Oxford: Elsevier, 2012.