

News Opinion Mining around Universitas Sebelas Maret Using Naive Bayes Algorithm

Muhammad Mukhlis Khoirudin
Informatika, Fakultas MIPA
Universitas Sebelas Maret
Jalan Ir. Sutami 36A Surakarta
muhammad.mk@student.uns.ac.id

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

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

Abstract

Opinion mining or sentiment analysis is a part of text mining and widespread topic nowadays. Opinion mining is the process of understanding, extracting, and processing textual data automatically to get sentiment information contained in a sentence. One of the opinion mining method that can be used to analyzed text documents is classification. This research aims to classify Indonesian news into three classes of positive, negative, and neutral using Multinomial Naïve Bayes.

To get optimal result, the author tries to add some feature selections using Document Frequency Thresholding (DF-Thresholding) and Term Weighting using Term Frequency-Inverse Document Frequency (TF-IDF).

The result showed that the classification using Multinomial Naïve Bayes obtained the highest accuracy with an average 92.44%, Multinomial Naïve Bayes with DF-Thresholding had an accuracy of 83.44%, and using Multinomial Naïve Bayes with Term Frequency-Inverse Document Frequency (TF-IDF) get an accuracy 78,33%. The actual purpose of using the feature selection in this research to add accuracy value, but the result show less influence in terms of accuracy. Using the selection feature can reduce the use of term dimension.

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

1. Pendahuluan

Universitas Sebelas Maret (UNS) merupakan salah satu universitas yang menempati 10 besar di Indonesia. Dalam penilaian yang dilakukan oleh 4 International Collages & Universities UNS menduduki peringkat ke-5, dan salah satu aspek yang dinilai dari website [1].

Banyaknya website yang dikelola oleh kampus ini dapat memperkaya konten dan aset kampus, salah satunya yaitu informasi berita khususnya koran. Berita yaitu laporan mengenai suatu kejadian atau peristiwa yang terbaru atau aktual, laporan mengenai fakta-fakta yang aktual, menarik perhatian, dinilai penting, atau bahkan luar biasa [2].

Berita koran yang dikelola oleh UNS berasal dari beberapa sumber, diantaranya yaitu: Jawa Pos, Suara Merdeka, Solopos, Joglo Semar, Kompas, Kedaulatan Rakyat, Koran Tempo, Media Indonesia, dan Seputar Indonesia. Redaksi yang berjumlah banyak dengan berbagai macam opini yang disampaikan membuat

penulis memproses berita UNS yang ada [3]. Tujuan klasifikasi tersebut yaitu untuk memudahkan admin media sosial dalam menyiarkan berita sentimen positif yang berasal dari website koran UNS.

Analisis berita secara umum berbentuk teks dapat dilakukan dengan *text mining*. *Text mining* dapat didefinisikan sebagai penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru dalam bentuk sebuah informasi, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda [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.

Ada berbagai macam *text mining* dan salah satunya yaitu *opinion mining*. *Opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek yang ditemukan oleh seseorang, apakah cenderung ke arah positif atau negatif [5].

Penelitian yang berkaitan tentang *opinion mining* ada beberapa, diantaranya Imam Fahrur Rozi, Sholeh Hadi Pramono dan Erfan Achmad Dahlan yang mengembangkan sistem *opinion mining* untuk menganalisis data opini publik pada perguruan tinggi. Dari hasil pengujian, didapatkan nilai uji dan data target dimana semua hasil pengujiannya menunjukkan nilai *precision* dan *recall* yang sama-sama tinggi. Untuk akurasi NBC dalam mengklasifikasikan opini tergolong baik, yaitu nilai *precision* yaitu 0.95 dan untuk nilai *recall* yaitu 0.94 [5]. Penelitian Daniel Pakpahan, Hilda Widyastuti mencoba melakukan penelitian terkait dengan *opinion mining* untuk menilai berita *online* yang berasal dari situs news.kompas.com dengan menggunakan algoritma Naive Bayes Classifier (NBC) [6]. Masalah yang umum ditemukan dalam proses klasifikasi dokumen adalah tingginya dimensi data, sehingga perlu dilakukan proses seleksi fitur untuk memilih beberapa fitur dan metode yang dapat digunakan untuk mewakili dokumen [7].

Berdasarkan hal tersebut, maka penelitian ini dilakukan untuk menganalisis berita dengan cara klasifikasi menggunakan metode *Naive Bayes* model *Multinomial Naive Bayes* merupakan salah satu metode di dalam *data mining* untuk mengklasifikasi data, cara

kerja NBC menggunakan perhitungan probabilitas [8]. Selain itu, dalam penelitian ini dilakukan proses seleksi fitur menggunakan metode *Document Frequency Thresholding* dan pembobotan dengan menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)* yang bertujuan untuk mendapatkan perbandingan hasil yang terbaik.

2. Dasar Teori

2.1. Text Mining

Text mining yaitu penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru dalam bentuk sebuah informasi, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit, yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda [4]. 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.

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* [7].

2.2. Text Preprocessing

Text preprocessing berfungsi sebagai proses pengubahan bentuk data teks yang belum terstruktur menjadi data yang terstruktur. Tahapan *text preprocessing* [6] 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 (*lowercase*).
- *Tokenization*
Tokenization merupakan proses pemotongan dokumen/ kalimat berdasarkan tiap kata yang menyusunnya.
- *Stemming*
Stemming merupakan proses pemotongan imbuhan atau pengembalian kata berimbuhan menjadi kata dasar. Dalam penelitian ini, penulis menggunakan *library* sastrawi, yaitu *library* berbasis PHP yang digunakan untuk *stemming* kata Bahasa Indonesia.
- *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.

2.3. Opinion Mining

Opinion mining atau sentimen analisis merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau

objek yang ditemukan oleh seseorang, apakah cenderung ke arah positif atau negatif [5].

Opinion mining adalah bidang studi yang menganalisis pendapat, sentiment, evaluasi, penilaian, sikap, pendapat, dan emosi terhadap entitas seperti produk, layanan, organisasi, individu, isu, peristiwa, dan topik yang berfokus pada opini yang mengarah pada sentiment positif atau negatif. [9]. Tujuan dari *opinion mining* untuk mengekstrak atribut dan komponen dari objek yang telah dikomentari pada setiap dokumen dan untuk menentukan apakah komentar tersebut positif atau negatif [9].

2.4. Document Frequency Thresholding

Document Frequency yaitu adalah jumlah dokumen dimana mengandung suatu *term* tertentu dan tiap term akan dihitung nilai *Document Frequency* (DF). Salah satu teknik seleksi fitur yang paling sederhana namun memiliki kinerja yang cukup baik adalah *Document Frequency Thresholding* yang bersifat *class independent* [10].

Term akan diseleksi berdasarkan nilai DF, apabila nilai DF dibawah *threshold* yang ditentukan, maka term tersebut akan dibuang. *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 prediksi secara keseluruhan [10].

2.5. Term Frequency - Inverse Document Frequency (TF-IDF)

Term weighting atau pembobotan kata yaitu bertujuan untuk memberikan bobot nilai pada setiap kata atau *term*. 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. Hal ini menunjukkan bahwa *term* yang muncul pada seluruh dokumen merupakan *term* yang tidak berguna untuk membedakan dokumen berdasarkan topik tertentu [10].

Rumus TF-IDF adalah sebagai berikut :

$$W_{dt} = tf_{dt} \times idf_t = tf_{dt} \times \log \left(\frac{N}{df_t} \right) \quad (2.1)$$

Dimana :

- W_{dt} = 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, dalam model ini mengambil jumlah kata yang muncul dalam sebuah dokumen. Misal terdapat dokumen d digambarkan sebagai distribusi multinomial kata dalam himpunan kelas c . Untuk memperhitungkan kelas dari dokumen d , maka dapat dihitung dengan persamaan :

$$P(c|\text{term dokumen } d) = P(c_i) \times P(t_1|c_i) \times P(t_2|c_i) \times P(t_3|c_i) \times \dots \times P(t_n|c_i) \quad (2.2)$$

Keterangan :

$P(c_i)$ = Probabilitas *prior* dari kelas c

t_n = Kata dokumen d ke- n

$P(c|\text{ term dokumen } d)$ = Probabilitas suatu dokumen termasuk kelas c

$P(t_n|c_i)$ = Probabilitas kata ke- n dengan diketahui kelas c

- Probabilitas *prior* kelas c ditentukan dengan persamaan:

$$P(c) = \frac{C_i}{N} \quad (2.3)$$

Keterangan :

C_i = Jumlah *term* kelas- i pada seluruh dokumen

N = Jumlah seluruh dokumen

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

$$PLaplace(i) = \frac{C_i + 1}{N + V} \quad (2.4)$$

Keterangan :

$PLaplace(w_i)$ = Probabilitas *Laplacian Smoothing*

C_i = Jumlah term ditemukan di seluruh data pelatihan dengan kelas c

N = Jumlah term di seluruh data pelatihan dengan kelas c

V = Jumlah seluruh term pada data pelatihan

Untuk persamaan Multinomial yang digunakan dengan pembobotan kata TF-IDF yaitu sebagai berikut :

$$P(t_n | c) = \frac{W_{ct} + idfn}{(\sum_{W' \in V} W'_{ct}) + \sum_i idfn} \quad (2.5)$$

Keterangan :

W_{ct} = Nilai pembobotan tfidf atau W dari *term* t di kategori c

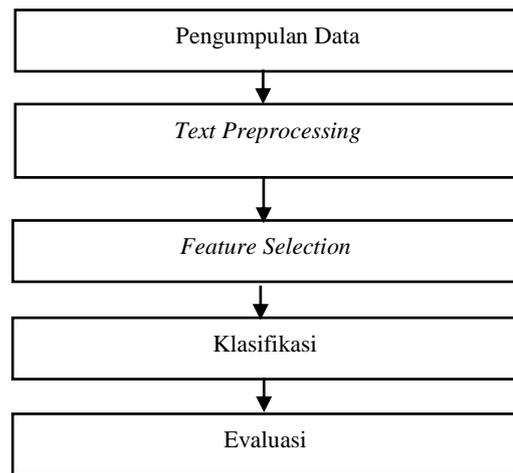
$idfn$ = Nilai invers *df* ke- n dari *term*

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

$\sum_i idfn$ = Jumlah total dari nilai *idf* ke- n dari suatu *term* terkait

3. Metodologi

Tahapan yang dilakukan oleh penulis dalam menyelesaikan penelitian ini dijelaskan pada Gambar 3.1 berikut ini :



Gambar 3.1 Metodologi Penelitian

3.1. Pengumpulan Data

Data yang digunakan oleh penulis dalam penelitian ini merupakan data berita yang diperoleh dari Badan Administrasi Perencanaan Sistem Informasi (BAPSI) Universitas Sebelas Maret (UNS). Data berbentuk teks menggunakan bahasa Indonesia dimana berisikan berita yang terkait seputar Universitas Sebelas Maret (UNS). Teks tersebut berasal dari website berita uns yaitu www.koran.uns.ac.id.

3.2. Text Preprocessing

Dalam tahap *text preprocessing* data akan diproses melalui beberapa tahapan, diantaranya yaitu *case folding*, *tokenization*, *stemming* dan *filtering*. *Case folding* mengubah semua huruf menjadi huruf kecil. *Tokenization* pemotongan tiap kata penyusun. *Stemming* mencari kata dasar dari kata. *Filtering* mengambil kata-kata penting hasil token, Hasil dari *text preprocessing* ini berupa database kata-kata yang akan digunakan untuk proses klasifikasi.

3.3. Feature Selection

Sebelum masuk proses klasifikasi, dilakukan terlebih dahulu proses seleksi fitur dengan menggunakan beberapa fitur, yaitu *Document Frequency Thresholding* (DF-Thresholding) dan *Term Frequency-Inverse Document Frequency* (TF-IDF). 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. Pada proses pembobotan, jumlah term akan dihitung berdasar kemunculannya (TF) dan seluruh dokumen (DF). Kemudian dihitung nilai inverse (IDF) dari nilai DF, kemudian dikalikan dengan nilai TF. Hasil pembobotan tersebut digunakan dalam proses klasifikasi

3.4. Klasifikasi

Klasifikasi teks adalah proses pengelompokan dokumen kedalam kelas berbeda. Dalam tahapnya, setiap dokumen d menunjuk pada satu kelas tertentu,

maka dibutuhkan proses untuk menggali informasi dari dokumen tersebut sehingga dapat mempresentasikan dari kelasnya dan mempunyai nilai. Proses klasifikasi dibagi menjadi dua proses, yaitu proses *training* dan proses *testing*.

Pada proses *training*, masing-masing data berita diproses dan setiap kata dihitung jumlah kemunculannya. Data-data ini yang kemudian akan digunakan sebagai acuan bahan pembelajaran pada proses selanjutnya yaitu 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 menggunakan *Term Frequency-Inverse Document Frequency (TF-IDF)*.

3.5. Evaluasi

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

Tabel 3.1 Confusion Matrix

Realita	Sistem			Total
	Kelas 1	Kelas 2	Kelas 3	
Kelas 1	True Positive	Error	Error	Total Kelas 1
Kelas 2	Error	True Positive	Error	Total Kelas 2
Kelas 3	Error	Error	True Positive	Total Kelas 3

Dengan persamaan perhitungan sebagai berikut :

$$Accuracy : \frac{TP(Kelas\ 1)+TP(Kelas\ 2)+TP(Kelas\ 3)}{Total\ (Kelas\ 1)+Total\ (Kelas\ 2)+Total\ (Kelas\ 3)} \quad (3.1)$$

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

$$Recall : \frac{TP(Kelas\ i)}{Total\ (Kelas\ i)} \quad (3.3)$$

4. Hasil Dan Pembahasan

4.1. Pengumpulan Data

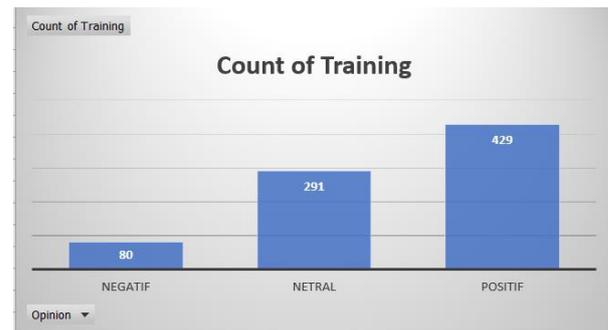
Data dalam penelitian ini berasal dari Biro Administrasi Perencanaan Sistem Informasi (BAPSI) Universitas Sebelas Maret (UNS) yang terdiri dari beberapa sumber, diantaranya yaitu: Jawa Pos, Suara Merdeka, Solopos, Joglo Semarang, Kompas, Kedaulatan Rakyat, Koran Tempo, Media Indonesia, dan Seputar Indonesia. Data berbentuk teks tersebut menggunakan bahasa Indonesia yang diambil dari bulan Januari 2016 sampai dengan bulan Oktober 2016.

Jumlah data berita yang digunakan yaitu sebanyak 1000 data teks, dengan rincian dari 1000 teks tersebut akan dibagi menjadi 800 data *training* dan 200 data *testing*. Dalam data *training* terdapat 3 kelas, yaitu positif, negatif dan netral, pembagiannya yaitu data dengan opini negatif berjumlah 80, opini netral 291, dan opini positif 429 berita. Penjelasan rinciannya terdapat pada Tabel 4.1 dan Gambar 4.1.

Tabel 4.1 Jumlah Data Training

Opinion class	Count of Training
Negatif	80
Netral	291
Positif	429

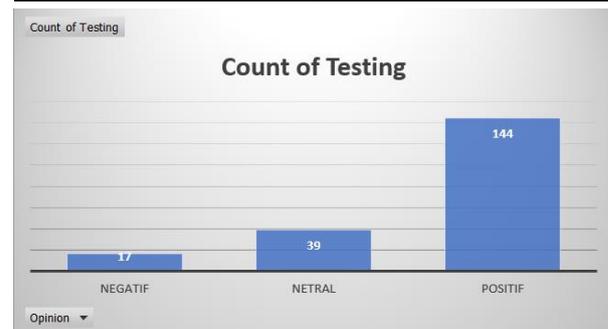
Gambar 4.1 Perbandingan Grafik Data Training



Jumlah alokasi data teks untuk data *testing* yaitu 200, dengan rincian sebagai berikut: 17 data opini negatif, 39 data opini netral, 144 data opini positif. Data tersebut dapat dilihat pada Tabel 4.2 dan Gambar 4.2.

Tabel 4.2 Jumlah Data Testing

Opinion class	Count of Testing
Negatif	17
Netral	39
Positif	144



Gambar 4.2 Jumlah Data Testing

Data dibagi menjadi dua bagian yaitu proses *training* dan proses *testing* seperti pada Gambar 4.2. Karena data berita dipengaruhi oleh waktu, maka data-data berita yang terbit lebih awal dijadikan sebagai data *training* diolah pada awal proses yaitu pada bulan Januari sampai Juli 2016. Data *testing* menggunakan data berita bulan Agustus sampai dengan Oktober 2016.

4.2. Text Preprocessing

Dalam *text preprocessing* ada beberapa tahap, diantaranya yaitu: *case folding*, *stemming*, *filtering*, dan *tokenizing*. Proses pertama yang dilakukan adalah *text*

preprocessing. 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 *stemming*, dalam tahapan ini merupakan proses untuk mencari kata dasar dari setiap kata. Proses ini menghilangkan imbuhan dan tambahan yang melekat pada kata. Proses *stemming* ini menggunakan *library* Sastrawi.

Tahap yang selanjutnya yaitu proses *filtering* atau biasa disebut *stopword removal*, merupakan tahapan untuk menghilangkan stopwords dan mengambil kata-kata yang penting. Tujuannya untuk mengurangi jumlah kata dan menghilangkan *stopword*.

Tahap yang terakhir yaitu *tokenization* yang bertujuan untuk mengubah bentuk *string* menjadi *token-token* atau mengubah kalimat dalam dokumen menjadi potongan kata. Semua hasil dari *text preprocessing* ini kemudian disimpan di dalam database kata.

Contoh tahapan *text preprocessing* ditunjukkan pada Gambar 4.3.



Gambar 4.3 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 ditunjukkan pada Tabel 4.1.

Tabel 4.1 Pengujian *DF-Threshold*

<i>Threshold</i>	Jumlah Kata	Akurasi
1 – 55	3.772	96.0 %
1 – 60	3.776	95.0 %
2 – 55	1.863	97.0 %
2 – 60	1.868	96.5 %

Threshold yang dipilih untuk digunakan pada proses selanjutnya adalah 2 – 55, jadi *term* atau kata yang jumlahnya kurang dari 2, dan lebih dari 55 akan dihilangkan. Penulis mengambil *threshold* 2-55 karena dengan dimensi jumlah kata yang minimal bisa memperoleh hasil akurasi yang maksimal.

4.4. Term Frequency-Invers Document Frequency (TF-IDF)

Pada proses pembobotan TF-IDF ini dimulai dengan menghitung tiap *term* (kata) yang ada pada setiap dokumen (TF). Selanjutnya yaitu menghitung jumlah dokumen yang memiliki *term* tertentu (DF). Kemudian dilanjutkan proses menghitung *Inverse Document Frequency* (IDF) dan yang terakhir nilai TF dikalikan dengan nilai IDF.

Contoh perhitungan TF-IDF ditampilkan pada Tabel 4.2 dan Tabel 4.3 berikut ini.

Tabel 4.2 Contoh Berita

No	Berita
1	universitas sebelas maret uns solo menargetkan akhir tahun telah berstatus perguruan negeri berbadan hukum ptnbh dan uns bisa meraup penghasilan minimal rp miliar
2	universitas sebelas maret solo siap laksana program computer based test cbt terima mahasiswa baru tahun akademik
3	mahasiswa universitas sebelas maret ikut program kuliah kerja nyata knn kelompok kota solo gelar giat sosialisasi

Tabel 4.3 Contoh Perhitungan TF-IDF

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

4.5. Klasifikasi

Dalam tahap klasifikasi, data *training* yang telah melewati seleksi fitur kemudian digunakan sebagai bahan pembelajaran pada proses *testing* untuk menentukan suatu data berita masuk pada kelas opini 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 diantaranya yaitu: Menghitung data *prior* masing-masing kelas dengan menggunakan Persamaan 2.3. Kemudian menghitung probabilitas kata ke-n data berita dengan menggunakan Persamaan 2.4 atau 2.5. Setelah itu menghitung probabilitas suatu dokumen masuk ke dalam suatu kelas dengan Persamaan 2.2. Yang terakhir menentukan kelas dokumen dengan memilih nilai probabilitas tertinggi.

4.6. Evaluasi

Hasil evaluasi seluruh data *testing* ditunjukkan pada Tabel 4.5, Tabel 4.6, dan Tabel 4.7 berikut ini.

Tabel 4.5 Tabel Evaluasi Multinomial Naive Bayes

Data Training	Data Testing	Akurasi	Precision	Recall
800	100	93,00	90,31	81,02
800	150	92,00	86,37	75,81
800	200	92,32	84,07	78,06

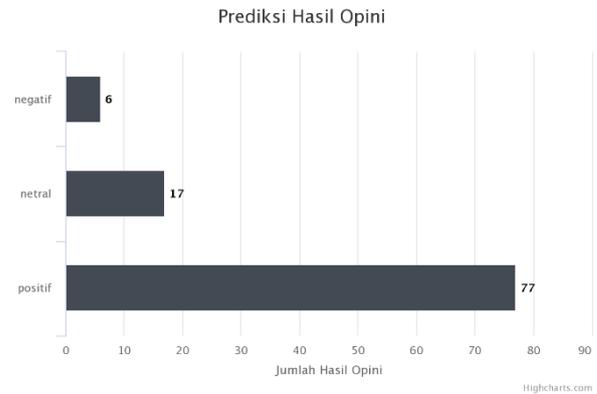
Tabel 4.6 Tabel Evaluasi Multinomial Naive Bayes dengan DF-Thresholding

Data Training	Data Testing	Akurasi	Precision	Recall
800	100	87,00	78,93	88,48
800	150	81,33	72,65	78,99
800	200	82,00	72,32	78,06

Tabel 4.7 Tabel Evaluasi Multinomial Naive Bayes dengan TF-IDF

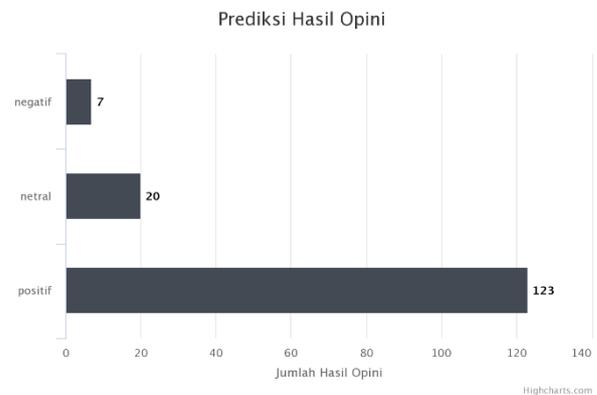
Data Training	Data Testing	Akurasi	Precision	Recall
800	100	80,00	69,11	78,79
800	150	78,00	61,61	71,47
800	200	77,00	62,71	71,20

Bagian prediksi hasil opini, terdapat 3 gambar diagram batang, yang membedakan yaitu adalah jumlah data *testing*. Berikut ini adalah hasil pengelompokan tiap kelas yaitu dibagi menjadi negatif, netral, dan positif. Grafik dengan 100 data *testing* bisa dilihat pada Gambar 4.4 berikut ini.



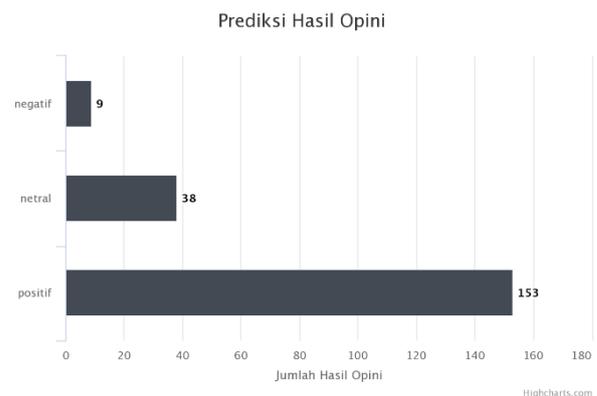
Gambar 4.4 Diagram batang 100 testing

Grafik dengan menggunakan 150 data *testing* dapat dilihat pada Gambar 4.5.



Gambar 4.5 Grafik 150 testing

Grafik dengan menggunakan 200 data *testing* dapat dilihat pada Gambar 4.6.



Gambar 4.6 Grafik 200 testing

Dari hasil proses pengujian, dapat dilihat bahwa untuk kelas yang paling tinggi dari ketiganya berasal dari kelas opini positif, kemudian kelas opini netral, dan yang terakhir kelas opini negatif.

Testing pertama dilakukan dengan 800 data *training* dan 100 data *testing*. Hasilnya yaitu 77 data

masuk kelas opini positif, 17 data masuk kelas opini netral, dan 6 data masuk kelas opini negatif.

Testing kedua dilakukan dengan 800 data *training* dan 150 data testing. Hasilnya yaitu 123 data masuk kelas opini positif, 20 data masuk kelas opini netral, dan 7 data masuk kelas opini negatif.

Testing ketiga dilakukan dengan 800 data *training* dan 200 data testing. Hasilnya menunjukkan 153 data masuk kelas opini positif, 38 data masuk kelas opini netral, dan 9 data masuk kelas opini negatif.

Dari perhitungan data Tabel 4.5, Tabel 4.6, Tabel 4.7, dapat disimpulkan bahwa semakin sedikit data *testing*, maka *accuracy*, *precision* dan *recall* dari ketiga metode cenderung semakin bagus, namun dengan catatan jumlah data *training* tetap. Hal ini disebabkan karena semakin banyak data *testing* yang diolah, maka akan semakin menambah jumlah kompleksitas *term*, sehingga variasi *term* dari *preprocessing text* data *testing* akan lebih banyak dan menimbulkan *term* baru. Penyebab yang mungkin yaitu *term* tersebut mungkin belum ada dalam data *training*, sehingga menyebabkan nilai *akurasi*, *presision*, dan *recall* cenderung menurun.

Untuk perbandingan metodenya, metode *Multinomial Naïve Bayes* mempunyai hasil yang lebih bagus dari segi *accuracy* dan *presision* dibanding dua metode yang lain. Sedangkan *DF-Thresholding* mempunyai hasil yang lebih baik dari segi *recall*. *Multinomial Naïve Bayes* dengan menggunakan TF-IDF paling rendah diantara ketiga metode yang dibandingkan. Hal ini mungkin disebabkan karena pembobotan *term* dalam penelitian tersebut antara data *training* dengan data *testing* kurang maksimal.

Penggunaan *DF-Thresholding* justru mengurangi nilai akurasi mungkin disebabkan karena adanya penghilangan beberapa *term* pada pemotongan *threshold* yang mewakili suatu dokumen tertentu dan tidak digunakan dalam proses klasifikasi.

5. Kesimpulan Dan Saran

5.1. Kesimpulan

Dari hasil penelitian yang telah dilakukan oleh penulis dapat disimpulkan bahwa penggunaan fitur seleksi dan penghapusan *redundancy* data dapat mengurangi banyaknya jumlah *term* dari 19.187 menjadi 3.809. Metode *Multinomial Naïve Bayes* dapat digunakan untuk menganalisis berita dalam teks Bahasa Indonesia ditunjukkan dengan hasil *accuracy* terbesar yaitu 93,00%, dengan rata-rata nilai 92,44 %.

Penggunaan fitur seleksi dengan metode *DF-Thresholding* dapat mengurangi kompleksitas jumlah *term*, ditunjukkan dengan pengurangan fitur dari yang semula 3.809 menjadi 1.863 dengan rasio pengurangan sebesar 48,9% dengan nilai akurasi akhir yang lebih kecil dari metode *Multinomial Naïve Bayes* yaitu paling besar 87,00 % dengan rata-rata nilai 83,44%.

Sementara penggunaan *TF-IDF* pada metode *Multinomial Naïve Bayes* menunjukkan hasil akurasi terbesar yaitu 80,00% dengan rata-rata nilai 78,33%. Penggunaan *DF-Threshold* pada *Multinomial Naïve Bayes* dan *TF-IDF* pun menunjukkan nilai akurasi akhir yang lebih rendah dari *Multinomial Naïve Bayes* normal, 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 sumber data berita dari berbagai sumber yang lain.
2. Mencari besar *thresholding* yang tepat untuk memperoleh hasil yang maksimal, sehingga bisa meningkatkan nilai *accuracy*, *precision* dan *recall*.
3. Uji coba dengan beberapa metode seleksi fitur selain *DF-Thresholding* dan pembobotan TF-IDF, seperti *Word Frequency (WF)*, *Mutual Information(MI)*, *Information Gain(IG)*, *Odds Ratio (OR)* dan *Chi Square*.

DAFTAR PUSTAKA

- [1] "Top Universities in Indonesia | 2017 Indonesian University Ranking." [Online]. Available: <http://www.4icu.org/id/>.
- [2] B. K, "Konsep Dasar Berita," *Scribd*. [Online]. Available: <https://www.scribd.com/doc/102009253/Konsep-Dasar-Berita>.
- [3] "Koran | Home." [Online]. Available: <http://koran.uns.ac.id/>. [Accessed: 30-Aug-2017].
- [4] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge; New York: Cambridge University Press, 2007.
- [5] I. F. Rozi, S. H. Pramono, and E. A. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi," *J. EECCIS*, vol. 6, no. 1, pp. 37–43, 2013.
- [6] D. Pakpahan and H. Widyastuti, "Aplikasi Opinion Mining dengan Algoritma Naïve Bayes untuk Menilai Berita Online," *J. Integrasi*, vol. 6, no. 1, pp. 1–10, 2014.
- [7] "Text Mining, Big Data, Unstructured Data." [Online]. Available: <http://www.statsoft.com/Textbook/Text-Mining>. [Accessed: 30-Aug-2017].
- [8] S. L. Ting, W. H. Ip, and A. H. Tsang, "Is Naive Bayes a good classifier for document classification," *Int. J. Softw. Eng. Its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.
- [9] K. M. Gandecha, V. S. Gondane, and V. R. Shelke, "A Survey on Opinion Mining."
- [10] D. P. Langgeni, Z. A. Baizal, and Y. F. AW, "Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection," in *Seminar Nasional Informatika (SEMNASIF)*, 2015, vol. 1.