E-Jurnal
MATEMATIKA

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Udayana

Vol 3, No 3 (2014)

Table of Contents

Articles

ISSN: 2303-1751

ANALISIS SENTIMEN MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER* DENGAN SELEKSI FITUR *CHI SQUARE*

Juen Ling[1], I Putu Eka N. Kencana[§2], Tjokorda Bagus Oka[3]

[1]Jurusan Matematika, Fakultas MIPA - Universitas Udayana [Email: juenling260292@gmail.com]
[2]Jurusan Matematika, Fakultas MIPA - Universitas Udayana [Email: i.putu.enk@gmail.com]
[3]Jurusan Matematika, Fakultas MIPA - Universitas Udayana [Email: tjokordabagusoka@gmail.com]
[§]*Corresponding Author*

ABSTRACT

Sentiment analysis is the computational study of opinions, sentiments, and emotions expressed in texts. The basic task of sentiment analysis is to classify the polarity of the existing texts in documents, sentences, or opinions. Polarity has meaning if there is text in the document, sentence, or the opinion has a positive or negative aspect. In this study, classification of the polarity in sentiment analysis using machine learning techniques, that is Naïve Bayes classifier. Criteria for text classification decisions, learned automatically from learning the data. The need for manual classification is still required because training the data derived from manually labeling, the label (feature) refers to the process of adding a description of each data according to its category. In the process of labeling, feature selection is used and performed by chi-square feature selection, to reduce the disturbance (noise) in the classification. The results showed that the frequency of occurrences of the expected features in the true category and in the false category have an important role in the chi-square feature selection. Then classification by Naïve Bayes classifier obtained an accuracy of 83% and a harmonic average of 90.713%.

Keywords: *chi square, classification, feature selection, machine learning technique, Naïve Bayes, sentiment analysis.*

1. PENDAHULUAN

Informasi dalam bentuk teks adalah informasi yang penting dan banyak didapatkan dari berbagai sumber seperti buku, surat kabar, situs web, ataupun pesan *e-mail*. Teks merupakan sebuah hamparan bahasa, baik dalam pembicaraan ataupun dalam tulisan, yang memiliki makna, bersifat praktis dan berguna untuk umum serta berhubungan dengan dunia nyata (Bolshakov & Gelbukh [2]). Sebuah teks dapat terdiri dari hanya satu kata ataupun susunan kalimat (Carter & McCarthy [3]). Pengambilan informasi dari teks (*text mining*) antara lain dapat meliputi kategorisasi teks atau dokumen, analisis sentimen (*sentiment analysis*), pencarian topik yang lebih spesifik (*search engine*), serta *spam filtering*. Gagasan umum *text mining* adalah untuk mengetahui cakupan atau topik dari permasalahan dalam teks (Maning, *et al*. [6]). T*ext mining* penting dalam analisis sentimen sebagai pengidentifikasi emosional suatu pernyataan, sehingga banyak studi tentang analisis sentimen dilakukan (Zhang, *et al*. [12]).

Analisis sentimen adalah studi komputasi dari opini-opini, sentimen, serta emosi yang diekspresikan dalam teks (Liu [5]). Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau pendapat. Polaritas mempunyai arti apakah teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif. Salah satu

teknik pembelajaran mesin untuk analisis sentimen adalah *Naïve Bayes classifier* (NBC). NBC merupakan teknik pembelajaran mesin yang berbasis probabilistik. NBC adalah metode sederhana tetapi memiliki akurasi serta performansi yang tinggi dalam pengklasifikasian teks (Routray, *et al*. [9]).

Banyak ide telah muncul selama beberapa tahun belakangan tentang teknik pembelajaran mesin untuk permasalahan analisis sentimen. Xhemali, *et al*. [11] berkonsentrasi pada perbandingan tiga metode. Metode-metode tersebut adalah Naïve Bayes, Pohon Keputusan, dan Neural Networks. Hasil penelitian secara keseluruhan menunjukkan bahwa Naïve Bayes *classifier* adalah pilihan terbaik untuk pelatihan domain. Metode berbasis leksikon untuk melakukan analisis sentimen pertama kali diterapkan oleh Zhang *et al*. [12]. Metode ini dapat memberikan presisi yang tinggi tapi *recall* yang rendah. Penelitian lain yang dilakukan Taboada, *et al*. [10] yang menerapkan pendekatan berbasis leksikon untuk mengekstrak sentimen dari teks yang menggunakan kamus kata-kata atau frase dijelaskan dengan orientasi semantik meliputi polaritas dan *strength* dari kata-kata, serta menggabungkan intensifikasi dan negasi. Routray *et al*. [9] dan Khairnar & Kinikar [4] membahas banyak pendekatan dari para peneliti yang berbeda, serta menyatakan bahwa metode pembelajaran mesin menjadi cara yang efisien untuk menganalisis sentimen.

Penggabungan yang dilakukan dalam tulisan ini adalah menggabungkan NBC dengan seleksi fitur. Penyeleksian fitur diperlukan dalam proses memilih subset dari fitur-fitur yang relevan untuk digunakan dalam konstruksi model probabilistik NBC. Penyeleksian fitur yang digunakan adalah seleksi fitur *chi square*. Dari data yang tersedia, sejumlah data akan digunakan untuk menguji hasil klasifikasi sistem NBC dengan penyeleksian fitur *chi square*.

2. KAJIAN PUSTAKA

2.1 *Pre-Processing*

Tokenization adalah tugas pemotongan urutan karakter dan sebuah set dokumen yang diberikan menjadi potongan-potongan kata atau karakter yang sesuai dengan kebutuhan sistem. Potongan-potongan tersebut dikenal dengan istilah token (Maning, et al. [6]).

Stemming merupakan salah satu proses dari mengubah token yang berimbuhan menjadi kata dasar, dengan menghilangkan semua imbuhan yang ada pada token tersebut. Pentingnya *stemming* dalam proses pembuatan sistem adalah untuk menghilangkan imbuhan pada awalan dan akhiran. Berdasarkan hasil proses tersebut, akan didapatkan sebuah informasi mengenai banyaknya fitur yang muncul dalam sebuah dokumen.

Stopwords dapat diartikan sebagai menghilangkan karakter, tanda baca, serta kata-kata umum yang tidak memiliki makna atau informasi yang dibutuhkan. *Stopwords* umumnya digunakan dalam pengambilan informasi salah satu contohnya adalah mesin pencari *Google*. Pengurangan ukuran indeks dalam teks dengan penghilangan beberapa kata kerja, kata sifat, dan kata keterangan lainnya dapat dimasukkan ke dalam daftar *stopwords*.

2.2 Seleksi Fitur *Chi Square*

Seleksi fitur dilakukan untuk mereduksi fitur-fitur yang tidak relevan dalam proses klasifikasi oleh NBC. Terdapat beberapa metode untuk penyeleksian fitur yaitu *Mutual Information* (MI), *chi square* (χ^2), dan yang umum digunakan adalah *frequency-based*. Seleksi fitur *frequency-based* menggunakan jumlah kemunculan *term* atau frekuensi *term* yang diurutkan dari yang paling banyak sampai paling sedikit dan diambil beberapa urutan atas untuk digunakan sebagai fitur. Seleksi fitur MI merupakan ukuran yang mengukur kehadiran atau ketidakhadiran sebuah *term* yang memberikan kontribusi kepada kategori yang tepat. Sedangkan seleksi

fitur *Chi Square* menggunakan teori statistika untuk menguji independensi sebuah *term* dengan kategorinya. Salah satu tujuan penggunaan seleksi fitur adalah untuk menghilangkan fitur pengganggu dalam klasifikasi.

Dalam seleksi fitur *Chi Square* berdasarkan teori statistika, dua peristiwa di antaranya adalah, kemunculan dari fitur dan kemunculan dari kategori, yang kemudian setiap nilai *term* diurutkan dari yang tertinggi berdasarkan perhitungan berikut [6]:

$$X^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

Penyeleksian fitur *Chi Square* dilakukan dengan cara mengurutkan setiap fitur berdasarkan hasil seleksi fitur *Chi Square* dari nilai yang terbesar hingga nilai yang terkecil. Nilai seleksi fitur *Chi Square* yang lebih besar dari nilai signifikan menunjukkan penolakan hipotesis independensi. Sedangkan jika dua peristiwa menunjukkan dependen, maka fitur tersebut menyerupai atau sama dengan label kategori yang sesuai pada kategori.

2.3 Naïve Bayes Classifier

Naïve Bayes *classifier* merupakan suatu metode klasifikasi yang menggunakan perhitungan probabilitas. Konsep dasar yang digunakan pada Naïve Bayes *classifier* adalah Teorema Bayes yang dinyatakan pertama kali oleh Thomas Bayes [1]. Nilai probabilitas yang digunakan dinyatakan secara sederhana sebagai berikut (Pop [8]):

$$p(C \mid D) = \frac{p(D \mid C)p(C)}{p(D)}.$$

2.4 Evaluasi Kinerja

Dua dasar ukuran yang sering digunakan untuk mengetahui efektivitas sistem adalah *precision* atau presisi dan *recall*. Presisi (*P*) adalah ukuran banyaknya dokumen yang ditemukan relevan, dinyatakan dalam pecahan sebagai berikut;

$$Precision = \frac{\#(Relevant\ item\ retrieved)}{\#(Retrieved\ item)}$$

sedangkan *recall* (*R*) adalah ukuran banyaknya dokumen yang relevan dapat ditemukan kembali, dinyatakan dalam pecahan sebagai berikut;

$$Recall = \frac{\#(Relevant\ item\ retrieved)}{\#(Relevant\ item)}$$

3. METODE PENELITIAN

Data pada penelitian ini berupa opini berbahasa Inggris tentang produk telepon genggam. Dari data yang tersedia, diambil secara acak sebanyak 200 buah opini yang terdiri dari 100 buah opini positif dan 100 buah opini negatif. Data tersebut digunakan sebagai data pembelajaran mesin dan data uji untuk mengevaluasi kinerja sistem.

Adapun langkah-langkah untuk perancangan analisis sentimen yang dibahas adalah sebagai berikut:

1. Tahap *pre-processing* data

Pada tahap *pre-processing* data, awal mula data mentah dilakukan proses *tokenizer*, *stemming*, serta *stopwords*. Hasil dari tahapan ini menghasilkan fitur yang digunakan sebagai data pembelajaran mesin oleh NBC.

2. Tahap penyeleksian fitur dengan seleksi fitur *Chi Square*

Penelitian ini menggunakan seleksi fitur *Chi Square*. Langkah awal, yaitu menentukan tabel kontingensi masing-masing fitur dengan Tabel 1. Langkah selanjutnya menghitung nilai seleksi fitur *Chi Square* dengan persamaan berikut (Maning, et al. [6]):

$$X^2(D,t,c) = \frac{(N_{00}+N_{11}+N_{10}+N_{01}) \times (N_{00}N_{11} - N_{10}N_{01})^2}{(N_{11}+N_{01}) \times (N_{11}+N_{10}) \times (N_{10}+N_{00}) \times (N_{01}+N_{00})}.$$

Seleksi fitur *Chi Square* digunakan untuk pegamatan kebersesuaian (*goodness of fit*) dari kategori dengan *terms*. Uji *Chi Square* dalam statistika diterapkan untuk menguji independensi dari dua peristiwa. Sedangkan

dalam seleksi fitur berdasarkan teori statistika, dua peristiwa tersebut di antaranya adalah, kemunculan dari fitur dan kemunculan dari kategori.

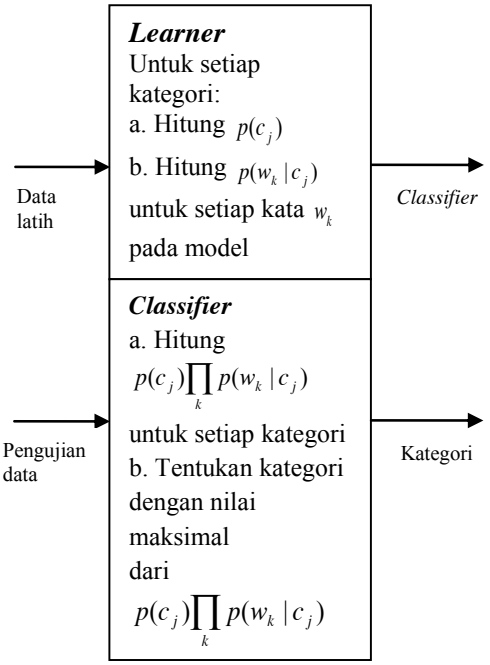
Tabel 1. Tabel Kontingensi Seleksi Fitur *Chi Square*

	$e_c = 1$	$e_c = 0$
$e_t = 1$	N_{11}	N_{10}
$e_t = 0$	N_{01}	N_{00}

Langkah terakhir yaitu mengurutkan semua hasil perhitungan seleksi fitur dari yang terbesar sampai yang terkecil. Penghapusan fitur dilakukan jika penerimaan hipotesis independen terpenuhi.

3. Tahap klasifikasi data

Naïve Bayes menganggap sebuah dokumen sebagai kumpulan dari kata yang menyusun dokumen tersebut. Naïve Bayes juga tidak memperhatikan urutan kemunculan kata pada dokumen. Adapun algoritma NBC [7] untuk klasifikasi data dapat dilihat pada Gambar 1.



Gambar 1. Proses Klasifikasi dengan NBC

4. Tahap evaluasi kinerja sistem

Pada tahap evaluasi sistem, perhitungan menggunakan tabel kontingensi yang diberikan pada Tabel 2.

Tabel 2. Tabel Kontingensi Evaluasi Kinerja Sistem

#	*Relevant*	*Not Relevant*
Retrieved	*True positive (TP)*	*False Positive (FP)*
Not Retrieved	*False negative (FN)*	*True Negative (TN)*

Alternatif yang jelas terlintas pada pikiran pembaca dalam menilai sebuah sistem adalah dengan akurasi. Akurasi adalah ketepatan suatu sistem melakukan klasifikasi yang benar. Perhitungan untuk Presisi (*P*), *Recall* (*R*), dan akurasi dapat dikalkulasi sebagai berikut:

$$P = \frac{TP}{TP + FP};$$

$$R = \frac{TP}{TP + FN};$$

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN}.$$

Sebuah ukuran yang digunakan sebagai rata-rata terbobot harmonik dari *P* dan *R* adalah sebagai berikut:

$$F_1 = \frac{2PR}{P + R}.$$

4. HASIL DAN PEMBAHASAN

Penggunaan data sebanyak 200 buah opini tentang opini berbahasa Inggris pengguna telepon genggam yang terbagi atas 100 buah opini positif dan 100 buah opini negatif. Sebagai data latih digunakan sejumlah 100 buah data yaitu data yang terbagi masing-masing 50 buah opini positif dan 50 buah opini.

Sisanya yaitu 100 buah opini yang terbagi sama rata antara opini positif dan opini negatif digunakan sebagai data uji.

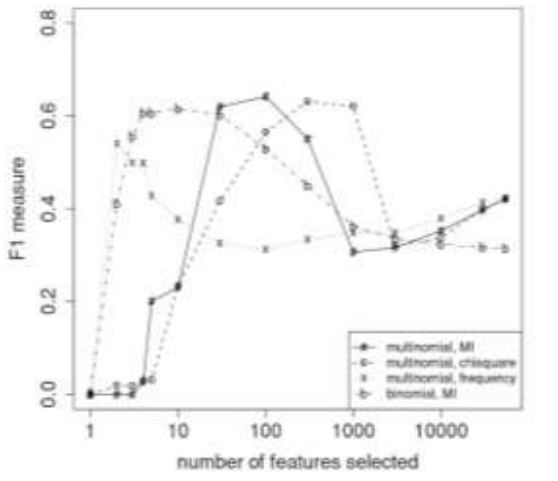
Perancangan basis data dengan menggunakan XAMPP, memuat tiga buah tabel yang independen, yaitu tabel datatraining, corpus, tb_feature. Tabel datatraining memuat keseluruhan data latih mentah. Tabel tb_feature memuat fitur-fitur yang telah diberi label positif atau negatif secara manual. Sedangkan tabel corpus memuat fitur-fitur yang telah diseleksi melalui tahapan yang telah diuraikan pada implementasi seleksi fitur *Chi Square*. Pada sistem, proses-proses atau urutan proses dirancang untuk mengklasifikasi data uji dengan melalui beberapa tahapan. Tahapan tersebut merupakan tahapan yang telah diuraikan pada tahap *pre-processing* dan implementasi Naïve Bayes *classifier*.

Langkah awal yaitu melakukan pelabelan secara manual dari 100 buah data latih. Pada pelabelan fitur tersebut dilakukan proses *stemming*. Pelabelan tersebut menghasilkan 117 fitur diantaranya 66 fitur negatif dan 51 fitur positif.

Setelah melakukan pelabelan dilanjutkan dengan proses penyeleksian fitur. Seleksi fitur dilakukan dari sebanyak 117 fitur yang diperoleh dari pelabelan. Pada tahap seleksi fitur ada tiga metode yang diketahui, yaitu *Mutual Information* (MI), *chi square* (χ^2), dan *frequency-based*. Seleksi fitur *frequency-based* menggunakan jumlah kemunculan *term* atau frekuensi *term* yang diurutkan dari yang paling banyak sampai paling sedikit dan diambil beberapa urutan atas untuk digunakan sebagai fitur.

Metode seleksi fitur *Frequency-based* memilih fitur yang paling umum di kategori. Metode seleksi fitur *Frequency-based* dapat didefinisikan baik sebagai frekuensi dokumen (jumlah dokumen di kategori *c* yang mengandung fitur *t*) atau sebagai koleksi frekuensi (jumlah token dari *t* yang muncul pada dokumen dalam *c*). Seleksi fitur MI merupakan ukuran yang mengukur kehadiran atau ketidakhadiran sebuah *term* yang memberikan kontribusi kepada kategori yang tepat. Sedangkan seleksi fitur *Chi Square* menggunakan teori statistika untuk menguji independensi sebuah *term* dengan kategorinya. Salah satu tujuan penggunaan seleksi fitur adalah untuk menghilangkan fitur pengganggu dalam klasifikasi. Maning, *et al*. [6] mengatakan untuk kasus seleksi fitur *frequency-based* memiliki kinerja yang buruk dibandingkan MI dan *Chi Square*.

Perbandingan dari peningkatan akurasi dapat diamati pada Gambar 2, dengan $\boldsymbol{F_{1\text{-measure}}}$ merupakan ukuran ketepatan klasifikasi oleh metode Naïve Bayes *Classifier* dengan dilakukannya beberapa metode seleksi fitur yaitu *MI*, *Chi Square* dan *frequency-based*. Pada Gambar 2, saat 100 buah fitur terpilih, ketepatan klasifikasi oleh metode seleksi fitur *frequency-based* memperoleh di bawah 40%, metode seleksi fitur *MI* memperoleh hasil di atas 60%, dan metode seleksi fitur *chi square* memperoleh hasil yang mendekati 60% [6]. Hal ini menunjukkan bahwa metode *frequency-based* tidak seharusnya digunakan.

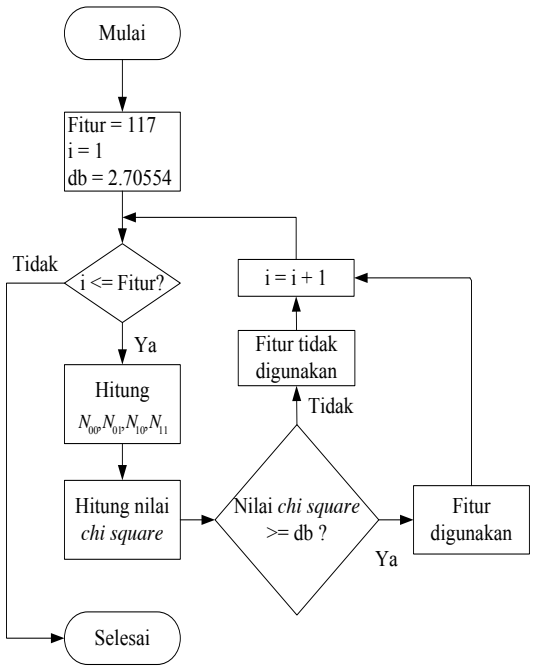


Gambar 2. Perbandingan Metode Seleksi Fitur dalam Ketepatan Klasifikasi

Hipotesis awal menyatakan bahwa *term t* independen terhadap kategori *c*. Sedangkan hipotesis akhir menyatakan bahwa *term t* dependen terhadap kategori *c*. Hasil seleksi fitur *Chi Square* dari 117 fitur terseleksi menjadi 30 fitur yang terdiri dari 14 fitur

negatif dan 16 fitur positif. Proses penyeleksian fitur sebanyak 117 fitur dengan seleksi fitur *Chi Square* yang dibangun dengan bahasa pemrograman Java membutuhkan waktu komputasi hanya dua detik.

Diagram alir untuk merepresentasikan langkah-langkah sistem untuk melakukan proses penyeleksian fitur dengan seleksi fitur *Chi Square* dapat dilihat pada Gambar 3 berikut:



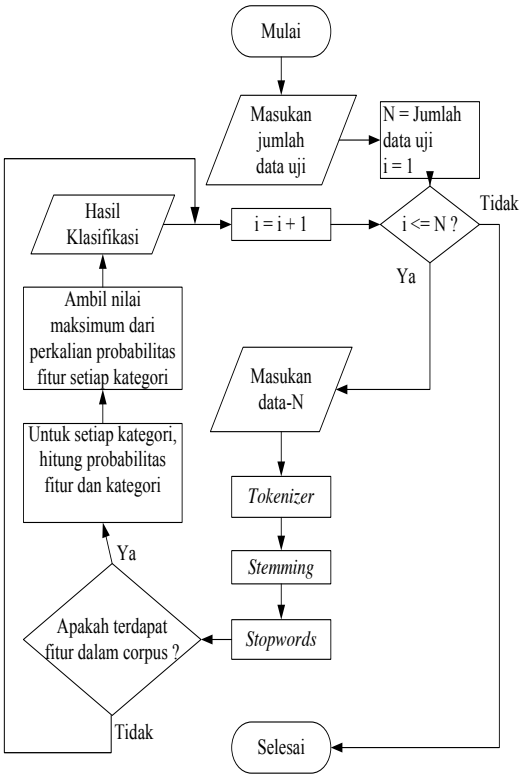
Gambar 3. Diagram Alir Penyeleksian Fitur dengan *Chi Square*

Dalam seleksi fitur *Chi Square* frekuensi pada fitur menjadi kurang penting bila fitur tersebut juga muncul beberapa kali pada kategori yang tidak diharapkan. Sehingga bila kedua metode tersebut dibandingkan seleksi fitur *Chi Square* akan lebih baik daripada *frequency-based*. Kemunculan frekuensi fitur pada kategori yang diharapkan dan kategori yang tidak diharapkan memiliki peranan penting dalam seleksi fitur *Chi Square*. Sedangkan pada *frequency-based* yang memiliki peranan penting hanya kemunculan frekuensi fitur pada kategori yang diharapkan. Apabila penolakan hipotesis awal terpenuhi atau penerimaan hipotesis akhir terpenuhi, fitur tersebut akan digunakan dalam proses klasifikasi NBC.

Mengacu pada konsep dasar Naïve Bayes *classifier* yaitu Teorema Bayes yang dinyatakan pertama kali oleh Thomas Bayes [1]:

$$p(C \mid D) = \frac{p(D \mid C)p(C)}{p(D)}.$$

Nilai probabilitas dihitung dari kemunculan opini yang setara dengan perkalian nilai probabilitas kemunculan fitur dalam opini tersebut. Adapun diagram alir untuk proses klasifikasi dengan NBC dapat dilihat pada Gambar 4.



Gambar 4. Diagram Alir Proses Klasifikasi dengan NBC

Klasifikasi oleh NBC pada data uji negatif memperoleh ketepatan sebesar 72% sedangkan untuk data uji positif memperoleh 96%. Secara keseluruhan data uji, klasifikasi oleh NBC

memperoleh akurasi sebesar 83% dan rata-rata harmonik ($F_{measure}$) sebesar 90,713%. Pada penelitian ini, NBC melakukan kesalahan klasifikasi sebanyak empat data uji, tidak dapat mengklasifikasi 13 data uji dari total keseluruhan 100 data uji.

NBC bekerja dengan baik dalam mengklasifikasi dan juga merupakan teknik pembelajaran mesin yang sederhana dengan hanya menggunakan kemunculan fitur serta frekuensi fitur pada tiap-tiap opini. Secara global klasifikasi oleh NBC dapat digunakan sebagai teknik analisis sentimen pasar produk seperti yang dilakukan pada penelitian ini.

Tabel 3. Peringkat Hasil Seleksi Fitur *Chi Square*

No	Fitur	Kategori	Frekuensi Fitur	Nilai chi square
1	*dissapoint*	negatif	29	40.84507042
2	*great*	positif	26	35.13513514
3	*love*	positif	22	28.20512821
4	*recommend*	positif	16	15.94613749
5	*good*	positif	16	15.94613749
6	*us*	negatif	10	11.11111111
7	*excel*	positif	10	11.11111111
8	*amaz*	positif	8	8.695652174
9	*perfect*	positif	8	8.695652174
10	*easi*	positif	8	8.695652174
11	*lot*	positif	10	8.273748723
12	*return*	negatif	7	7.52688172
13	*broken*	negatif	7	7.52688172
14	*bad*	negatif	7	7.52688172
15	*upset*	negatif	6	6.382978723
16	*fix*	negatif	4	4.166666667
17	*highli*	positif	4	4.166666667
18	*best*	positif	4	4.166666667
19	*awesom*	positif	4	4.166666667
20	*pai*	negatif	3	3.092783505
21	*poor*	negatif	3	3.092783505
22	*cheat*	negatif	3	3.092783505
23	*old*	negatif	3	3.092783505
24	*damag*	negatif	3	3.092783505
25	*over*	negatif	3	3.092783505
26	*sure*	positif	3	3.092783505
27	*beauti*	positif	3	3.092783505
28	*like*	positif	10	2.990033223
29	*unlock*	negatif	5	2.836879433
30	*fast*	positif	5	2.836879433

Daftar hasil perhitungan memuat nilai seleksi fitur *Chi Square* yang berdasarkan hipotesis independensi, dengan hipotesis awal menyatakan bahwa *term t* independen terhadap kategori *c*. Apabila nilai seleksi fitur *Chi Square* lebih besar daripada nilai signifikan, sehingga penolakan hipotesis awal akan terpenuhi. Hipotesis akhir yang diperoleh menyatakan bahwa *term t* dependen terhadap kategori *c*.

5. SIMPULAN

Berdasarkan hasil yang diperoleh dapat disimpulan bahwa kemunculan frekuensi fitur pada kategori yang diharapkan dan kategori yang tidak diharapkan memiliki peranan penting dalam seleksi fitur *Chi Square*, oleh karena itu seleksi fitur *Chi Square* baik digunakan dalam penyeleksian fitur dibandingkan dengan metode *frequency-based*. Serta pembangunan sistem analisis sentimen menggunakan metode NBC dengan bahasa pemrograman Java memperoleh akurasi sebesar 83% dan rata-rata harmonik sebesar 90,713%. Terdapat kesalahan klasifikasi karena pada data uji terdapat fitur yang muncul pada bukan kategorinya. Untuk penelitian selanjutnya yaitu dapat dibandingkan hasil seleksi fitur dari *chi square* terhadap hasil seleksi fitur dari *Mutual Information* berdasarkan segi waktu komputasi dan segi ketepatan klasifikasi. Serta menggabungkan teknik pembelajaran mesin NBC dengan beberapa model *n*-gram untuk meneliti hasil ketepatan klasifikasi yang diperoleh.

DAFTAR PUSTAKA

[1] Aldrich, J., 2008. R. A. Fisher on Bayes and Bayes' Theorem. *Bayesian Analysis*, 3(1), pp. 161-170.

[2] Bolshakov, I. A. & Gelbukh, A., 2004. *Computational Linguistics*. 1st ed.

Mexico: Instituto Politécnico Nacional.
[3] Carter, R. & McCarthy, M., 2006. *Cambridge Grammar of English*. Cambridge: Cambridge Univ. Press.
[4] Khairnar, J. & Kinikar, M., 2013. Machine Learning Algorithms for Opinion Mining and Sentiment Classification. *International Journal of Scientific and Research Publications*, June, 3(6), pp. 1-6.
[5] Liu, B., 2012. *Sentiment Analysis and Opinion Mining*. San Rafael: Morgan & Claypool Publishers.
[6] Maning, C., Raghavan, P. & Schutze, H., 2008. *Introduction to Information Retrieval*. London: Cambridge University Press.
[7] Mitchell, T. M., 1997. *Machine Learning*. 1st ed. New York: McGraw-Hill.
[8] Pop, I., 2006. An approach of the Naive Bayes classifier for the document classification. *General Mathematics*, 14(4), p. 135–138.
[9] Routray, P., Swain, C. K. & Mishra, S. P., 2013. A Survey on Sentiment Analysis. *International Journal of Computer Applications*, Agustus, 70(10), pp. 1-8.
[10] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M., 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), pp. 267-307.
[11] Xhemali, D., J. Hinde, C. & G. Stone, R., 2009. Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science Issues*, 4(1), pp. 16-23.
[12] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. & Liu,B., 2011. *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis*, Chicago: Hewlett-Packard Development Company, L.P.