Analisis Sentimen Menggunakan *Naive Bayes Classifier* dengan *Chi-Square Feature* Selection Terhadap Penyedia Layanan Telekomunikasi

Ainun Nisa[1], Eko Darwiyanto,S.T.,M.T.[2], Ibnu Asror,S.T.,M.T.[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]ainunnisa@students.telkomuniversity.ac.id, [2]ekodarwiyanto@telkomuniversity.ac.id,
[3]iasror@telkomuniversity.ac.id

Abstrak

Pendapat masyarakat terhadap penyedia layanan telekomunikasi merupakan sesuatu yang dapat digunakan sebagai bahan pertimbangan untuk membuat keputusan, baik bagi pengguna maupun pihak perusahaan. Analisis sentimen merupakan bidang studi yang meneliti tentang opini terhadap suatu objek, dimana opini tersebut dapat diklasifikasikan berdasarkan polaritas yang terkandung di dalamnya. Penelitian ini melakukan klasifikasi menggunakan metode naive bayes terhadap opini masyarakat tentang penyedia layanan telekomunikasi. Dimensionalitas data yang tinggi pada klasifikasi menggunakan naive bayes dapat dikurangi dengan seleksi fitur *chi square*. Hasil penelitian menunjukkan rata-rata performansi tertinggi didapatkan oleh klasifikasi menggunakan metode naive bayes dengan seleksi fitur *chi square* dengan tingkat signifikansi 0,01 yaitu akurasi 85,5%, presisi 83%, *recall* 86% dan *f1-score* 84%. Seleksi fitur *chi square* tidak memberikan perbedaan yang signifikan terhadap klasifikasi menggunakan naive bayes.

Kata kunci : analisis sentimen, *naive bayes classifier, chi square*

Abstract

Public opinion on telecommunications service providers is something that can be used as a consideration for making decisions, both for users and the company. Sentiment analysis is a field of study that examines opinions on an object, where opinions can be classified based on the polarity contained in them. This research classifies use Naive Bayes method on public opinion about telecommunications service providers. High data dimensions in classification using Naive Bayes can be reduced by the chi square feature selection. The results showed that the highest average performance obtained by classification using naive bayes with chi square feature selection with significance level of 0,01, get 85.5% accuracy, 83% precision, 86% recall and 84% f1-score. Chi square feature selection did not give a significant difference to the classification using naive bayes.

Keywords: sentiment analysis, naive bayes classifier, chi square

1. Pendahuluan

1.1 Latar Belakang

Analisis sentimen adalah bidang studi yang menganalisis opini, penilaian dan sikap terhadap entitas seperti produk, layanan, organisasi, individu, topik, dan atributnya [1]. Analisis sentimen merupakan klasifikasi yang mengelompokkan teks berdasarkan opini atau sentimen yang terkandung di dalamnya [2]. Proses klasifikasi dapat dilakukan dengan menggunakan metode naive bayes. Konsep naive bayes dapat dikatakan sederhana dan intuitif, metode yang efisien dan memberikan akurasi yang tinggi [2, 3, 4].

Di samping penggunaan metode naive bayes yang terbukti memberikan hasil klasifikasi yang baik pada banyak penelitian, data opini yang digunakan adalah data tidak terstruktur, dan banyaknya jumlah fitur dapat menimbulkan masalah tingginya dimensionalitas data. Dimensionalitas data yang tinggi dapat menyebabkan data *overload*, menyimpang dan dapat mempengaruhi proses klasifikasi [5]. Maka dari itu diperlukan proses untuk memilih fitur yang relevan, dengan kriteria untuk mendapatkan fitur yang optimal [6]. Seleksi fitur dapat dilakukan dengan melihat nilai berdasarkan hasil evaluasi dengan kriteria tertentu [5]. Seleksi fitur yang pada beberapa penelitian [7, 8, 9] memberikan performansi yang baik adalah metode chi-square dengan konsep untuk melihat ketergantungan antar dua variabel. Oleh karena itu pada penelitian ini seleksi fitur akan dilakukan dengan *chi square*, dan hasilnya akan dipakai untuk klasifikasi sentimen menggunakan naive bayes.

Sehubungan dengan analisis sentimen, berdasarkan data hingga bulan Januari 2019, pengguna ponsel terdaftar di Indonesia mencapai 355,5 juta atau 133% dari populasi Indonesia yang berjumlah 268,2 juta jiwa [10]. Berdasarkan data tersebut, sentimen masyarakat tentang penyedia layanan telekomunikasi dapat dijadikan data untuk dianalisis menggunakan metode klasifikasi naive bayes dan seleksi fitur *chi square*.

1.2 Topik dan Batasan

Berdasarkan latar belakang, masalah yang dapat dirumuskan pada penelitian ini adalah bagaimana pengaruh seleksi fitur *chi square* terhadap metode naive bayes apabila diterapkan pada klasifikasi opini berbahasa Indonesia terhadap penyedia layanan telekomunikasi. Studi kasus yang dilakukan terbatas pada penyedia layanan telkomunikasi terkemuka di Indonesia yang dilansir dari situs *Finder* [11] yaitu Telkomsel, IM3 Ooredoo, XL Axiata dan Tri. Data yang digunakan berupa tweet terhadap akun resmi penyedia layanan telekomunikasi terpilih dengan username Twitter @Telkomsel, @IM3Ooredoo, @triindonesia, dan @myXLCare. Data teks yang digunakan sebagai dataset klasifikasi berjumlah 1200 tweet, dimana berdasarkan penelitian sebelumnya, klasifikasi dengan naive bayes memberikan akurasi tinggi dengan 70% data latih dan 30% data uji [12]. Metode naive bayes, berdasarkan persebaran datanya dapat dibedakan menjadi tiga jenis [13], yaitu bernoulli naive bayes, gaussian naive bayes, dan satu lagi yang digunakan pada penelitian ini yaitu multinomial naive bayes yang cocok digunakan dalam pengklasifikasian data teks atau dokumen [14]. Selanjutnya, pengukuran yang digunakan untuk melihat performansi metode yang diterapkan adalah akurasi, presisi, *recall dan f1-score*.

1.3 Tujuan

Pendapat seseorang terhadap penyedia layanan telekomunikasi sekarang makin mudah untuk disampaikan secara langsung ke akun resmi Twitter penyedia telekomunikasi terkait. Pendapat tersebut dapat dijadikan data untuk selanjutnya diklasifikasikan berdasarkan polaritas pendapat yang terkandung di dalamnya. Naive Bayes adalah metode klasifikasi dengan performansi yang baik [2, 3, 4], dengan asumsi ketidaktergantungan atau independensi yang kuat. *Chi square* merupakan metode statistik yang mengukur seberapa dependen atau tingkat ketergantungan antara dua variabel, dimana pada kasus ini berarti ketergantungan antara *term* dengan kategori [15] yang pada beberapa penelitian sebelumnya [7, 8, 9] dapat meningkatkan performansi untuk proses klasifikasi.

Tujuan dari penelitian ini adalah untuk mengetahui bagaimana pengaruh seleksi fitur *chi square* terhadap metode klasifikasi naive bayes untuk data opini berbahasa Indonesia tentang penyedia layanan telekomunikasi.

1.4 Organisasi Tulisan

Organisasi penulisan ini dimulai dengan latar belakang, topik, batasan dan tujuan penelitian yang termasuk dalam pendahuluan. Selanjutnya ada studi terkait yang berisi tentang penelitian-penelitian yang berhubungan dengan analisis sentimen dan metode yang digunakan. Ketiga, sistem yang dibangun, gambaran umum sistem beserta penjelasan setiap tahapnya . Keempat, dituliskan hasil implementasi dan analisis berdasarkan penelitian yang dilakukan. Kelima adalah poin kesimpulan serta saran.

2. Studi Terkait

2.1 Analisis Sentimen

Analisis sentimen merupakan bidang studi yang menganalisis pendapat, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya [1]. Definisi lain juga mengatakan analisis sentimen adalah klasifikasi teks yang mengelompokkan teks berdasarkan opini yang terkandung di dalamnya [2].

Analisis sentimen merupakan salah satu bagian penting dari *Natural Language Processing* [2]. *NLP* adalah bidang ilmu komputer dan kecerdasan buatan yang terutama berhubungan dengan interaksi bahasa manusia-komputer [2]. Analisis sentimen memerlukan penggunaan data latih dan data uji untuk kinerjanya, dan kualitas data latih tersebut memiliki peran dalam evaluasi teks yang akurat.

2.2 Naive Bayes

Naive Bayes Classifier (NBC) merupakan metode dengan algoritma yang dapat melakukan klasifikasi data pada kelas tertentu. Setiap atribut memiliki bobot yang sama penting, saling lepas satu sama lain, dan memiliki peran untuk mendukung pengambilan keputusan [16]. Metode naive bayes yang cocok digunakan dalam pengklasifikasian teks atau dokumen adalah multinomial naive bayes, dimana tidak hanya melihat kata yang muncul namun juga jumlah kemunculannya [14].

$$C_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{x \in X} P(t_k|c) \quad (1)$$

Rumus *Maximum a Posterior (MAP)* di atas digunakan untuk menentukan kelas suatu data uji dengan mengambil nilai maksimum probabilitas setiap dokumen [13]. Dimana $P(c)$ adalah *prior probability* yang dihitung dengan membagi suatu kelas dalam data latih (N_c) dengan total kelas yang digunakan (N) [13]:

$$P(c) = \frac{N_c}{N} \quad (2)$$

Dan $P(t_k|c)$ sebagai *conditional probability* yang menghitung kemungkinan munculnya kata dalam setiap kelas, yang dapat dilakukan dengan perhitungan berikut :

$$P(t_i|c_j) = \frac{count(t_i, c_j) + 1}{(\sum_{w \in V} count(t, c)) + |V|} \quad (3)$$

Untuk mendapatkan C_{MAP}, *conditional probability* dikalikan, yang jika pembaginya sangat besar, dapat menyebabkan *floating point underflow*. Oleh karena itu, probabilitas kata terlebih dahulu digunakan logaritma untuk menghindari *floating point underflow* [13] :

$$C_{MAP} = \underset{c \in C}{\operatorname{argmax}} \left[\log P(c) + \sum_{1 \le k \le nd} \log P(t_k|c) \right] \quad (4)$$

2.3 Seleksi Fitur *Chi Square*

Seleksi fitur merupakan proses untuk memilih fitur yang relevan, dengan kriteria untuk mendapatkan fitur yang optimal, karena pada data dengan dimensi tinggi, menemukan subset fitur yang optimal adalah tugas yang cukup sulit [6]. Beberapa manfaat yang dapat diperoleh dari pemilihan fitur antara lain : mengurangi kebutuhan penyimpanan (*storage*), mengurangi *training*, membantah kutukan dimensi (*curse of dimensionality*), dan meningkatkan performansi [17].

Chi square merupakan seleksi fitur yang melihat ketergantungan *term* dengan kategorinya, diamana beberapa syarat uji *chi square* dapat digunakan yaitu [18] :

i. Tidak ada sel dengan nilai *actual count* atau frekuensi kenyataan (F_0) yang 0 (nol);
ii. Apabila tabel kontingensi berbentuk 2x2, makatidak boleh ada sel dengan *expected count* atau frekuensi harapan (F_h) kurang dari 5;
iii. Apabila bentuk tabel lebih dari 2 x 2, misal 2 x 3, maka jumlah sel dengan frekuensi harapan (F_h) yang kurang dari 5 tidak boleh lebih dari 20%.

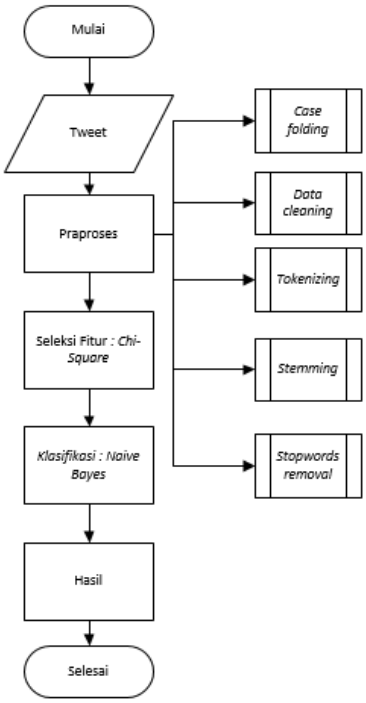
Uji *chi square* dapat dirumuskan sebagai berikut [18] :

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (5)$$

dengan
O_i = nilai observasi ke-i
E_i = nilai ekspektasi ke-i

3. Sistem yang Dibangun

Sistem pada penelitian ini terdiri dari beberapa proses. Adapun gambaran umum sistem yang dibangun ditunjukkan oleh gambar di bawah :



Gambar 1. Gambaran umum sistem

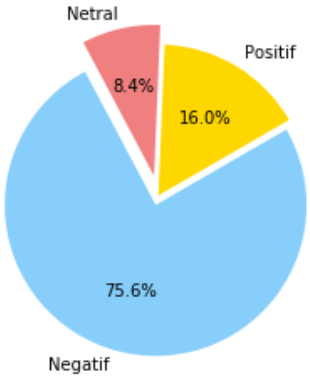
i. Dataset

Sebelum masuk ke tahap klasifikasi, data terlebih dahulu diberi label dengan angka -1, 0, dan 1, untuk menunjukkan masing-masing kelas atau kategori. Adapun arti dari label tersebut adalah angka -1 menunjukkan opini termasuk kategori negatif, 0 adalah netral, dan 1 menunjukkan opini positif. Berikut adalah contoh dataset yang digunakan :

Tabel 1. Contoh dataset

Tweet	**Label**	**Provider**
@myXLCare xl bbrp hr ni lemot trus signalnya, tanda mo mati ya ??l	-1	XL
Baru saja mengirim foto @ Telkomsel Graha Pena Fajar, Urip Sumoharjo no.20 https://t.co/oXEavaHt7i	0	Telkomsel
@vickylaurentina @IM3Ooredoo Di beberapa titik tetap bagus. Di rumah ga masalah sih, tetap lancar	1	IM3Ooredoo
@3CareIndonesia Yang pertamaa dong min. Tapi tri baik bangeeet, masa aktif pulsaku masih sampe Maret 2025 hehe \ud83d\ude02	1	Tri

Persentase data tweet yang diperoleh ditunjukkan oleh gambar berikut :



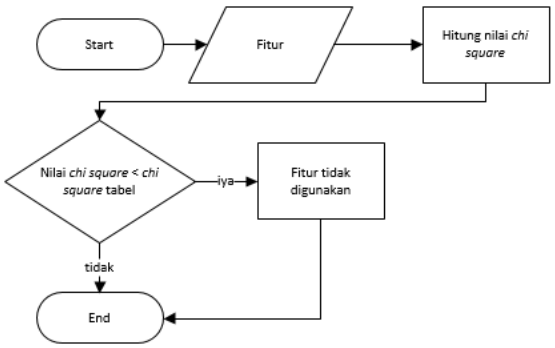
Gambar 2. Persentase kategori dataset

Total data yang digunakan adalah 1200 tweet berbahasa Indonesia dengan jumlah kelas positif sebanyak 203, kelas negatif 928, dan untuk kelas netral sebanyak 69 tweet.

ii. Preprocessing

Opini yang disampaikan melalui tweet mengandung banyak kata yang dianggap sebagai noise dalam proses klasifikasi. Preprocessing atau praproses bertujuan untuk mempersiapkan data mentah untuk mengurangi kata yang dianggap kurang cocok untuk dilanjutkan dalam proses klasifikasi. Tahapan praproses pada penelitian ini adalah *case folding*, *data cleaning*, *tokenizing*, *stemming*, dan *stopword removal*. *Case folding* akan mengkonversi semua data tweet menjadi huruf kecil. *Data cleansing* terdiri dari penghapusan *url*, *username*, *character* dan *short words.* Dilanjutkan dengan *tokenizing* yang membagi data menjadi kata per kata, kemudian dilakukan *stemming* atau penghapusan awalan dan / atau akhiran kata dengan aturan bahasa Indonesia menggunakan sastrawi *stemmer*. Selanjutnya adalah menghapus *stopword*, atau kata-kata yang sering muncul dalam dokumen tetapi dianggap tidak berguna untuk proses klasifikasi.

iii. Chi square feature selection



Gambar 3. Cara kerja seleksi fitur *chi square*

Setelah melalui praproses, selanjutnya adalah seleksi fitur *chi square*. Contoh tabel kontingensi yang digunakan pada proses seleksi fitur *chi square*, dengan ukuran b x k ditunjukkan oleh Tabel 2 :

Tabel 2. Tabel kontingensi

	$Kolom_1$	$Kolom_2$	...	$Kolom_j$	Jumlah
$Baris_1$	O_{11}	O_{12}	...	O_{1j}	b_1
$Baris_2$	O_{21}	O_{22}	...	O_{2j}	b_2
...	...	...	...	...	...
$Baris_i$	O_{i1}	O_{i2}	...	O_{ij}	b_i
Jumlah	k_1	k_2	...	k_j	N

O_{ij} merupakan *actual count* atau nilai sebenarnya dari dua variabel yang diamati, pada kasus ini yaitu fitur sebagai baris, dan kelas (negatif, netral, positif) sebagai kolom. Proses selanjutnya dalam uji *chi square* adalah mencari *expected count*. Tabel *expected count* ditunjukkan oleh tabel berikut :

Tabel 3. Tabel kontingensi *expected count*

	$Kolom_1$	$Kolom_2$	...	$Kolom_j$
$Baris_1$	E_{11}	E_{12}	...	E_{1j}
$Baris_2$	E_{21}	E_{22}	...	E_{2j}
...	...	...	...	...
$Baris_i$	E_{i1}	E_{i2}	...	E_{ij}

Mencari *expected count* dari O_{ij} , atau mengisi sel E_{ij} , dilakukan dengan perhitungan berikut [19]:

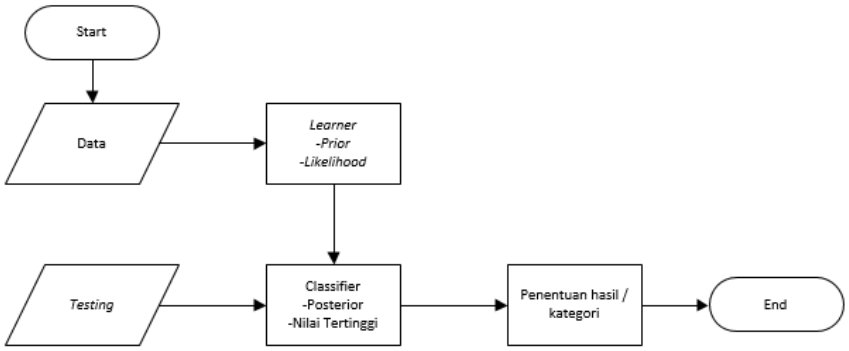
$$E_{ij} = \frac{b_i k_j}{N} \quad (6)$$

Berdasarkan Tabel 2 dan Tabel 3 :
O_{ij} = *actual count* dari baris i dan kolom j
E_{ij} = *expected count* dari baris i dan kolom j
b_i = penjumlahan baris ke i
k_j = penjumlahan kolom ke j
N = total seluruh nilai

Penggunaan *alpha* memberikan pengaruh terhadap proses seleksi fitur chi square, karena semakin kecil nilai *level of significance* maka semakin sedikit fitur yang lolos seleksi. Hal ini dikarenakan semakin kecil *nilai level of significance*, semakin tinggi *critic score* yang menyebabkan proses seleksi menjadi lebih ketat [20]. Pada penelitian ini *alpha* yang digunakan sebesar 0,01. Untuk contoh penerapan proses seleksi fitur dengan *chi square* pada penelitian ini ada pada Lampiran 2.

iv. Naive bayes classifier

Setelah melakukan seleksi fitur *chi square*, selanjutnya adalah proses klasifikasi menggunakan naive bayes. Hasil prediksi terhadap data uji dimunculkan dalam bentuk *confussion matrix* yang mendukung perhitungan performansi terhadap metode yang digunakan. Berikut ilustrasi proses klasifikasi dengan metode naive bayes :



Gambar 4. Cara kerja naive bayes

Contoh cara kerja penggunaan naive bayes terdapat pada Lampiran 3.

v. Hasil

Hasil klasifikasi ditampilkan dalam bentuk *confussion matrix* yang selanjutnya dilakukan perhitungan untuk mengetahui persentase akurasi, presisi, *recall* dan *f1-score* yang digunakan untuk melihat performansi metode yang digunakan. Adapun perhitungannya adalah sebagai berikut :

$$Akurasi = \frac{Jumlah\ data\ dengan\ prediksi\ benar}{Total\ semua\ data\ yang\ digunakan} \quad (7)$$

$$Presisi = \frac{Jumlah\ data\ dengan\ prediksi\ benar}{Jumlah\ data\ yang\ diprediksi} \quad (8)$$

$$Recall = \frac{Jumlah\ data\ dengan\ prediksi\ benar}{Jumlah\ data\ sebenarnya} \quad (9)$$

$$F1 - score = \frac{2\ \times presisi\ \times recall}{presisi\ + recall} \quad (10)$$

4. Evaluasi

4.1 Hasil Pengujian

Pengujian ini dilakukan dengan dua skenario, yaitu :

i. Klasifikasi menggunakan metode naive bayes tanpa menggunakan seleksi fitur *chi square* terhadap data teks berbahasa Indonesia tentang penyedia layanan telekomunikasi
ii. Klasifikasi menggunakan metode naive bayes dengan menggunakan seleksi fitur *chi square* (dengan *alpha* 0,05; 0,01)terhadap data teks berbahasa Indonesia tentang penyedia layanan telekomunikasi

Tabel 4. Hasil klasifikasi

Metode	**Rata-rata**			
	Akurasi	**Presisi**	**Recall**	**F1-score**
Naive Bayes	84,4%	80%	84%	82%
Naive Bayes dengan Seleksi fitur *chi square* (*aplha* = 0,05)	84,7%	80%	85%	82%
Naive Bayes dengan Seleksi fitur *chi square* (*aplha* = 0,01)	85,5%	83%	86%	84%

Hasil pengujian menggunakan seleksi fitur *chi square* dengan tingkat signifikansi atau *alpha* 0,01 memberikan akurasi yang lebih tinggi 1,1% dibandingkan dengan seleksi fitur tanpa menggunakan naive bayes

Tabel 5. Hasil klasifikasi naive bayes dengan seleksi fitur *chi square*

Prediksi	**Negatif**	**Netral**	**Positif**
Sebenarnya			
Negatif	268	7	1
Netral	19	0	2
Positif	20	3	40

$$Akurasi = \frac{268 + 40}{268 + 19 + 20 + 7 + 3 + 1 + 2 + 40} = 85{,}5\%$$

Presisi, recall dan f1-score untuk hasil klasifikasi naive bayes dengan seleksi fitur *chi square* ditunjukkan oleh Tabel 5 :

Tabel 6. Performansi naive bayes dengan seleksi fitur *chi-square*

	presisi	***recall***	***f-measure***
Negatif	87%	97%	92%
Netral	0%	0%	0%
Positif	93%	63%	75%
Rata-rata	83%	86%	84%

4.2 Analisis Hasil Pengujian

Berdasarkan skenario pengujian yang dilakukan, didapatkan performansi yang baik dengan rata-rata diatas 80%. Performansi tersebut dapat dipengaruhi oleh ketidak seimbangan data, dimana dari 1200 tweet yang digunakan, lebih dari 70% data memiliki kelas / kategori negatif. Terlihat dari banyaknya data netral dan positif yang salah dikategorikan menjadi negatif. Konsep multinomial naive bayes yang juga memperhatikan kemunculan kata pada suatu kelas, membuat data dengan kelas negatif memberikan performansi yang lebih tinggi jika dibandingkan dengan dua kelas lainnya yaitu kelas positif dan netral. Pelabelan tweet masih dilakukan secara manual, dan untuk mengurangi subjektivitas, pelabelan ini dilakukan bukan hanya oleh 1 orang, namun oleh 3 orang *native speaker* bahasa Indonesia dengan diambil modus (nilai terbanyak) untuk diterapkan pada proses klasifikasi. Perbedaan pendapat terhadap label pada beberapa tweet menyebabkan ketidakkonsistenan data yang juga dapat berakibat pada performansi sistem, karena kata yang terdapat pada tweet tersebut dapat mempengaruhi perhitungan untuk klasifikasi.

Chi square sebagai seleksi fitur pada penelitian ini mengimplementasikan nilai *alpha* 0,05 dan 0,01. Hasil yang diperoleh dengan *alpha* 0,01 memberikan akurasi lebih tinggi. Semakin kecil tingkat signifikansi (*alpha*) yang digunakan, semakin tinggi *critic score chi square*, sehingga proses seleksi fitur semakin ketat, dan yang lolos adalah yang dianggap memiliki ketergantungan kuat, pada penelitian ini, *alpha* yang lebih kecil, memberikan akurasi yang lebih baik.

Perbedaan performansi antar dua skenario uji menunjukkan hasil yang tidak signifikan, yang bisa disebabkan karena pemilihan fitur yang cenderung sedikit, dari total 2404 fitur, seleksi fitur *chi square* pada penelitian ini dapat mereduksi fitur menjadi 2399. Namun, pada data yang digunakan untuk klasifikasi pada penelitian sebelumnya [20], dengan fitur awal 3224 menjadi 1229, penggunaan atribut yang banyak juga tidak memberikan perubahan yang signifikan dengan perbedaan akurasi tidak lebih dari 1%. Pemilihan atribut juga menunjukkan beberapa fitur pada penyedia layanan telekomunikasi yang sering dikeluhkan oleh masyarakat Indonesia, seperti sinyal, jaringan dan pulsa.

5. Kesimpulan dan Saran

Penerapan seleksi fitur *chi square* dengan tingkat signifikansi 0,01 terhadap klasifikasi menggunakan metode naive bayes pada penelitian ini memberikan akurasi 85,5%, dan *f1-score* 84%, sedangkan tanpa menggunakan seleksi fitur *chi square* didapatkan akurasi 84,4%, dan *f1-score* 82%. Penggunaan seleksi fitur *chi square* tidak memberikan pengaruh yang signifikan terhadap klasifikasi menggunakan naive bayes, namun bisa mereduksi fitur yang dianggap kurang relevan untuk proses klasifikasi.

Saran yang dapat dipertimbangkan untuk penelitian selanjutnya adalah untuk menyeimbangkan jumlah kelas pada data yang digunakan dan menyeimbangkan distribusi atribut untuk proses seleksi fitur *chi square*. Hal ini disarankan untuk mengetahui bagaimana pengaruhnya terhadap hasil klasifikasi yang akan dilakukan.

Daftar Pustaka

[1] B. Liu, Sentiment Analysis and Opinion Mining, Johnson: Morgan & Claypool, 2012.

[2] Devika, Sunitha and A. Ganesh, "Sentiment Analysis : A Comparative Study on Different Approaches," in *Fourth International Conference on Recent Trends in Computer Science & Engineering*, Chennai, 2016.

[3] A. F. Hidayatullah and M. R. Ma'arif, "Penerapan Text Mining pada Klasifikasi Judul Skripsi," *Seminar Nasional Aplikasi Teknologi Informasi (SNATI),* pp. 33-36, 2016.

[4] D. Xhemali, C. J. Hinde and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *IJCSI International Journal,* vol. 04, no. Computer Science, pp. 16-24, 2009.

[5] M. Danubianu and S. Pentiuc, "Data Dimensionality Reduction for Data Mining : A Combined Filter-Wrapper Framework," *INT J Comput Commun,* vol. VII, pp. 824-831, 2012.

[6] V. Kumar and S. Minz, "Feature Selection : A Literature Review," *Samrt Computing Review,* pp. 211-229, 2014.

[7] F. Thabtah, "Naive Bayesian Based on Chi Square to Categorize Arabic Data," *Communications of the IBIMA* , vol. 10, no. 2009, pp. 158-163, 2009.

[8] I. Sofiana and dkk, "Analisis Pengaruh Feature Selection Menggunakan Information Gain dan Chi-Square untuk Kategorisasi Teks Berbahasa Indonesia," *Proceding of Engineering,* p. 33, 2012.

[9] C. F. Suharno, M. A. Fauzi and R. S. Perdana, "Klasifikasi Teks Bahasa Indonesia pada Dokumen Pengaduan Sambat Online Menggunakan K-Nearest Neighbors dan Chi-Square," *SYSTEMIC,* vol. 03, pp. 25-32, 2017.

[10] "Digital 2019 : Indonesia," 01 06 2019. [Online]. Available: https://datareportal.com/reports/digital-2019-indonesia.

[11] Finder, "Traveling to Indonesia? Check out our guide to finding the best prepaid SIM card," 17 August 2018. [Online]. Available: https://www.finder.com/best-prepaid-sim-card-indonesia.

[12] Y. Pramitarini, P. I. K. Eddy and M. H. Purnomo, "Analisa Rekam Medis untuk Menentukan Status Gizi Anak Balita Menggunakan Naive Bayes Classifier," in *Prosiding Seminar Nasional Manajemen Teknologi XVII,* Surabaya, 2013.

[13] G. Berliana, Shaufiah and S. Sa'adah, "Klasifikasi Posting Tweet mengenai Kebijakan Pemerintah menggunakan Naive Bayesian Classification," *e-Proceeding of Engineering,* vol. 5, pp. 1562-1569, 2018.

[14] D. H. Kalokasari, I. M. Shofi and A. H. Setyaningrum, "Implementasi Algoritma Multinomial Naive Bayes Classifier pada Sistem Klasifikasi Surat Keluar," *Jurnal Teknik Informatika,* vol. 10, pp. 109-118, 2017.

[15] C. F. Suharno, M. A. Fauzi and R. S. Perdana, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan K-Nearest Neighbors dan Shi-Square," *Systemic,* pp. 25-32, 2017.

[16] S. Kusumadewi, "Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification," *CommIT, Vol. 3 No. 1*, pp. 6-11, 2009.

[17] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research 3*, pp. 1157-1182, 2003.

[18] I. C. Negara and A. Prabowo, "Penggunaan Uji Chi-Square untuk Mengeahui Pengaruh Tingkat Pendidikan dan Umur terhadap Pengetahuan Penasun Mengenai HIV-AIDS di Provinsi DKI Jakarta," in *Prosiding Seminar Nasional Matematika dan Terapannya 2018*, Purwokerto, 2018.

[19] W. Pramesti, "Tabel Kontingensi untuk Mengetahui Hubungan Antara Jenis Penyakit, Jenis Kelamin, Usia, Lama Rawat dan Keadaan keluar Pasien," *J-Statistika*, vol. 4, pp. 15-26, 2012.

[20] R. A. Aziz, M. S. Mubarok and Adiwijaya, "Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes," *Ind. Symposium on Computing*, pp. 139-148, 2016.