Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 1, Januari 2019, hlm. 447-453 http://j-ptiik.ub.ac.id

Penentuan *Rating Review* Film Menggunakan Metode *Multinomial Naïve Bayes Classifier* dengan *Feature Selection* berbasis *Chi-Square* dan *Galavotti-Sebastiani-Simi Coefficient*

Thio Marta Elisa Yuridis Butar Butar[1], Mochammad Ali Fauzi[2], Indriati[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]thaz070196@student.ub.ac.id, [2]moch.ali.fauzi@ub.ac.id, [3]indriati.tif@ub.ac.id

Abstrak

Pada era saat ini terdapat berbagai ragam film, meskipun cara pendekatannya berbeda-beda, semua film dapat dikatakan mempunyai satu sasaran, yaitu menarik perhatian orang terhadap muatan-muatan masalah yang dikandung. Dari muatan-muatan film tersebut terdapat banyak respons dari penulis dan menuliskannya dalam *review* singkatnya. Dengan adanya *review* tersebut bisa membantu penonton untuk lebih selektif lagi dalam memilih suatu film. Dan dari pihak produksi bisa terbantu untuk mengukur seberapa jauh kualitas film yang mereka hasilkan. Namun dari pihak produksi sendiri terkadang mengalami kesulitan dalam memilah dan mengkategorikan *review*, apakah produk tersebut kualitasnya tergolong bagus, cukup bagus, tidak bagus, dan sebagainya. Dalam penelitian ini peniliaian suatu film berdasarkan *review* yang diberikan adalah *Rating*. Sehingga dibutuhkan sebuah sistem prediksi *Rating* untuk memprediksi dan menentukan *Rating* yang tepat berdasarkan *review* yang diberikan oleh pengguna terhadap suatu film. Untuk mendukung sistem yang dibangun dibutuhkan metode untuk menyelesaikan permasalahan tersebut, dalam penelitian ini peneliti menggunakan metode *Multinomial Naïve Bayes* serta *Chi-Square* dan *Galavotti-Sebastiani-Simi Coefficient*. *Multinomial Naïve Bayes* adalah metode untuk klasifikasi sedangkan *Chi-Square* dan *Galavotti-Sebastiani-Simi Coefficient* adalah *feature selection* untuk lebih mengoptimalkan hasil dari klasifikasi. Dari hasil pengujian, didapat tingkat akurasi terbaik pada saat penggunaan *feature* sebesar 90%, dan 100% dengan tingkat akurasi sebesar 36%. Hasil tersebut adalah hasil terbaik dari hasil dengan prosentase penggunaan *feature* yang lain. Dari hasil tersebut *CHI-GSS* terbukti bisa melakukan pemilihan kata yang dianggap relevan maupun tidak relevan untuk dilakukan klasifikasi.

Kata kunci: prediksi *rating*, *review* film, *Multinomial Naïve Bayes*, *Chi-Square*, *Galavotti-Sebastiani-Simi Coefficient*

Abstract

In the current era there are various kinds of movies, although the way of approach varies, all movies can be said to have one goal, namely to attract people's attention to the contents of the problem. From the contents of the movie there are many responses from the author and write them in a short review. With the review can help consumers to be more selective again in choosing a movie. And from the production side can be helped to measure how far the quality of the movies they produce. But from the production itself sometimes have difficulty in sorting and categorize the review, whether the movie is good quality, good enough, not good, and so forth. In this study the assessment of a moview based on the review given is Rating. So it takes a Rating prediction system to predict and determine the right Rating based on the reviews given by the users of a movie. To support the system built required methods to solve the problem, in this study researchers used the method of Multinomial Naïve Bayes along Chi-Square and Galavotti-Sebastiani-Simi Coefficient. Multinomial Naïve Bayes is a method for classification whereas Chi-Square and Galavotti-Sebastiani-Simi Coefficient is a feture selection to futher optimize the results of classification. From the test results, obtained the best accuracy level when the use features by 90%, and 100% with an accuracy of 36%. These results are the best results of the results with other features usage percentages. From these results CHI-GSS proven to make the selection of words that are considered relevant or irrelevant to do classification.

***Keywords*:** *rating prediction, movie review, Multinomial Naïve Bayes, Chi-Square, Galavotti-Sebastiani-Simi Coefficient*

1. PENDAHULUAN

Menurut Effendy (1986: 134), film adalah media komunikasi yang bersifat audio visual untuk menyampaikan suatu pesan kepada sekelompok orang yang berkumpul di suatu tempat tertentu. Pesan film pada komunikasi massa dapat berbentuk apa saja tergantung dari misi film tersebut. Akan tetapi, umumnya sebuah film dapat mencakup berbagai pesan, baik itu pesan pendidikan, hiburan dan informasi. Pesan dalam film adalah menggunakan mekanisme lambang – lambang yang ada pada pikiran manusia berupa isi pesan, suara, perkataan, percakapan dan sebagainya.

Film juga dianggap sebagai media komunikasi yang ampuh terhadap massa yang menjadi sasarannya, karena sifatnya yang audio visual, yaitu gambar dan suara yang hidup. Dengan gambar dan suara, film mampu bercerita banyak dalam waktu singkat. Ketika menulis film penulis seakan-akan dapat menembus ruang dan waktu yang dapat menceritakan kehidupan dan bahkan dapat mempengaruhi audiens.

Oleh begitu pesatnya perkembangan teknologi dewasa ini banyak juga suatu situs konten film yang berisikan tentang bagaimana film tersebut dan *review* penulis yang pernah menyaksikan film tersebut. Sehingga dari *review* tersebut bisa merekomendasi pecinta film untuk menyaksikan film tertentu. Salah satu contoh situs web yang dimaksud ialah seperti pada https://montasefilm.com dan http://www.ulasanpilem.com.

Pada era yang ini terdapat berbagai ragam film, meskipun cara pendekatannya berbeda-beda, semua film dapat dikatakan mempunyai satu sasaran, yaitu menarik perhatian orang terhadap muatan-muatan masalah yang dikandung. Dari muatan-muatan film tersebut terdapat banyak respons dari penulis dan menuliskannya dalam *review* singkatnya. Dari *review* singkat tersebut dapat dikategorikan film tersebut termasuk dalam kategori *Rating* 1, 2, 3, 4, atau 5. Namun dari beberapa *review* singkat dari penulis tersebut, ada *review* yang tidak begitu jelas termasuk kategori yang mana. Jadi, kami akan membuat sebuah aplikasi yang menerapkan *Text Mining* menggunakan metode *Naïve Bayes* yang dapat menentukan *review* singkat termasuk kategori *Rating* 1, 2, 3, 4, atau 5.

Pada penelitian sebelumnya yang berkaitan dengan *Chi-Square* (*CHI*) yang berjudul ***Study of Feature Selection Algorithms for Text-Categorization*** oleh Kandarp Dave, 2011 menyatakan bahwa pengkategorisasian teks sangat penting, tetapi permasalahan *feature selection* sama banyak atau lebih penting daripada kategori teks. Dalam penelitian tersebut dibahas banyak topik penting mulai dari pengumpulan data, hingga pemrosesan data dan akhirnya menggunakan data yang telah diproses tersebut untuk dilakukan tes secara efisien menggunakan algoritme *feature selection*. Penelitian tersebut menunjukkan beberapa peningkatan dramatis menggunakan hasil tersebut. Dan metode pemilihan *feature* harus diteliti lebih lanjut, pada data skala sangat besar. Jumlah pelatihan dan dokumen uji yang digunakan dalam penelitian tersebut sangat kecil dibandingkan apa yang tersebar luas di dunia maya. Selain itu, dalam penelitian tersebut, hanya ada 9 kategori. Di dunia nyata, ada ratusan kategori. Untuk memiliki pengkategorisasi berskala beasar. Algoritme *feature selection* harus secara kuat dikembangkan. Dan topik ini dapat diteliti dan diuji lebih lanjut. Dengan melakukan penelitian lebih lanjut tentang topik yang disebutkan tersebut akan membantu, karena pada akhirnya dapat membantu mengkategorikan semua dokumen di dunia.

Pada penelitian sebelumnya yang berkaitan dengan *feature selection* serta mencakup *Galavotti-Sebastiani-Simi Coefficient* (*GSS*) di dalamnya yang berjudul ***Feature Selection for Text Categorisation*** oleh Øystein Løhre Garnes, 2009 membahas tentang langkah-langkah untuk membangun pengklasifikasian menggunakan kumpulan file vektor (*feature* yang dipilih), mengevaluasi kinerjanya, dan menyimpan hasilnya di file. File berisi satu baris untuk setiap run, fold, dan dataset. Oleh karena itu dalam kasus penelitian tersebut, kalau hasil file dari percobaan *Naïve Bayes* berisi 600 baris hasil: 6 set ukuran *feature* (dari 500 hingga 10.000 *feature*) kali 10 kali fold 10 kali run (setiap kali lipat dijalankan), sedangkan hasil file dari percobaan Support Vector Machine berisi 150 baris hasil: 3 set ukuran *feature* (500, 1000, dan 2000 *feature*) kali 10 kali fold 5 kali run. Maka, dapat disimpulkan bahwa *Naïve Bayes* lebih cepat dibandingkan Support Vector Machine.

Pada penelitian sebelumnya yang berkaitan

dengan *feature selection* yang berjudul ***Categorical Proportional Difference***: ***A Feature Selection Method for Text Categorization*** oleh Simeon, 2008 menyatakan bahwa sebuah *feature* dapat meningkatkan akurasi proses perhitungan. Metode *feature selection* adalah metode yang sangat populer dalam berbagai penelitian. *feature selection* digunakan untuk mengurangi dimensi dan mempercepat proses perhitungan. Selain itu *feature selection* juga mampu meningkatkan efisiensi dan akurasi dalam proses *document extraction* yang subset dengan pemilihan *feature* yang dianggap lebih relevan.

2. DATA DAN METODE

2.1 Data

Data yang digunakan pada penelitian ini adalah review dari berbagai film yang langsung diambil dari situs web https://montasefilm.com dan http://www.ulasanpilem.com data yang digunakan sebagai data latih sebanyak 250 data dengan komposisi *Rating* 1 sebanyak 50 data, *Rating* 2 sebanyak 50 data, *Rating* 3 sebanyak 50 data, *Rating* 4 sebanyak 50 data, *Rating* 5 sebanyak 50 data. Sedangkan data uji yang digunakan sebanyak 50 data.

2.2 *Naïve Bayes Classifier* (*NBC*)

Naïve Bayes Classifier merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan *Teorema Bayes* dengan asumsi ketidaktergantungan (independent) yang tinggi. *Teorema Bayes* adalah teorema yang dipakain dalam statistika untuk menghitung peluang untuk suatu hipotesis. *Bayes Optimal Classier* menghitung peluang dari suatu kelas dari masing-masing kelompok atribut yang ada, menentukan kelas mana yang paling optimal.

2.2.1 *Multinomial Naïve Bayes*

Multinomial Naïve Bayes digunakan untuk mengasumsikan independensi kemunculan kata dalam dokumen. Metode ini tidak memperhitungkan urutan kata dan information-context dalam dokumen, namun memperhitungkan jumlah kata dalam dokumen (Destuardi dan Sumpeno, 2009).

Dengan persamaan:

$$P(F, C_k) = \frac{count\ (F,C_k)+1}{(\sum_{F \in V} count(F,C_k))+|V|} \quad (1)$$

$count\ (F, C_k)$ merupakan jumlah feature F yang muncul dalam suatu kategori C_k, penambahan nilai 1 untuk menghindari nilai *zero*, $\sum_{F \in V} count(F, C_k)$ merupakan jumlah seluruh feature F pada kategori C_k, dan $|V|$ merupakan jumlah seluruh feature unik di seluruh kategori.

2.3 *Chi-Square* (*CHI*)

Chi-Square adalah metode statistik yang pada awalnya digunakan dalam analisis statistik untuk mengukur bagaimana hasil pengamatan berbeda (yaitu independen) dari hasil yang diharapkan sesuai dengan hipotesis awal (nilai yang lebih tinggi menunjukkan kemandirian yang lebih tinggi). Dalam konteks statistik klasifikasi teks digunakan untuk mengukur seberapa independen sebuah kata dan sebuah kelas (Gulden Uchyigit, 2012).

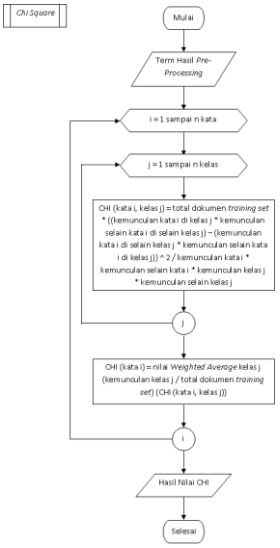
Dengan persamaan:

$$x^2(F, C_k) = \frac{N \times ((N_{F,C_k} \times N_{\bar{F},\overline{C_k}}) - (N_{F,\overline{C_k}} \times N_{\bar{F},C_k}))^2}{N_F \times N_{\bar{F}} \times N_{C_k} \times N_{\overline{C_k}}} \quad (2)$$

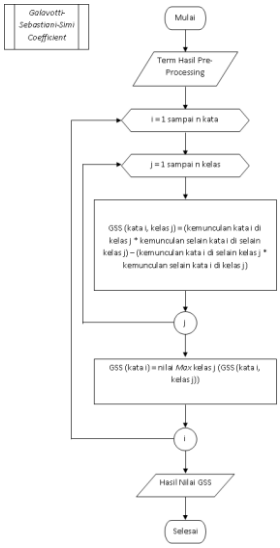
Dan persamaan untuk menghitung *weighted average* adalah sebagai berikut:

$$x^2(F) = \sum_{k=1}^{|C|} \frac{N_{C_k}}{N} x^2(F, C_k) \quad (3)$$

N merupakan jumlah total dokumen dalam training set, N_{C_k} merupakan jumlah dokumen dalam kategori C_k, $N_{\overline{C_k}}$ merupakan jumlah dokumen yang tidak ada dalam kategori C_k, N_F merupakan jumlah dokumen yang berisi feature F, $N_{\bar{F}}$ merupakan jumlah dokumen yang tidak berisi feature F, N_{F,C_k} merupakan jumlah dokumen yang berisi feature F dalam kategori C_k, $N_{\bar{F},C_k}$ merupakan jumlah dokumen yang tidak berisi feature F dalam kategori C_k, $N_{F,\overline{C_k}}$ merupakan jumlah dokumen yang berisi feature F tidak ada dalam kategori C_k, dan $N_{\bar{F},\overline{C_k}}$ merupakan jumlah dokumen yang tidak berisi feature F tidak ada dalam kategori C_k.



Gambar 1. Alur Proses Penyelesaian CHI



Gambar 2. Alur Proses Penyelesaian GSS

2.4 *Galavotti-Sebastiani-Simi Coefficient* (*GSS*)

GSS Coefficient adalah metode yang diusulkan oleh Galavotti dkk. Yang merupakan sebuah statistik *Chi-Square* yang disederhanakan. Mereka menghapus faktor pada pembilang karena sama untuk semua pasang dan karena itu menjadi berlebihan. Mereka kemudian menghapus ini memiliki nilai rendah untuk kata-kata langka yang berarti bahwa kata-kata tersebut diberi skor tinggi (karena ini adalah bagian dari penyebut). Tapi kata-kata langka yang ditunjukkan oleh adalah yang paling tidak efektif dalam klasifikasi teks. Akhirnya, mereka menghapus faktor dari penyebut karena akan menekankan kategori yang sangat langka (yaitu kategori dengan contoh yang sangat sedikit) (Gulden Uchyigit, 2012).

Dengan persamaan:

$$GSS(F, C_k) = N_{F,C_k} N_{\overline{F},\overline{C_k}} - N_{F,\overline{C_k}} N_{\overline{F},C_k} \quad (4)$$

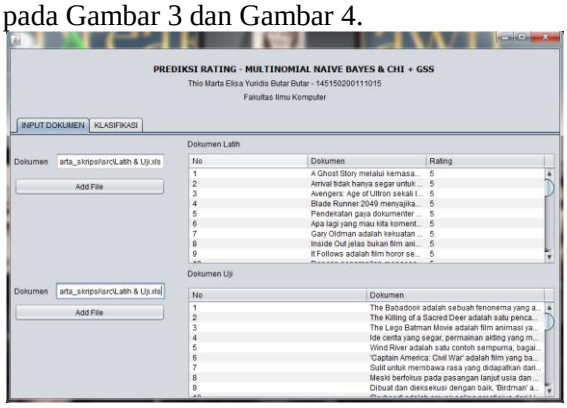
Dan persamaan untuk menghitung nilai max adalah sebagai berikut:

$$GSS(F, C_k) = \max_{k=1}^{|C|} GSS(F, C_k) \quad (5)$$

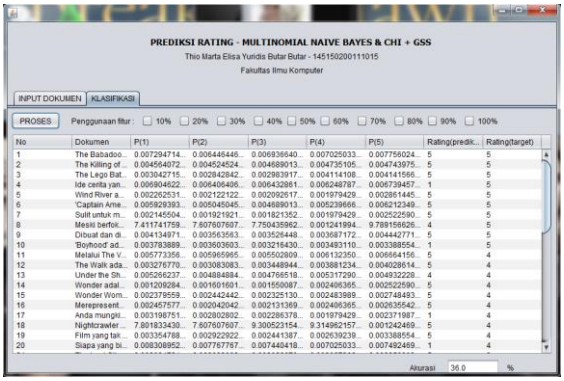
N merupakan jumlah total dokumen dalam training set, N_{F,C_k} merupakan jumlah dokumen yang berisi feature F dalam kategori C_k, $N_{\overline{F},C_k}$ merupakan jumlah dokumen yang tidak berisi feature F dalam kategori C_k, $N_{F,\overline{C_k}}$ merupakan jumlah dokumen yang berisi feature F tidak ada dalam kategori C_k, $N_{\overline{F},\overline{C_k}}$ merupakan jumlah dokumen yang tidak berisi feature F tidak ada dalam kategori C_k.

3. IMPLEMENTASI

Lingkungan implementasi diantaranya menggunakan perangkat keras dan perangkat lunak. Perangkat keras yang digunakan memiliki spesifikasi meliputi: Memory (RAM) 8 GB, Processor Intel Core i3 2.40 GHz, Harddisk 1.5 TB, Power Supply 600 W sedangkan perangkat lunak yang digunakan meliputi: Sistem operasi Microsoft Windows 7 64-bit, editor pemrograman Netbeans IDE, editor dokumentasi Microsoft Office 2007, dan Microsoft Excel 2007. Implementasi antarmuka pada sistem ini berbasis desktop dan menggunakan bahasa pemrogramana Java. Berikut halaman antarmuka sistem ditunjukkan pada Gambar 3 dan Gambar 4.



Gambar 3. Halaman Input Data



Gambar 4. Halaman Pengujian

4. HASIL PENGUJIAN DAN ANALISIS

Pada tahap pengujian dilakukan dua skenario pengujian, yang pertama adalah pengujian dengan menggunakan metode *Multinomial Naïve Bayes* saja tanpa menggunakan seleksi feature. Yang kedua adalah pengujian dengan menggunakan *Multinomial Naïve Bayes*-CHI-GSS artinya pada pengujian ini akan dilakukan pengurangan feature berdasarkan nilai CHI-GSS pada masing-masing term.

4.1 Skenario Pengujian Klasifikasi Klasifikasi *Naïve Bayes* Tanpa *Feature Selection*

Pengujian ini dilakukan untuk mengetahui tingkat akurasi pada klasifikasi prediksi *Rating* dengan menggunakan metode *Multinomial Naïve Bayes* dan tanpa menggunakan feature seleksi. Kumpulan *term* yang dihasilkan dari proses pre-processing akan langsung dilakukan klasifikasi tanpa harus dikurangi.

Pada pengujian ini, data yang diuji benar-benar data asli *review* film yang diambil pada situs web https://montasefilm.com dan http://www.ulasanpilem.com. data uji yang dipakai pada pengujian ini sebanyak 50 data dengan komposisi data random untuk *Rating* 1 sampai *Rating* 5, sedangkan untuk data latih yang digunakan sebanyak 250 data dengan komposisi data 50 data *Rating* 1, 50 data *Rating* 2, 50 data 3, 50 data 4, 50 data 5. Hasil pengujian untuk tingkat akurasi menggunakan *Multinomial Naïve Bayes* adalah sebesar 36%.

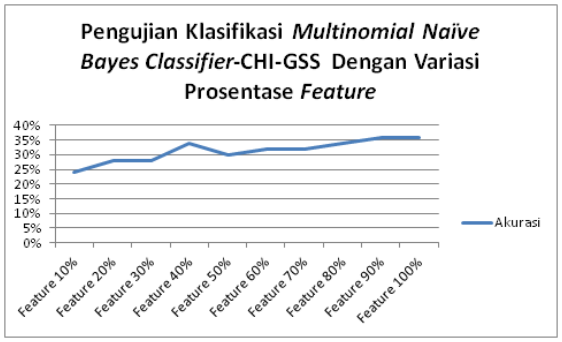
4.2 Skenario Pengujian Klasifikasi Klasifikasi *Naïve Bayes* Tanpa Variasi Prosentase Feature

Pengujian ini menjelaskan tentang pengujian klasifikasi dengan menggunakan metode *Multinomial Naïve Bayes*-CHI-GSS dengan variasi prosentase feature yang digunakan pada saat klasifikasi. Pada pengujian ini akan dilakukan pengurangan dimensi feature atau *term* hasil pre-processing yang digunakan pada saat klasifikasi adalah sebanyak prosentase yang yang telah ditetapkan yaitu sebesar 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, dan 100%. Untuk hasil pengujian ini dapat dilihat pada Tabel 2.

Tabel 2. Pengujian Klasifikasi *Naïve Bayes* Dengan Variasi Prosentase Feature

Prosentase	Akurasi
10%	24%
20%	28%
30%	28%
40%	34%
50%	30%
60%	32%
70%	32%
80%	34%
90%	36%
100%	36%

Pada pengujian ini sudah dilakukan pengurangan dimensi feature sesuai prosentase yang sudah ditetapkan oleh peneliti. Dengan data yang sama pengujian ini mendapatkan nilai akurasi terbaik sebanyak 36% pada penggunaan feature sebanyak 90% dan 100%. Dengan kata lain penggunaan CHI-GSS di sini dapat mempengaruhi nilai akurasi dengan mengurangi atau membuang *term-term* yang dianggap tidak relevan untuk diklasifikasi.



Gambar 5. Grafik Akurasi *Multinomial Naïve Bayes*-CHI-GSS Dengan Variasi Prosentase Feature

5. KESIMPULAN

Berdasarkan hasil pengujian dan analisis dari prediksi *Rating* pada *review* film menggunakan Metode *Multinomial Naïve Bayes Classifier* dengan *feature selection* berbasis *Chi-Square* dan *Galavotti-Sebastiani-Simi Coefficient* dapat disimpulkan sebagai berikut:

1. Metode *Multinomial Naïve Bayes Classifier* dengan *feature selection* berbasis *Chi-*

Square dan *Galavotti-Sebastiani-Simi Coefficient* dapat diterapkan pada proses prediksi *Rating* pada *review* film. Data latih yang diambil bersumber pada *review* film pada situs web https://montasefilm.com dan http://www.ulasanpilem.com dengan menggunakan data latih sebanyak 250 data dan data uji sebanyak 50 data dapat memberikan hasil yang lebih baik ketika penggunaan *feature* sebesar 90%, dan 100%. Sebelum dilakukan proses klasifikasi dokumen perlu melalui berbagai tahapan meliputi *tokenizing*, *filtering*, *case folding*, dan *stemming* untuk lebih memaksimalkan hasil klasifikasi.

2. Klasifikasi menggunakan *Naïve Bayes Classifier* dengan *feature selection* berbasis *Chi-Square* dan *Galavotti-Sebastiani-Simi Coefficient* dapat memberikan hasil yang lebih baik daripada menggunakan *Naïve Bayes Classifier* biasa. Hasil terbaik yang didapat pada saat penggunaan *feature* sebanyak 90%, dan 100%. dengan tingkat akurasi sebesar 36%. Hal tersebut membuktikan *CHI-GSS* berhasil melakukan pemilihan kata yang lebih diprioritaskan untuk dilakukan klasifikasi dan membuang kata-kata yang dianggap tidak relevan untuk dilakukan klasifikasi.
3. Pengurangan dimensi *feature* dengan menggunakan *Chi-Square* dan *Galavotti-Sebastiani-Simi Coefficient* yang diterapkan dengan menggunakan data dari situs web https://montasefilm.com dan http://www.ulasanpilem.com tidak menjamin ketika semakin kita memperkecil dimensi *feature* yang digunakan maka akan memberikan tingkat akurasi yang lebih baik. Karena pada pengujian yang dilakukan pada saat penggunaan *feature* sebesar 10% sistem memberikan nilai akurasi yang paling rendah yaitu hanya sebesar 24%, artinya adalah *term-term* yang diproses pada saat klasifikasi tidak memberikan hasil yang maksimal karena bisa saja terdapat kata-kata yang sebelumnya relevan untuk dilakukan klasifikasi tetapi ikut terbuang sehingga sistem kekurangan informasi untuk memberikan hasil yang maksimal.

6. DAFTAR PUSTAKA

Adel, A., Omar, N., & Al-Shabi, A. (2014). *A Comparative Study of Combined Feature selection Methods For Arabic Text Classification*. *Journal of Computer Science*, *10*(*11*), *2232-2239*.

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M. M., Williams, H. E. (2007). *Stemming* Indonesian: A Confix-Stripping Approach. ACM J. Educ. Resourch. Computer. 6, 4, Article 13, 33 pages.

Agusta, Ledy. (2009). Perbandingan Algoritma *Stemming* Porter Dengan Algoritma Nazief & Adriani Untuk *Stemming* Dokumen Teks Bahasa Indonesia. Konferensi Nasional Sistem dan Informatika 2009: Bali, November 14, 2009.

Andilala. (2016). Movie *Review* Sentimen Analisis Dengan Metode Naïve Bayes Base On *Feature selection*. *Journal Pseudocode*, III(1).

Bhoir, P. & Kolte, S. (2015). *Sentiment Analysis of Movie Reviews Using Lexicon Approach*. *IEEE International Conference on Computational Intelligence and Computing Research*.

Dave, Kandarp. (2011). *Study of Feature selection Algorithms for Text-Categorization*. *Las Vegas*: *University of Nevada*.

Destuardi & Surya, S. (2009). Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode *Naïve Bayes*. Teknik Elektro, Institut Teknologi Sepuluh Nopember, Surabaya.

Effendy, Onong Uchjana. (1986). Dimensi-Dimensi Komunikasi. Bandung: Alumni.

Feldman, R. & Sanger, J. (2007). *The Text Mining Handbook*: *Advanced Approaches In Analyzing Unstructured Data*. *Cambridge University Press*

Forman, G. (2008). *Feature selection for text classification, in H. Liu and H. Motoda, eds, 'Computational Methods of Feature selection', Chapman and Hall/CRC, pp. 257–276*.

Garnes, Øystein Løhre. (2009). *Feature selection for Text Categorization*. *Norwegian University of Science and Technology*.

Guo, Q. (2010). *An Effective Algorithm for Improving the Performance of Naïve Bayes for Text Classification*. *Cambridge University Press*.

Jong, J. (2011). *Predicting Rating with*

Sentiment Analysis.

Kurniawan, B., Fauzi, M. A., & Widodo, A. W. (2017). Klasifikasi Berita Twitter Menggunakan Metode *Improved Naïve Bayes*

Medhat, W., Hassan, A., & Korashy, H. (2014). *Sentiment Analysis Algorithms and Applications*: *A Survey*. *Ain Shams Engineering Journal*, *5*(*4*), *1093-1113*.

Mustafa, A., Akbar, A., & Sultan, A. (2009). *Knowledge Discovery Using Text mining*: *A Programmable Implementation on Information Extraction and Categorization*. *International Journal of Multimedia and Ubiquitous Engineering*, *4*(*2*), *183-188*.

Rosi, F., Fauzi, M. A., & Perdana, R. S. (2018). Prediksi *Rating* Pada *Review* Produk Kecantikan Menggunakan Metode *Naïve Bayes* dan *Categorical Proportional Difference* (CPD)

Sahu, T. P. & Ahuja, S. (2016). *Sentiment Analysis of Movie Reviews*: *A study on Feature selection & Classification Algorithms*. *IEEE*.

Uchyigit, Gulden. (2012). *Experimental Evaluation of Feature selection Methods for Text Classification*. *IEEE 9th International Conference on Fuzzy Systems and Knowledge Discovery*.

Wijaya, M. C., Tjiharjadi, S. (2010). Aplikasi Klasifikasi Dokumen Menggunakan Metoda *Naïve Baysian*.

Zheng, Z., Srihari, R., Srihari, S. (2003). *A Feature selection Framework for Text Filtering*. *Proceedings of the Third IEEE International Conference on Data Mining*.