

Ekstraksi Ulasan Sentimen Film dari Twitter dengan Naïve Bayes pada Situs Web Media Sosial Penggemar Film

Adri Gabriel Sooi, *Fakultas Teknik, Prodi Ilmu Komputer Universitas Katolik Widya Mandira Kupang*,
Melania Laniwati, *Departemen Teknik Informatika Institut Sains dan Teknologi Terpadu Surabaya*.

Abstrak— Film dianggap sebagai bentuk seni serta merupakan sumber hiburan yang populer. Pembuatan penelitian ini diharapkan bisa membantu orang Indonesia untuk mendapatkan informasi tentang film serta membaca review dari film. Review film yang ada pada website ini didapatkan dari user-user lokal maupun dari Twitter. Sistem mengekstraksi dan mengkategorikan isi sentiment dari sebuah barisan teks tweet dengan menggunakan metodologi Basic Unified Process. Proses klasifikasi sentiment yang ada bertujuan untuk mengklasifikasi review sebagai positif/negatif. Seluruh tweet akan diproses melalui Feature Reduction dan Normalisasi. Proses Feature Reduction akan menghapus hashtag, username, link, dan tanda baca pada tweets. Pada proses Normalisasi, seluruh singkatan dan kata bukan baku pada tweets akan diganti. Penelitian ini menggunakan sistem Rule-Based dalam menentukan apakah tweet tersebut merupakan review film atau bukan. Penulis menggunakan algoritma Naïve Bayes untuk mengklasifikasi sentiment (positif/negatif) dari review. Penulis telah melakukan 8 buah pengujian, masing-masing 4 kali untuk pengujian sistem Rule-Based dan Naïve Bayes Classifier. Total data tweet yang diujicobakan adalah sebanyak 6.323, dan hasil akhir paling optimal yang didapatkan oleh sistem terhadap Rule-Based System menghasilkan akurasi sebesar 82,64% dan terhadap Naïve Bayes Classifier sebesar 74,09%. Dari hasil pengujian paling optimal ini, sistem mendapatkan nilai recall dan presisi masing-masing sebesar 71,44% dan 77,92% untuk Rule-Based System, serta 83,77% dan 77,65% untuk Naïve Bayes.

Kata Kunci—Naïve Bayes Classifier, Natural Language Processing, Rule-Based System, Sentiment Analysis.

I. PENDAHULUAN

Saat ini dunia teknologi semakin berkembang, terutama sejak meluasnya perkembangan internet. Hal yang mendorong perkembangan tersebut sebenarnya adalah peningkatan kebutuhan manusia akan berbagai bentuk dan jenis informasi, yang akan digunakan untuk berbagai macam keperluan manusia seperti halnya untuk mendapatkan informasi. Pada kota-kota besar di Indonesia di mana tantangan

semakin besar dan tingkat stress semakin tinggi, manusia semakin membutuhkan sarana hiburan yang tepat untuk meredakan tingkat stress. Salah satu sarana hiburan yang sehat dan populer saat ini adalah menonton film. Menonton film kini semakin populer di masyarakat Indonesia kini, didukung pula dengan fakta bahwa perfilman lokal semakin lebih baik.

Hingga saat ini, masyarakat di Indonesia tidak sepenuhnya bisa menentukan apakah suatu film layak ditonton di teater atau tidak, mengingat bahwa Cinema 21 yang merupakan penguasa pangan teater di Indonesia, tidak menawarkan sistem rating dan review sebuah film. Karena kelayakan suatu film untuk ditonton di teater tidak bisa ditentukan hanya berdasarkan sinopsis saja, masyarakat sering kali tertipu dan akhirnya merasa kecewa setelah menonton suatu film di teater [1]. Hal ini didukung pula dengan kenyataan bahwa banyak film lama yang baru saja ditampilkan di teater Indonesia. Masyarakat memang bisa saja melihat rating dari film secara manual melalui website atau aplikasi IMDB ataupun Metacritic, tetapi mengingat bahwa selera bangsa tidak selalu sepadan dengan bangsa lain, banyak film yang memiliki rating tinggi tetapi tidak diminati oleh masyarakat.

Sejauh ini, masih belum banyak terdapat aplikasi ataupun website lokal yang memperbolehkan pengguna untuk berbagi rating dan review mereka terhadap suatu film. Para penggemar film masih belum memiliki wadah untuk menyalurkan pengalaman mereka selama menonton suatu film. Tujuan utama dari dibuatnya penelitian ini adalah untuk mengatasi masalah di atas. Pada website yang akan dibuat ini, pengguna bisa membagikan pengalaman dan kesan pengguna selama menonton film dengan meninggalkan rating dan review terhadap suatu film. Pengguna bisa pula melihat film yang sedang tayang dan akan tayang, beserta dengan rating dari film yang sedang tayang, baik dari rating lokal (rating dari sesama pengguna) maupun rating yang diambil dari IMDB dan Metacritic, serta ekstraksi banyaknya respons negatif/positif dari komentar film yang ada di Twitter.

II. TINJAUAN PUSTAKA

Pada berikut ini akan menjelaskan mengenai landasan-landasan teori apa saja yang dijadikan sumber dalam pembuatan penelitian ini. Ada beberapa teori penunjang yang akan digunakan untuk membuat sekaligus mengembangkan aplikasi yang dibuat. Pada beberapa bagian akan disertakan

Adri Gabriel Sooi, Fakultas Teknik, Prodi Ilmu Komputer. Universitas Katolik Widya Mandira, Kupang, Indonesia (email: adrigabriel@unwira.ac.id)
Melania Laniwati, Departemen Teknik Informatika, Sekolah Tinggi Teknik Surabaya, Surabaya, Jawa Timur, Indonesia

beberapa contoh dari segmen program atau pun rumus dari teori-teori penunjang yang ada.

A. *Web Crawling*

Web Crawling adalah suatu proses di mana skrip yang dibuat bertujuan untuk melakukan scan ke halaman-halaman internet lainnya [2]. Kegunaan dari web crawling adalah untuk mendapatkan informasi yang diinginkan dari halaman-halaman internet yang dituju. Demi mencapai tujuan ini, maka dibutuhkan sebuah skrip di mana skrip tersebut dapat dengan sendirinya berjalan otomatis untuk mendapatkan informasi yang dibutuhkan tersebut [3]. Salah satu cara untuk melakukan web crawling adalah dengan melakukan parsing HTML, dimana skrip dari halaman web yang dituju akan diambil, dan dari sana akan diekstraksi informasi-informasi yang diinginkan dari halaman HTML tersebut.

Sebuah web crawler adalah bot Internet yang secara otomatis menelusuri World Wide Web, biasanya untuk tujuan pengindeksan web (web spidering) [4]. Beberapa mesin pencari web dan beberapa situs lain menggunakan web crawling untuk memperbarui konten web mereka atau indeks konten web lain. Web crawler dapat menyalin semua halaman yang dikunjungi untuk diproses kemudian oleh developer. Web crawler mengkonsumsi sumber daya pada sistem yang dikunjungi dan sering mengunjungi situs tanpa persetujuan. Web crawler dapat memvalidasi hyperlink dan kode HTML [5].

Sebuah crawler web dimulai dengan daftar URL untuk dikunjungi. Jika crawler bekerja untuk melakukan pengarsipan website itu salinan dan menyimpan informasi sebagai kelanjutannya. Arsip biasanya disimpan sedemikian rupa sehingga dapat dilihat, dibaca dan dinavigasikan karena arsip berada di web [6]. Arsip dikenal sebagai repositori dan dirancang untuk menyimpan dan mengelola koleksi halaman web. Repositori hanya menyimpan halaman HTML dan halaman ini disimpan sebagai file yang berbeda. Sebuah repositori mirip dengan sistem lain yang menyimpan data, seperti database modern. Satu-satunya perbedaan adalah bahwa repositori tidak perlu semua fungsi yang ditawarkan oleh sistem database. Repositori menyimpan versi terbaru dari halaman web diambil oleh crawler. Jumlah kemungkinan URL yang dikunjungi yang dihasilkan oleh perangkat lunak server-side juga telah membuat sulit bagi web crawler untuk menghindari mengambil duplikat konten. kombinasi tak berujung HTTP GET (berdasarkan URL) parameter yang ada, dari yang hanya pilihan kecil benar-benar akan kembali konten yang unik [7].

B. *Rule-Based System*

Dalam ilmu komputer, sistem rule-based digunakan sebagai cara untuk menyimpan dan memanipulasi data pengetahuan untuk menginterpretasikan informasi dalam cara yang bermanfaat [8]. Algoritma rule-based ini sering digunakan dalam aplikasi kecerdasan buatan dan penelitian. Sebuah

```

For each attribute,
  For each value of the attribute, make a rule as follows:
    count how often each class appears
    find the most frequent class
    make the rule assign that class to this attribute-value
  Calculate the error rate of the rules
  Choose the rules with the smallest error rate
    
```

Gambar. 1. Contoh dari algoritma Rule-Based System.

aplikasi yang menerapkan sistem rule-based memiliki sederatan aturan (rule) yang menjadi dasar dari penentuan (knowledge base) [9].

Teori ini menggunakan teknik yang sederhana, yang dimulai dengan dasar aturan yang berisi semua pengetahuan dari permasalahan yang dihadapi yang kemudian dikodekan ke dalam aturan IF-THEN dan sebuah tempat penyimpanan yang mengandung data, pernyataan dan informasi awal. Sistem akan memeriksa semua aturan kondisi (IF) yang menentukan subset, set konflik yang ada. Jika ditemukan, maka sistem akan melakukan kondisi THEN. Perulangan atau looping ini akan terus berlanjut hingga salah satu atau dua kondisi bertemu, jika aturan tidak diketemukan maka sistem tersebut harus keluar dari perulangan [10].

C. *Naïve Bayes Classification*

Menggunakan teorema Bayes untuk metode pengklasifikasian dengan menggunakan metode probabilitas dan statistic [11]. Memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya, dengan ciri utama memiliki asumsi yang sangat kuat dan memiliki independensi dari masing-masing kondisi. Naïve Bayes Classifier bekerja lebih baik dibanding model classifier lainnya dan membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian [12].

Algoritma Naïve Bayes adalah sebuah algoritma sederhana yang bekerja baik pada kategorisasi teks. Algoritma ini memanfaatkan metode probabilitas dan statistic yang dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya [13]. Dasar dari teori Naïve Bayes dalam pemrograman adalah rumus berikut ini:

$$P(c|x) := \frac{P(c)*P(x|c)}{P(x)} \tag{1}$$

Di mana:

P(x) adalah nilai input fitur yang digunakan

P(c) adalah frekuensi kelas yang ada

P(c|x) adalah nilai probabilitas input fitur yang dicari jika diketahui kelasnya

Apabila diberikan k atribut yang saling bebas, nilai probabilitas dapat diberikan sebagai berikut:

$$P(x_1, x_k|C) := P(x_k|C) * ... * P(x_1|C) \tag{2}$$

Algoritma yang digunakan juga adalah klasifikasi Naïve Bayes, dengan rumusan:

$$c^* = \operatorname{argmax}_c P(c|x) \quad (3)$$

Algoritma ini mencari nilai maksimum dari probabilitas yang didapat dari masing-masing kelas. Nilai probabilitas yang paling maksimum adalah nilai yang paling benar untuk kelas itu.

III. METODE DAN INTI PENELITIAN

Untuk menyelesaikan penelitian ini, dilakukan proses pengembangan aplikasi dengan menggunakan metodologi BUP (Basic Unified Process), yang merupakan versi RUP (Rational Unified Process) milik IBM yang lebih simple. Dalam BUP, development bagian-bagian sistem dikembangkan dalam waktu dan kecepatan yang berbeda, dan diintegrasikan setelah bagian-bagian sistem tersebut selesai dibuat. Developer akan kembali ke bagian yang telah selesai dibuat untuk memperbaiki dan meningkatkan bagian tersebut. BUP ini memiliki tujuan untuk menangani kelemahan yang ditemukan pada waterfall model, di mana BUP sendiri memiliki kemampuan yang lebih untuk mengakomodasi perubahan selama pengembangan software karena proses kerja yang berbentuk sebuah siklus. BUP memiliki deskripsi yang lebih simple dan menuju inti dari suatu permasalahan. BUP juga menyediakan proses yang terstruktur karena memiliki syarat yang spesifik dalam masing-masing tahapan pengerjaannya.

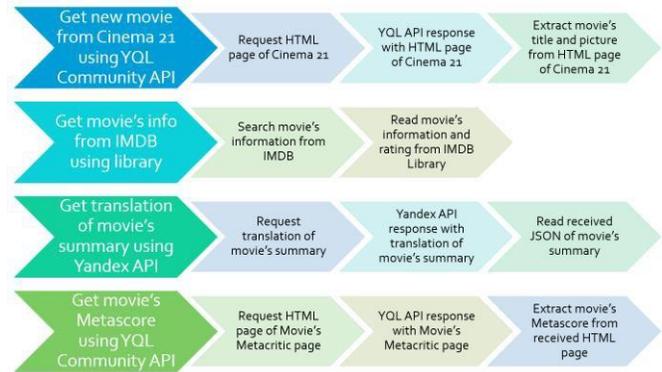
A. Proses Web-Crawling

Proses web-crawling ini merupakan proses di mana sistem akan mencari informasi dari sumber lain (website lainnya) dan menyimpan informasi-informasi tersebut ke dalam database. Sumber lain yang dimaksud adalah: website Cinema 21, IMDB, dan Metacritic. Untuk proses web-crawling ini, penulis menggunakan beberapa tools, yaitu:

- Library Curl: untuk mengirim request kepada layanan API dan membaca response JSON dari layanan API tersebut.
- Library IMDB: untuk mencari film pada IMDB dan mendapatkan informasi-informasi film pada IMDB.
- YQL Community API: untuk mendapatkan kode dari halaman web Cinema 21 dan Metacritic.
- Yandex Translation API: untuk mendapatkan terjemahan dari sinopsis film dari IMDB (yang berbahasa Inggris) ke dalam Bahasa Indonesia.

Untuk pengambilan informasi dari halaman web Cinema 21 dan Metacritic, sistem akan menggunakan metode web scraping, di mana sistem akan membaca dan memotong kode dari halaman web tersebut untuk mendapatkan informasi yang dibutuhkan. Metode ini membutuhkan layanan YQL Community API dan Library Curl, karena tanpa YQL API, Cinema 21 hanya akan menampilkan pesan error (Cookies not found). Library Curl ini juga dibutuhkan agar sistem bisa membaca pesan balasan dari YQL Community API yang berisi kode dari halaman HTML web.

Pada proses web-crawling ini sendiri, pertama-tama, sistem akan meminta kode halaman Cinema 21 kepada YQL



Gambar. 2. Proses web-crawl yang meliputi YQL Community API, IMDB Library, Yandex Translation API, dan Curl Library.

Community API, untuk kemudian mengambil judul film dan link dari poster film. YQL Community API ini diperlukan karena tanpa layanan API YQL ini, sistem hanya akan mendapatkan header dan footer dari halaman web Cinema 21.

Setelah sistem mendapatkan judul-judul film dari Cinema 21, maka sistem akan mencari informasi dari film-film tersebut dari IMDB dengan menggunakan Library IMDB. Adapun sistem hanya akan menggunakan library IMDB ini apabila judul film tersebut belum tercatat sebelumnya pada database. Setelah sistem selesai mendapatkan informasi-informasi film dari database, sistem akan meminta terjemahan dari sinopsis film kepada Yandex Translation API. Karena sinopsis film ini didapatkan dari IMDB, maka sinopsis film ini berbahasa Inggris. Tujuan pasar dari aplikasi web yang dibuat ini adalah bagi masyarakat Indonesia, dan karena itu, sinopsis dari film yang berbahasa Inggris ini harus diterjemahkan terlebih dahulu ke Bahasa Indonesia.

Apabila sistem telah menerima hasil terjemahan sinopsis film dari Yandex API, maka sistem akan kemudian mengambil Metascore dari film. Karena tidak ada layanan API atau library yang bisa digunakan untuk mengambil Metascore film, maka penulis harus mendapatkan informasi Metascore film ini dengan cara web-scraping. Sistem akan meminta kode halaman web Metacritic menggunakan YQL Community API, dan kemudian mengekstrak Metascore film dari barisan HTML web. Sama dengan proses pengambilan informasi dari IMDB sebelumnya, proses pengambilan Metascore film ini hanya akan dilakukan oleh sistem apabila judul film tersebut belum tercatat pada database.

B. Proses Pemilahan dan Penilaian Tweet

Proses pemilahan dan penilaian tweet ini merupakan langkah-langkah yang dilakukan oleh sistem demi mencapai tujuan akhir, yaitu mendapatkan sentiment dari tweet. Setiap tweet yang akan diproses akan diubah mengikuti format UTF-8, dan ditransformasikan menjadi huruf kecil seluruhnya menggunakan fungsi `strtolower()`. Proses ini terdiri dari tiga buah sub-proses, di mana ke-3 buah sub-proses itu sendiri adalah: Pre-processing data, Rule-Based System, dan Naïve Bayes Classifier. Pada proses pemilahan dan penilaian tweet ini, sistem akan menggunakan beberapa tools, yaitu:

- Sentiment Lexicon dataset, sebagai dataset

TABEL I
CONTOH PROSES PEMILAHAN DAN PENILAIAN TWEET

Input	wow filmnya #hangout bagus banget gan @radityadika gue kagum
Feature Reduction	wow filmnya HASHTAG bagus banget gan USERNAME gue kagum
Normalization	wow filmnya HASHTAG bagus sangat sahabat USERNAME saya kagum
Rule-Based	wow filmnya HASHTAG bagus sangat sahabat USERNAME saya kagum
Final Text	wow filmnya HASHTAG bagus sangat sahabat USERNAME saya kagum
Hasil Test	Review Film: Ya Sentimen: Positif

Contoh dari langkah-langkah yang dilakukan oleh sistem terhadap tweet untuk mendapatkan hasil akhir, yaitu sentimen tweet.

perbandingan pada Rule-Based sistem.

- Common Words Review dan Non-Review data, sebagai data perbandingan pada Rule-Based sistem.
- Library Sentiment Analyzer, untuk menentukan sentiment akhir dari tweet (positif/negatifnya suatu tweet).
- Movie Review Sentiment Polarity Dataset, sebagai training dataset untuk library Sentiment Analyzer [16].

Pada langkah pertama (pre-processing data), sistem akan melakukan dua buah tahapan, yaitu: Feature Reduction dan Normalization. Pada tahap Feature Reduction ini, sistem akan mengganti hashtag menjadi kata "HASTHTAG", link menjadi kata "URL", dan username menjadi kata "USERNAME". Sistem juga akan mengganti judul film menjadi "JUDULFILM", serta menghapus seluruh tanda baca. Pada tahap Normalization, sistem akan mengganti seluruh kata bukan baku dan kata singkatan yang terdapat pada tweets berdasarkan data kata singkatan yang telah disusun oleh penulis. Daftar kata singkatan dan kata bukan baku yang telah disusun oleh penulis ini terdiri dari 1991 baris.

Selanjutnya pada tahap Rule-Based System, sistem akan membandingkan tweet dengan beberapa data, yaitu: Sentiment Lexicon dataset, Common Words Review data, dan Common Words Non-Review data. Tahap Rule-Based ini sendiri digunakan oleh sistem untuk menentukan status dari tweet, apakah tweet tersebut merupakan review film atau bukan. Tiap kata/frase yang sama pada tweets ditemukan pada ke-3 data ini, akan diberi poin-poin tertentu sesuai poin dari kata/frase itu sendiri pada data tersebut.

Apabila total dari poin yang didapatkan oleh tweet tersebut berjumlah sama dengan atau lebih dari 10, maka sistem akan menilai tweet tersebut sebagai review film. Setelah itu, sistem juga akan membandingkan tweet dengan data sentiment negatif dan data sentiment positif yang umumnya ditemukan pada review film. Data sentiment positif dan negatif yang disusun ini memiliki nilai/poin yang berbeda-beda untuk masing-masing kata/frase. Tiap kata/frase yang sama pada tweet dengan masing-masing data sentiment ini, akan disimpan pada variable positivity dan negativity, untuk kemudian digunakan selanjutnya pada proses ke-3.

Pada proses ke-3, yaitu tahap Naïve Bayes Classifier, sistem akan mulai menentukan sentiment dari teks tweets [17]. Seluruh tweets yang telah ditentukan sebagai review akan masuk ke dalam proses ke-3 ini. Untuk proses ini, sistem akan

TABEL II
CONTOH HASIL PENGUJIAN

Text Akhir	Review		Sentimen	
	Sistem	Admin	Sistem	Admin
kelemahan HASHTAG karakter2nya konsep fantasinya ter develop konklusinya anti klimaks anyone lihat cerita	✓	✓	+	-
JUDULFILM puashati bagus menyesal bagus bagus bagus gila 10 10 well done	✓	X	+	-
HASHTAG furious8 oke pembahasan	✓	✓	+	+
JUDULFILM URL USERNAME USERNAME	X	X	-	-
both are good weh type spooky japanese culture	✓	✓	+	+
JUDULFILM or my neighbor totoro bagus kaget menghina cerita habis lihat JUDULFILM bodoh deh lihat URL	✓	✓	-	-
sadis kejam keren john wick chapter 2 URL	✓	✓	-	+
lihat JUDUL temen2 bilang rugi lihat bioskop video playlist USERNAME	✓	✓	-	-
URL JUDULFILM 2017 full movies download	X	X	-	-
JUDULFILM aja URL	X	X	-	-

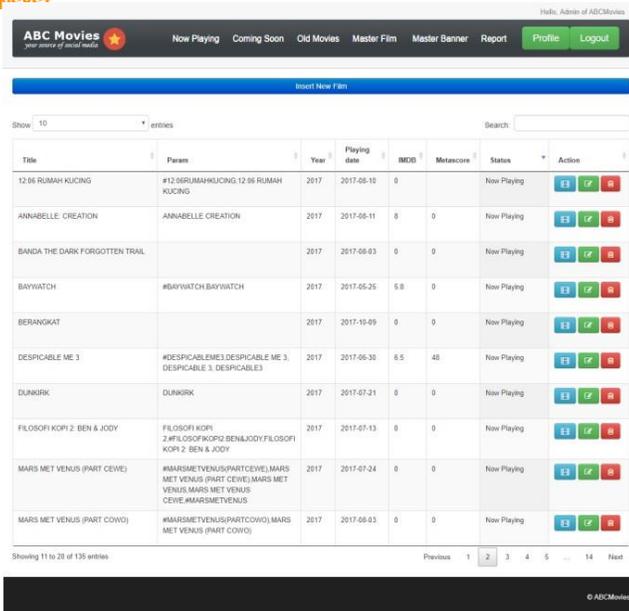
Beberapa contoh tweet yang telah mengalami proses pemilahan dan penilaian tweet serta nilai kebenaran dari tweet tersebut.

menggunakan Library Sentiment Analyzer, yang bisa memudahkan penulis dalam menggunakan algoritma Naïve Bayes. Apabila tweets telah diproses oleh library Sentiment Analyzer, sistem akan menambahkan hasil dari analisa library tersebut ke dalam variable negativity dan positivity. Apabila hasil dari variable negativity lebih besar dari positivity, maka sistem akan menilai tweets tersebut sebagai negatif, dan begitu pula sebaliknya.

IV. HASIL EKSPERIMEN DAN PENELITIAN

Data uji yang digunakan adalah beberapa judul film, seperti Beauty and The Beast, dan The Guys. Data pengujian ini (tweets) didapatkan dari Twitter dengan menggunakan Twitter Search API [18]. Data ujicoba tersebut kemudian akan diproses sesuai dengan sistem Rule-Based dan Naïve Bayes, yang kemudian hasil akhirnya (output) akan dibandingkan dengan data yang telah diklasifikasikan secara manual oleh penulis sebelumnya. Setelah uji coba selesai dilakukan, akan dilakukan penghitungan persentase kebenaran algoritma. Untuk menghitung persentase kebenaran klasifikasi data sendiri, yang dihitung tiga buah evaluasi data, yaitu: akurasi, presisi, dan recall. Apabila data FP terlalu banyak, maka nilai dari presisi algoritma akan menjadi jelek. Apabila data FN terlalu banyak, maka nilai dari recall algoritma akan menjadi jelek [19].

Akan dijabarkan mengenai uji coba yang telah dilakukan dalam rangka menguji kemampuan serta keakurasian algoritma yang diimplementasikan pada penelitian ini. Pada table 1 di atas telah diberikan hasil dari beberapa data uji yang



Gambar. 3. Contoh user interface dari web serta hasil crawling yang didapatkan dari website Cinema21, Metacritic, IMDB, dan Twitter.

telah diproses. Kolom sistem akan menunjukkan status apa yang ditentukan oleh sistem terhadap tweet tersebut (review/non-review dan positif/negatif), sedangkan kolom admin akan menunjukkan status sebenarnya dari tweet tersebut (menurut manusia).

Beberapa contoh tweets seperti pada table 1 di atas, telah mengalami pengklasifikasian tweet mana yang merupakan review film dan mana yang bukan, dengan menggunakan sistem Rule-Based. Setelah didapatkan tweet-tweet yang merupakan review, maka tweet-tweet tersebut akan diproses melalui Naïve Bayes Classification untuk menentukan sentiment dari review. Masing-masing algoritma yang digunakan (Rule-Based System dan Naïve Bayes Classifier), akan dihitung persentase kebenarannya. Untuk tiap-tiap pengujian yang akan dijelaskan berikut ini, digunakan jumlah tweet yang sama untuk tiap kali uji coba, yaitu sebanyak 6.323 tweet.

Pada pengujian Rule-Based System, dimana algoritma ini sendiri digunakan untuk menentukan apakah tweet termasuk film atau bukan, diberikan tindakan komparasi teks tweet dengan dataset lexicon dan common words. Pada tabel 1 di bawah ini dapat dilihat hasil evaluasi dari pengujian akhir terhadap Rule-Based System, di mana hasil presisi memiliki nilai akhir 82% dan recall memiliki nilai 74%. Hasil ini didapatkan dari pengujian sebanyak 4x dengan perlakuan yang berbeda-beda. Hasil akhir ini didapatkan ketika tiap common words diberi nilai untuk masing-masing kata/frase, sesuai dengan seringnya kata/frase tersebut digunakan pada tweet yang merupakan review film.

Pada pengujian Naïve Bayes Classifier, yang dimana algoritma ini digunakan untuk menentukan apakah tweet tersebut memiliki sentiment positif/negatif, diberikan tindakan komparasi teks tweet dengan common words sentiment dan juga penambahan training dataset dari Twitter sendiri. Dengan

TABEL III
EVALUASI OUTPUT PENGUJIAN
Rule-Based System

	Klasifikasi Review (TP) 1212	Klasifikasi Non-review (FN) 780	Total Tweets 1992
Tweets Review	(TP) 1212	(FN) 780	1992
Tweets Non-review	(FP) 336	(TN) 3324	3660
Total data yang diuji			5652
Akurasi = (TP + TN) / (TP + TN + FP + FN)			80,255 %
Precision = TP / (TP + FP)			78,294 %
Recall = TN / (TN + FN)			80,994 %
F-Measure = (2 * Precision * Recall) / (Precision + Recall)			79,621 %
Naïve Bayes Classifier			
	Klasifikasi Positif (TP) 692	Klasifikasi Negatif (FN) 95	Total Tweets 787
Tweets Positif	(TP) 692	(FN) 95	787
Tweets Negatif	(FP) 148	(TN) 277	425
Total data yang diuji			1212
Akurasi = (TP + TN) / (TP + TN + FP + FN)			79,950 %
Precision = TP / (TP + FP)			82,381 %
Recall = TN / (TN + FN)			74,462 %
F-Measure = (2 * Precision * Recall) / (Precision + Recall)			78,222 %

Kalkulasi kebenaran terhadap algoritma Rule-Based yang menentukan review/non-review dan Naïve Bayes yang menentukan positif/negatif uji coba akhir yang memiliki nilai recall sebesar 74% dan nilai presisi sebesar 82% ini, diberikan tindakan penambahan 400 buah training dataset dari Twitter. Hasil akhir ini didapatkan pula dengan penambahan komparasi teks tweet dengan common words sentiment, dimana tiap kata/frase yang terdapat pada data common words sentiment ini memiliki nilai yang berbeda-beda, sesuai dengan nilai dari positif/negatifnya kata/frase itu sendiri dalam sebuah review film.

V. KESIMPULAN

Pada pembuatan aplikasi web ini, Rule-Based System menghasilkan hasil akurasi akhir sebesar 80,26% sedangkan Naïve Bayes Classifier menghasilkan nilai akurasi sebesar 79,95%. Pada Rule-Based System, hasil akurasi ini didapatkan karena adanya penggantian kata-kata bukan baku dan penggantian kata-kata singkatan pada tweets yang didapatkan. Faktor lainnya yang mempengaruhi adalah karena adanya data Common Words yang berisi daftar kata-kata/frase yang sering digunakan pada review film dan Common Words yang berisi daftar kata-kata/frase yang sering digunakan pada tweets yang jelas bukan merupakan review film. Data pada Common Words ini berpengaruh besar pada hasil akurasi pada Rule-Based System, karena masing-masing baris data pada Common Words mengandung nilai-nilainya sendiri berdasarkan sering/tidaknya kata tersebut digunakan pada review/non-review.

Hasil akurasi yang besar pada Naïve Bayes Classifier dipengaruhi oleh data Common Words positif dan negatif dan juga training dataset dari Twitter. Sama dengan Common Words review/non-review pada Rule-Based System, data Common Words negatif/positif ini digunakan untuk menentukan sentiment dari tweets yang juga memiliki nilai-nilai sendiri. Training dataset dari Twitter sendiri juga diperlukan, karena tanpa training dataset dari Twitter, hasil presisi dan recall akan berkurang sebanyak 4%, di mana apabila hasil presisi dan recall turun, akan mempengaruhi hasil

dari akurasi dan F-Measure pula.

DAFTAR PUSTAKA

- [1] C. Basu, H. Hirsh, W. Cohen, and others, "Recommendation as classification: Using social and content-based information in recommendation," in *Aaai/iaai*, 1998, pp. 714–720.
- [2] C. Olston, M. Najork, and others, "Foundations and Trends® in Information Retrieval," *Found. Trends® in Inf. Retr.*, vol. 4, no. 3, pp. 175–246, 2010.
- [3] C. De Groc, "Babouk: Focused web crawling for corpus compilation and automatic terminology extraction," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, vol. 1, pp. 497–498.
- [4] J. Rennie and A. McCallum, "Efficient web spidering with reinforcement learning," 1999.
- [5] S. C. Herring, "Web content analysis: Expanding the paradigm," in *International handbook of Internet research*, Springer, 2009, pp. 233–249.
- [6] H. W. Park and M. Thelwall, "Hyperlink analyses of the World Wide Web: A review," *J. Comput. Commun.*, vol. 8, no. 4, p. JCMC843, 2003.
- [7] S. M. Schneider and K. A. Foot, "The web as an object of study," *New media & Soc.*, vol. 6, no. 1, pp. 114–122, 2004.
- [8] J. A. Bernard, "Use of a rule-based system for process control," *IEEE Control Syst. Mag.*, vol. 8, no. 5, pp. 3–13, 1988.
- [9] E. Lembessis, "Dynamic learning behaviour of a rule-based self-organising controller.," Queen Mary, University of London, 1984.
- [10] J. Efstathiou, "Rule-based process control using fuzzy logic," in *Approximate reasoning in intelligent systems, decision and control*, Elsevier, 1987, pp. 145–158.
- [11] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 129–136.
- [12] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *arXiv Prepr. cs/0212032*, 2002.
- [13] D. D. Lewis, "Representation and learning in information retrieval," University of Massachusetts Amherst, 1992.
- [14] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*, 1998, pp. 4–15.
- [15] R. B.-R. U. P. Content, "Basic Unified Process: A Process for Small and Agile Projects."
- [16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv Prepr. cs/0205070*, 2002.
- [17] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment analysis of review datasets using naive bayes and k-nn classifier," *arXiv Prepr. arXiv1610.09982*, 2016.
- [18] M. Mertiya and A. Singh, "Combining naive bayes and adjective analysis for sentiment detection on Twitter," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 2, pp. 1–6.
- [19] L. L. Dhande and G. K. Patnaik, "Analyzing sentiment of movie review data using Naive Bayes neural classifier," *Int. J. Emerg. Trends & Technol. Comput. Sci.*, vol. 3, no. 4, pp. 313–320, 2014.