Jurnal Teknik Elektro Vol. 8 No. 2 ISSN 1411 - 0059
Juli - Desember 2016

Prediksi *Rating* Film Menggunakan Metode Naïve Bayes

Riszki Wijayatun Pratiwi[1] dan Yusuf Sulistyo Nugroho[2]

Program Studi Informatika Fakultas Komunikasi dan Informatika, Universitas Muhamammdiyah Surakarta
Jl. A Yani Tromol Pos 1 Pabelan, Kartasura, 57102*, Jawa Tengah, Indonesia*
riszkiwp@gmail.com[1], yusuf.nugroho@ums.ac.id[2]

Abstrak— Pada saat ini perkembangan dunia perfilman sudah sangat pesat, contohnya dengan banyaknya film-film yang silih berganti untuk ditayangkan. Para penikmat film juga membutuhkan film-film yang mempunyai kualitas gambar, suara, alur cerita dan nilai positif yang baik dalam sebuah film, agar mereka tetap antusias dalam mengikuti film-film yang terbaru. Namun film-film yang ada tidak semuanya dapat dinikmati dan tidak semua kalangan menyukai semua film. Agar suatu film dapat terus berkembang, tentunya membutuhkan penilaian-penilaian dari para penikmat film, untuk mengetahui selera film yang sesuai dengan para penikmat film. Untuk itu dibutuhkan analisis agar dapat mengetahui bagaimana minat penikmat film yaitu dengan membuat penilaian-penilaian yang nantinya digunakan untuk mengetahui *rating* suatu film menggunakan metode naïve bayes yaitu metode yang melakukan pendekatan statistika yang fundamental dalam pengenalan pola (*pattern recognition*). Pendekatan ini didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan resiko yang ditimbulkan dalam keputusan-keputusan tersebut. Metode tersebut merupakan salah satu metode dari *data mining*, dengan atribut yang sudah ditentukan, yaitu meliputi *genre* film, aktor film, bahasa,warna, durasi film, negara, dan lainnya yang dapat digunakan sebagai tolak ukur sutradara untuk membuat film.

Kata kunci— analisa, *data mining*, film, naïve bayes

I. PENDAHULUAN

Setiap bentuk kesenian, meliputi seni musik, seni tari, seni sastra, seni rupa ataupun seni peran perlu sebuah apresiasi dari penikmatnya masing-masing. Secara umum, apresiasi seni mempunyai makna penghargaan terhadap kehadiran sebuah karya seni, sebuah karya seni mengalami suatu perkembangan dari tahun ke tahun sehingga pada akhirnya tercipta sebuah perpaduan yang imbang dan juga harmonis antara seni sastra, seni musik, seni peran dan juga komedi yang dibungkus dalam bentuk film. Film adalah sarana baru yang dipergunakan untuk menyebarkan suatu hiburan yang telah menjadi kebiasaan terdahulu, dan menyajikan cerita peristiwa, musik, drama, lawak, dan sajian teknis lainnya kepada masyarakat umum [1].

Namun tidak semua masyarakat menyukai film, atau ada beberapa jenis (*genre*) film yang disukai oleh masyarakat. Oleh karena itu, agar film berkembang terus-menerus maka dibutuhkan penilaian-penilaian dari masyarakat penikmat film untuk mengetahui selera film berjenis apa yang diinginkan oleh masyarakat penikmat film. Untuk itu dibutuhkan analisa agar dapat mengetahui minat penikmat film dengan cara menganalisis *rating* suatu film menggunakan teknik data mining yaitu dengan menganalisis data dari pendapat yang berbeda dan merangkumnya untuk memperoleh suatu informasi yang bermanfaat. Informasi yang didapatkan dari hasil *data mining* bisa digunakan untuk meningkatkan pendapatan atau mengurangi biaya produksi [2]. Pengumpulan data perlu dilakukan terlebih dahulu, biasanya data yang diperoleh bersifat *big data* yang berarti pengumpulan data dari berbagai macam sumber yang relevan [3].

Metode yang digunakan adalah naïve bayes yaitu metode yang mempunyai perhitungan matematik dasar yang sangat kuat serta dalam efisiensi klasifikasinya juga stabil, namun kekurangannya adalah parameter model naïve bayes perlu diperkirakan dan kurang peka terhadap data yang sudah hilang. Model naïve bayes memiliki tingkat kesalahan yang sangat minimum jika dibandingkan dengan algoritma klasifikasi lainnya [4]. Metode naïve bayes ini merupakan salah satu metode yang populer untuk pengkategorian teks dengan frekuensi kata sebagai fitur. Hal ini dapat disimpulkan bahwa fitur-fitur yang independen dapat dibuktikan dalam algoritma klasifikasi menjadi lebih efektif [5].

II. METODE

A. Penentuan Obyek Observasi

Observasi ini bertujuan untuk memprediksi *rating* sebuah film. Observasi ini dipilih karena untuk tolak ukur sebuah rumah produksi film ketika nantinya akan membuat film serta mengetahui selera film yang sesuai dengan para penikmat film.

B. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sampel dari berbagai situs di internet. Data tersebut digunakan sebagai data *training* dan juga data *testing*.

1) Data *Training*

Data *training* adalah data yang digunakan untuk perhitungan probabilitas dari data berdasarkan data

pembelajaran yang dilakukan. Data *training* yang digunakan adalah data sampel yang di dapat dari situs di internet, yaitu dari situs https://www.kaggle.com/.

2) Data *Testing*

Data *testing* merupakan data yang akan atau sedang terjadi dan dipergunakan sebagai bahan uji yang sebelumnya sudah didapatkan pada data *training*. Data *testing* tersebut juga menggunakan data sampel yang diperoleh dari situs di internet, yaitu dari situs https://www.kaggle.com/.

C. Penentuan Atribut

Atribut-atribut yang digunakan untuk proses data mining ini mengacu pada tujuan penelitian. Ada dua jenis variabel yang ditentukan [7],yaitu :

1) Variabel *dependen (Y)*

Variabel dependen (Y) merupakan variabel yang nilainya terikat, bisa disebut variabel terikat. Variabel Y yang digunakan yaitu imdb_score.

2) Variabel *independen (X)*

Variabel independen (X) merupakan variabel yang nilainya tidak tergantung pada nilai dari variabel lainnya atau bisa disebut sebagai variabel bebas. Variabel X yang digunakan terdiri dari :

- Variabel X1 Color
- Variabel X2 Director_name
- Variabel X3 Director_facebook_likes
- Variabel X4 Duration
- Variabel X5 Actor_1_name
- Variabel X6 Actor_1_facebook_likes
- Variabel X7 Actor_2_name
- Variabel X8 Actor_2_facebook_likes
- Variabel X9 Actor_3_name
- Variabel X10 Actor_3_facebook_likes
- Variabel X11 Gross
- Variabel X12 Genres
- Variabel X13 Movie_title
- Variabel X14 Num_voted_users
- Variabel X15 Cast_total_facebook
- Variabel X16 Facenumber_in_poster
- Variabel X17 Plot_Keywords
- Variabel X18 Num_users_for_reviews
- Variabel X19 Language
- Variabel X20 Country
- Variabel X21 Content_rating
- Variabel X22 Budget
- Variabel X23 Title_year
- Variabel X24 Movie_facebook_like
- Variabel X25 Num_critic_for_reviews

D. *Data Cleaning*

Pembersihan data perlu dilakukan supaya data yang digunakan valid sesuai kebutuhan. Sehingga dari nilai *class* data film dalam atribut tidak terjadi ketidakkonsistenan data dalam pengujian.

E. Penggunaan Metode Naïve Bayes

Naïve Bayes adalah sebuah pengelompokan statistik yang bisa di dipakai untuk memprediksi probabilitas anggota suatu *class*. Naïve Bayes juga mempunyai akurasi dan kecepatan yang sangat kuat ketika diaplikasikan pada *database* dengan *big data* [6].

Berikut rumus *naive bayes* [7] ditunjukkan pada persamaan 1.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

Keterangan :

X : Data dengan *class* yang belum diketahui
Y : Hipotesis data yaitu suatu class spesifik
P(Y|X) : Probabilitas hipotesis berdasar kondisi X (*posteriori probability*)
P(Y) : Probabilitas hipotesisY (*prior probability*)
P(X|Y) : Probabilitas X saat kondisi hipotesis Y
P (X) : Probabilitas X

III. Hasil dan Pembahasan

Data yang dipakai adalah data set yang diambil dari situs https://www.kaggle.com/ yang akhirnya dipakai untuk data *training* dan data *testing*. Kemudian dilakukan *prepocessing* pada data tersebut dengan cara menerapkan *metode* cleaning. Lalu data tersebut melalui pengolahan, pengolahan data menggunakan *software rapid miner* yang nantinya bisa menghasilkan sebuah prediksi *rating* film.

TABEL I. DATA SEBELUM *PREPOCESSING*

color	director_name	num_critic_for_reviews	Duration	director_facebook_likes	actor_3_facebook_likes
Color	James Cameron	723	178	0	855
Color	Gore Verbinski	302	169	563	1000
Color	Sam Mendes	602	148	0	161
Color	Christopher Nolan	813	164	22000	23000
Color	Andrew Stanton	462	132	475	530
Color	Sam Raimi	392	156	0	4000
Color	Nathan Greno	324	100	15	284
Color	Joss Whedon	635	141	0	19000
Color	David Yates	375	153	282	10000
Color	Zack Snyder	673	183	0	2000

TABEL II. DATA SESUDAH *PREPOCESSING*

color	director_name	num_critic_for_reviews	duration	Director facebook_like	actor_3_facebook_like
Color	James Cameron	tinggi	panjang	rendah	rendah
Color	Gore Verbinski	sedang	panjang	sedang	sedang
Color	Sam Mendes	tinggi	panjang	rendah	rendah
Color	Christopher Nolan	tinggi	panjang	tinggi	tinggi
Color	Andrew Stanton	tinggi	panjang	rendah	sedang
Color	Sam Raimi	sedang	panjang	rendah	rendah
Color	Nathan Greno	sedang	pendek	rendah	rendah
Color	Joss Whedon	tinggi	panjang	rendah	rendah
Color	David Yates	sedang	panjang	rendah	rendah
Color	Zack Snyder	tinggi	panjang	rendah	rendah

TABEL III. DATA *TRAINING*

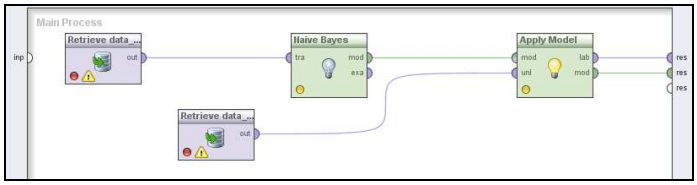
color	Director name	num_critic_for_reviews	duration	director_facebook_like	actor_3_facebook_like
Color	Barry Levinson	rendah	panjang	rendah	rendah
Color	Brad Silberling	rendah	pendek	rendah	rendah
Black and White	Quentin Tarantino	sedang	pendek	tinggi	tinggi
Color	Frank Oz	rendah	pendek	rendah	rendah
Black and White	Quentin Tarantino	sedang	panjang	tinggi	tinggi
Color	Andrey Konchalovskiy	rendah	pendek	rendah	rendah
Color	Robert Zemeckis	rendah	pendek	rendah	rendah
Color	Tom Dey	rendah	pendek	rendah	rendah
Color	Stuart Baird	rendah	panjang	rendah	rendah
Color	Mark Waters	sedang	pendek	rendah	rendah

TABEL IV. DATA *TESTING*

color	director_name	num_critic_for_reviews	duration	director_facebook_like	actor_3_facebook_like
Color	James Cameron	tinggi	panjang	rendah	rendah
Color	Gore Verbinski	sedang	panjang	sedang	sedang
Color	Sam Mendes	tinggi	panjang	rendah	rendah
Color	Christopher Nolan	tinggi	panjang	tinggi	tinggi
Color	Andrew Stanton	tinggi	panjang	rendah	sedang
Color	Sam Raimi	sedang	panjang	rendah	rendah
Color	Nathan Greno	sedang	pendek	rendah	rendah
Color	Joss Whedon	tinggi	panjang	rendah	rendah
Color	David Yates	sedang	panjang	rendah	rendah
Color	Zack Snyder	tinggi	panjang	rendah	rendah

A. Implementasi *Software* Rapid Miner

Penggunaan *software* ini bisa mengimpor sebuah informasi yang terdapat dari berbagai macam sumber *database* untuk diperiksa dan dianalisa didalam sebuah aplikasi. Rapid Miner sebagai solusi untuk memprediksi dan menganalisa komputasi statistik [3]. Gambar 1 hingga Gambar 4 merupakan hasil penelitian yang dilakukan menggunakan aplikasi Rapid Miner.



Gambar 1. Klasifikasi Naïve Bayes

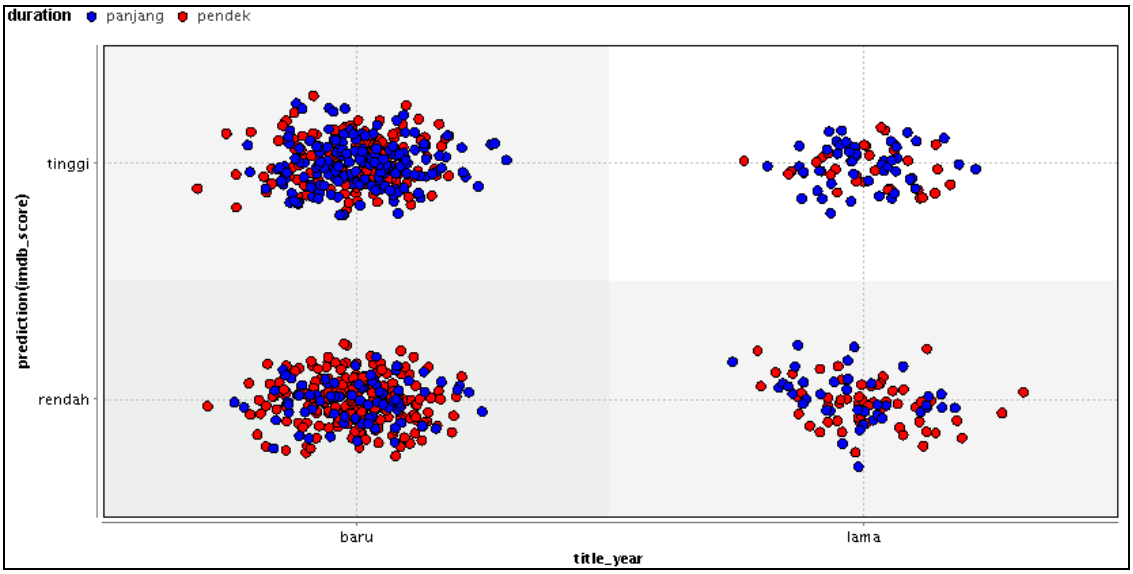
SimpleDistribution

```
Distribution model for label attribute imdb_score


Class rendah (0.707)
25 distributions

Class tinggi (0.293)
25 distributions
```

Gambar 2. Hasil Naïve Bayes pada *text view*



Gambar 3. Hasil Naïve Bayes pada *plot view*

accuracy: 55.80% +/- 2.20% (mikro: 55.80%)

	true rendah	true tinggi	class precision
pred. rendah	1263	469	72.92%
pred. tinggi	857	411	32.41%
class recall	59.58%	46.70%	

Gambar 4. Tingkat akurasi pada hasil *Naïve Bayes*

Gambar 1 menunjukan model distribusi *Naïve Bayes*. Pada hasil *naïve bayes* bisa dilihat bahwa model distribusi nilai *class* "RENDAH" sebanyak 0,707, sedangkan *class* "TINGGI" sebanyak 0,293, ditunjukan pada Gambar 2. Berdasarkan Gambar 3 menunjukan penentuan *rating* film (*imdb_score*) tinggi apabila tahun film (*title_year*) baru serta *duration* film panjang.

B. Pengujian

Hasil klasifikasi yang didapat tersebut kemudian dihitung nilai *accuraty, precision, recall.* Berdasarkan Gambar 4 nilai *accuracy* 55,80%, *precision* 32,41% dan *recall* 46,70 %

IV. PENUTUP

Naïve Bayes adalah salah satu klasifikasi berjenis teks, contoh nya adalah prediksi *rating* film. Naïve Bayes sederhana dan efisien, serta sangat familiar jika dipakai untuk pengklasifikasian teks dan mempunyai performa yang bagus pada banyak domain. Hasil penelitian menunjukkan bahwa hasil prediksi *rating* film menggunakan metode naïve bayes memiliki *accuracy* 55,80%, *precision* 32,41%, dan *recall* 46,70%, Berdasarkan analisa yang didapat menggunakan data set dari situs https://www.kaggle.com/ menunjukan bahwa mayoritas prediksi *rating* film RENDAH.

REFERENSI

[1] Mudjiono, Y. (2011). Kajian Semiotika Dalam Film. Jurnal Ilmu Komunikasi, Vol. 1, No.1, April 2011 ISSN: 2088-981X KAJIAN, 1(1), 123.

[2] Sharma, P., Singh, D., & Singh, A. (2015). Classification Algorithms on a Large Continuous Random Dataset Using Rapid. IEEE Sponsored 2nd International Conference on Electronics and Communication System(ICECS 2015), (Icecs), 704–709.

[3] Utmal, M., & Pandey, R. K. (2015). Taxonomy on the Integration of Hadoop and Rapid Miner for Big Data Analytics. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (hal. 890–893). http://doi.org/10.1109/CICN.2015.175.

[4] Liu, J., Tian, Z., Liu, P., Jiang, J., & Li, Z. (2016). An Approach of Semantic Web Service Classification Based on Naive Bayes. 2016 IEEE International Conference on Services Computing An. http://doi.org/10.1109/SCC.2016.53.

[5] Chandrasekar, P., & Qian, K. (2016). The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier. 2016 IEEE 40th Annual Computer *Software* and Applications Conference (COMPSAC), 2, 618–619. http://doi.org/10.1109/COMPSAC.2016.205

[6] Widiastuti, N. A., Santosa, S., & Supriyanto, C. (2014). Algoritma Klasifikasi Data Mining Naïve Bayes Berbasis Particle Swarm Particle Swarm Optimization Untuk Deteksi Penyakit Jantung. Jurnal Pseudocode, 1, 11–14.

[7] Nugroho, Y. S., & Setyawan. (2014). Klasifikasi Masa Studi Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta Menggunakan Algoritma C4.5. KomuniTi, Vol. VI, No. 1 Maret 2014, VI(1), 84–91.