e-ISSN. 2685-4066
DOI 10.30812/bite.v2i1.804

Jurnal BITe
Vol.2 No.1 (Juni) 2020, Hal 40-44

Analisis Sentimen *Movie Review* Menggunakan *Naive Bayes Classifier* Dengan Seleksi Fitur *Chi Square*

Ahmad Zuli Amrullah[1], Andi Sofyan Anas[2], Muh. Adrian Juniarta Hidayat[3]

[1,2,3] Fakultas Teknik dan Desain, Universitas Bumigora
zuli@universitasbumigora.ac.id, andi.sofyan@universitasbumigora.ac.id,
m.adrian@universitasbumigora.ac.id

Abstrak

Ulasan film berisi opini atau pandangan penonton terhadap suatu karya *film*, dalam hal ini gambaran secara umum dan detail sebuah *film*. Banyaknya respon dari penonton terhadap suatu *film* belum bisa dikategorikan secara langsung menjadi sebuah s*entiment*, untuk itu perlunya sebuah sentimen analisis. Analisis sentimen adalah subjek utama dalam *machine learning* yang bertujuan untuk mengekstrak subjektif informasi dari ulasan tekstual. Pada penelitian ini akan melakukan analisis sentiment pada *movie review* yang didapat dari IMDB untuk menganalisis respon penonton terhadap *film* yang mereka tonton kedalam dua kelompok; respon positif dan negatif. Proses analisis dilakukan dengan menggunakan *text mining* dalam mengekstraksi informasi yang diperoleh dan diklasifikasi dengan *Naïve Bayes*. Sentimen respon akan diuji dengan *Chi Square*.

Kata kunci: *Sentiment Analysis, text mining, movie, naïve bayes, chi square*

Abstract

Movie reviews contain the viewer's opinion or perspective of a movie, based on a general and detailed description. The number of responses from the audience to a movie cannot be categorized directly into a sentiment; for this reason, an analysis of sentiment is needed. Sentiment analysis is the main subject in machine learning which aims to extract subjective information from textual reviews. In this study, the dataset of movie reviews is obtained from IMDB to analyze audience responses to the movies they watch in two groups; positive and negative. The analysis conducted text mining in extracting information is obtained and classified with Naïve Bayes. Response sentiments will be tested with Chi-Square.

Keywords: *Sentiment Analysis, text mining, movie, naïve bayes, chi square*

I. INTRODUCTION

Text mining, mengacu pada proses mengambil informasi berkualitas tinggi dari teks. Informasi berkualitas tinggi biasanya diperoleh melalui peramalan pola dan kecenderungan melalui sarana seperti pembelajaran pola statistik. Proses *text* mining yang khas meliputi kategorisasi teks, *text clustering*, ekstraksi konsep/entitas, produksi taksonomi *granular*, *sentiment analysis*, penyimpulan dokumen, dan pemodelan relasi entitas.

Sebuah teks dapat terdiri dari hanya satu kata ataupun susunan kalimat [1]. Pengambilan informasi dari teks (*text mining*) antara lain dapat meliputi kategorisasi teks atau dokumen, analisis sentimen (*sentiment analysis*), pencarian topik yang lebih spesifik (*search engine*), serta spam filtering. Gagasan umum *text mining* adalah untuk mengetahui cakupan atau topik dari permasalahan dalam teks [2]. Text mining penting dalam analisis sentimen sebagai pengidentifikasi emosional suatu pernyataan, sehingga banyak studi tentang analisis sentimen dilakukan [3].

Analisis sentimen adalah studi komputasi dari opini-opini, sentimen, serta emosi yang diekspresikan dalam teks [4]. Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau pendapat. Polaritas mempunyai arti apakah teks yang ada dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif.

Salah satu metode klasifikasi yang dapat digunakan adalah metode *Naïve Bayes* yang sering disebut dengan *Naïve Bayes Classifier* (NBC). Kelebihan NBC adalah sederhana tetapi memiliki akurasi yang tinggi. Berdasarkan hasil eksperimen,

NBC terbukti dapat digunakan secara efektif untuk mengklasifikasikan berita secara otomatis dengan akurasi mencapai 90.23%. Algoritma NBC yang sederhana dan kecepatannya yang tinggi dalam proses pelatihan dan klasifikasi membuat algoritma ini menarik untuk digunakan sebagai salah satu metode klasifikasi (Yudi Wibisono, 2008).

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

II. METODOLOGI

Data dalam penelitian ini berupa opini berbahasa inggris tetang film (*movie review*). Dari data yang tersedia terdapat sebanyak 1400 buah opini, terdiri dari 700 buah opini positif dan 700 buah opini negative. Data tersebut digunakan sebagai data pada *machine learning* dan data uji untuk mengevaluasi kinerja sistem. Adapun langkah-langkah untuk perancangan analisis sentiment yang dibahas adalah sebagai berikut.

2.1. Teks Mining

Text mining dapat didefinisikan secara luas sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya dalah kategorisasi. Text mining bisa dianggap subjek riset yang tergolong baru. Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokan dan menganalisa unstructured text dalam jumlah besar[5].

2.2. Teks Preprocessing

Tahap *text preprocessing* adalah tahap awal dari text mining. Tahap ini mencakup semua rutinitas, dan proses untuk mempersiapkan data yang akan digunakan pada operasi knowledge discovery sistem text mining. Tindakan yang dilakukan pada tahap ini adalah toLowerCase, yaitu mengubah semua karakter huruf menjadi huruf kecil dan Tokenizing[6].

Secara garis besar tokenisasi adalah tahap memecah sekumpulan karakter dalam suatu teks kedalam satuan kata. Sekumpulan karakter tersebut dapat berupa karakter whitespace, seperti enter, tabulasi, spasi. Namun untuk karakter petik tunggal („), titik (.), semikolon (;), titik dua (:) atau lainnya, juga dapat memiliki peran yang cukup banyak sebagai pemisah kata. Sebuah titik (.) biasanya untuk tanda akhir kalimat, tapi dapat juga muncul dalam singkatan, inisial orang, alamat internet,dll. Kemudian tanda hypen (-) biasanya muncul untuk menggabungkan dua token yang berbeda untuk membentuk token tunggal. Tapi dapat pula ditemukan untuk menyatakan rentang nilai, kata berulang, dsb. Atau karakter slash (/) sebagai pemisah file atau direktori atau url ataupun untuk menyatakan “dan atau”[7].

2.3. *Fitur Selection*

Tahap seleksi fitur (feature selection) bertujuan untuk mengurangi dimensi dari suatu kumpulan teks, atau dengan kata lain menghapus kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dokumen sehingga proses pengklasifikasian lebih efektif dan akurat [5]. Pada tahap ini tindakan yang dilakukan adalah menghilangkan stopword (stopword removal) dan stemming terhadap kata yang berimbuhan [5].

Stopword adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen [5]. Misalnya “di”, “oleh”, “pada”, “sebuah”, “karena” dan lain sebagainya. Stemming adalah proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya (stem) [6]. Tujuan dari proses stemming adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata.

Setelah melalui proses stopword removal tindakan selanjutnya adalah yaitu proses stemming. Stemming adalah proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya. Tujuan dari proses stemming adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata.

2.4. Sentiment Analysis

Sentiment analysis atau opinion mining mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan text mining yang bertujuan menganlisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik , produk, layanan, organisasi, individu, ataupun kegiatan tertentu.

Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudia menentukan pendapat yang dikemukakan dalam kaliamat atau dokumen tersebut apakah bersifat positif, negatif atau netral. Sentiment analysis juga dapat menyatakan perasaan emosional sedih, gembira, atau marah. Kita dapat mencari pendapat tentang produk-produk, merek atau orang-orang dan

menentukan apakah mereka dilihat positif atau negatif di web.

2.5. Naïve Bayes

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

$$p(C \mid D) = \frac{p(D \mid C)\, p(C)}{p(D)} \quad (1)$$

2.6. Chi Square

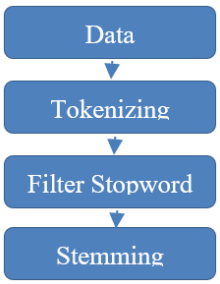
Seleksi fitur dilakukan untuk mereduksi fitur-fitur yang tidak relevan dalam proses klasifikasi oleh NBC. Terdapat beberapa metode untuk penyeleksian fitur yaitu *Mutual Information* (MI), chi square ($x2$), dan yang umum digunakan adalah frequency-based. Seleksi fitur frequency-based menggunakan jumlah kemunculan term atau frekuensi term yang diurutkan dari yang paling banyak sampai paling sedikit dan diambil beberapa urutan atas untuk digunakan sebagai fitur. Seleksi fitur MI merupakan ukuran yang mengukur kehadiran atau ketidakhadiran sebuah term yang memberikan kontribusi kepada kategori yang tepat. Sedangkan seleksi fitur Chi Square menggunakan teori statistika untuk menguji independensi sebuah term dengan kategorinya. Salah satu tujuan penggunaan seleksi fitur adalah untuk menghilangkan fitur pengganggu dalam klasifikasi[9].

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

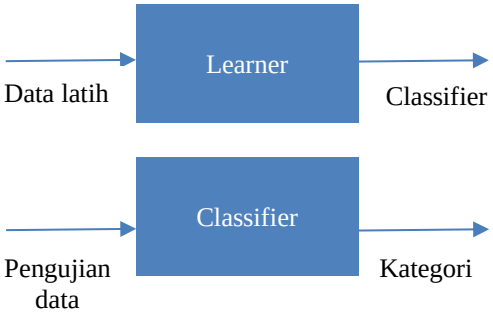
$$x^2(D,t,c) = \sum_{et\in\{1,0\}} \sum_{ec\in\{1,0\}} \frac{(N_{et\ el} - E_{et\ el})}{E_{et\ el}} \quad (2)$$

2.7. Tahap preprocessing data

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



Gambar 1 Proses *Preprocessing*



Gambar 2. Proses Naïve bayes

2.8. Tahap penyeleksian fitur dengan Chi Square

Tahap ini akan melakukan seleksi fitur dengan menggunakan *Chi Square*. Pertama menentukan menentukan table kontingensi masing-masing fitur dengan dengan table 1.

Tabel 1 Tabel kontingensi seleksi fitur Chi Square

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

Langkah selanjutnya menghitung nilai seleksi fitur Chi Square dengan persamaan berikut.

$$x^2(D,t,c) = \frac{(N_{00}+N_{11}+N_{10}+N_{01}) \times (N_{00}N_{11} - N_{10}N_{01})}{(N_{00}+N_{11}) \times (N_{00}+N_{11}) \times (N_{00}+N_{11}) \times (N_{00}+N_{11})} \quad (3)$$

Seleksi fitur *Chi Square* digunakan untuk pengamatan berkesesuaian (*goodness of fit*) dari kategori dengan *term*. Uji Chi Square dalam statistika diterapkan untuk menguji independensi dari dua peristiwa. Sedangkan dalam seleksi fitur berdasarkan teori statistika, dua peristiwa tersebut diantaranya adalah, kemunculan dari fitur dan kemunculan dari kategori.

2.9. *Tahapan Klasifikasi Data*

Naïve Bayes menganggap sebuah dokumen sebagai kumpulan dari kata yang menyusun dokumen tersebut. Naïve Bayes juga tidak

memperhatikan urutan kemunculan kata pada dokumen.

III. HASIL DAN PEMBAHASAN

Penggunaan data sejumlah 1400 data opini tentang review film berbahasa inggris. Opini tersebut terbagi menjadi 700 data opini positif dan 700 data opini negative. Sebagai data latih diugunakan 1000 data yaitu terbagi menjadi 500 buah opini positif dan 500 buah opini negative. Sisanya 400 data yang terdiri dari opini positif dan negative digunakan sebagai data uji. Pada penelitian ini dilakukan 2 pengujian dengan membagi data uji dengan data training.

Tabel 2 Proses pembagian data training dan uji

No	Data Training	Data Uji
1	200 opini negative dan 200 opini positif	500 opini negative dan 500 opini positif
2	500 opini negative dan 500 opini positif	200 opini negative dan 200 opini positif

Hasil dari percobaan diatas menunjukkan signifikasi hasil akurasi dengan klasifikasi Naïve Bayes ketika data training ditambahkan.

Tabel 3 Hasil akurasi

No	Akurasi
1	56.40%
2	64.40%

Implementasi seleksi fitur Chi Square pada sistem, proses-proses atau urutan proses dirancang untuk mengklasifikasi data uji dengan melalui beberapa tahapan. Tahapan tersebut merupakan tahapan yang telah diuraikan pada tahap pre-processing dan implementasi Naïve Bayes classifier.

Langkah awal dengan melakukan pelabelan secara manual untuk menentukan labeling. Pada tahap penyeleksian fitur ada 2 metode yang digunakan, chi-square dan frequency based. Seleksi fitur *frequency-based* menggunakan jumlah kemunculan *term* atau frekuensi *term* yang diurutkan dari yang paling banyak sampai paling sedikit dan diambil beberapa urutan atas untuk digunakan sebagai fitur.

Metode seleksi fitur *Frequency-based* memilih fitur yang paling umum di kategori.Metode seleksi fitur *Frequency-based* dapat didefinisikan baik sebagai frekuensi dokumen (jumlah dokumen di kategori c yang mengandung fitur t) atau sebagai koleksi frekuensi (jumlah token dari t yang muncul pada dokumen dalam c). Sedangkan seleksi fitur Chi Square menggunakan teori statistika untuk menguji independensi sebuah term dengan kategorinya. Salah satu tujuan penggunaan seleksi fitur adalah untuk menghilangkan fitur pengganggu dalam klasifikasi. Maning, et al.

Hipotesis awal menyatakan bahwa *term t* independen terhadap kategori *c*. Sedangkan hipotesis akhir menyatakan bahwa *term t* dependen terhadap kategori *c*.

Tabel 4 Peringkat Hasil seleksi Chi Square

No	Fitur	Nilai chi square	Kategori
1	bad	43.17881473	negative
2	wast	40.34444444	negative
3	stupid	24.0456621	negative
4	enjoy	24.04463119	positive
5	great	22.74286879	positive
6	bore	21.43856877	negative
7	terribl	20.51396316	negative
8	fail	19.8683394	negative
9	strong	19.613711	positive
10	hilari	19.31933029	positive
11	deal	19.23705395	positive
12	obvious	18.95894328	positive
13	want	18.73329677	positive
14	worst	18.66482493	negative
15	perfect	18.61688312	positive
16	avoid	18.46336996	negative
17	dull	17.58361566	negative
18	poor	17.13932956	negative
19	delight	17.11733333	positive
20	fun	15.82780369	positive

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

IV. KESIMPULAN

Berdasarkan hasil yang diperoleh dapat disimpulan bahwa kemunculan frekuensi fitur pada kategori yang diharapkan dan kategori yang tidak diharapkan memiliki peranan penting dalam seleksi fitur *Chi Square*, oleh karena itu seleksi fitur *Chi Square* baik digunakan dalam penyeleksian fitur. pembangunan sistem analisis *sentiment* menggunakan metode NBC merupakan suatu yang baik namun terdapat kesalahan klasifikasi karena pada data uji terdapat fitur yang muncul pada bukan kategorinya. Pada penelitian kedepannya dapat menggunakan algoritma yang lain, seperti SVM.

REFERENSI

[1] M. Carter, R. & McCarthy, *Cambridge Grammar of English*. Cambridge: Cambridge Univ. Press, 2006.

[2] H. Maning, C., Raghavan, P. & Schutze, *Introduction to Information Retrieval.* London: Cambridge University Press., 2008.

[3] H. Zhang, L., Ghosh, R., Dekhil, M. and B. M. & Liu, *Combining Lexiconbased and Learning-based Methods for Twitter Sentiment Analysis*. Chicago: Hewlett-Packard Development Company, L.P., 2011.

[4] B. Liu, *Sentiment Analysis and Opinion Mining*. Rafael: Morgan & Claypool Publishers, 2012.

[5] J. Feldman, Ronen., Sanger, "No Title," *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. [Online]. Available: http://www.books24x7.com/marc.asp?bookid=23164.

[6] B. Kurniawan, S. Effendi, and O. S. Sitompul, "Klasifikasi Konten Berita Dengan Metode Text Mining," *J. Dunia Teknol. Inf.*, vol. 1, no. 1, pp. 14–19, 2012.

[7] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah," no. 5.

[8] J. Aldrich, *R. A. Fisher on Bayes and Bayes' Theorem. Bayesian Analysis, 3(1)*. 2006.

[9] J. Ling, I. putu E. N. Kencana, and T. B. Oka, "Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square," *E-Jurnal Mat.*, vol. 3, no. 3, pp. 92–99, 2014.