REPOSITOR, Vol. 1, No. 2, Desember 2019, Pp. 125-130
ISSN : 2714-7975
E-ISSN : 2716-1382

Klasifikasi Sinopsis Novel Menggunakan Metode Naïve Bayes Classifier

Vinna Rahmayanti Setyaning Nastiti [*1], Setio Basuki[2], Hilman[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
vinastiti@umm.ac.id[*1], setio_basuki@umm.ac.id[2], hlmnft@gmail.com[3]

Abstrak

Tidak dapat dipungkiri kemajuan teknologi berkembang dengan sangat cepat terutama di bidang komputer, saat ini dengan komputer pekerjaan yang mulanya dikerjakan oleh manusia dapat di ambil alih komputer guna membantu pekerjaan manusia itu sendiri, seperti halnya studi kasus pada penelitian ini yaitu sebuah sistem yang dapat mengklasifikasikan teks berupa sinopsis kedalam kelompok genrenya. Genre adalah gaya cerita dalam sebuah novel, terdapat banyak genre pada novel diantaranya genre romantis, komedi, misteri, horor dan lain-lain, dengan mengetahui genre novel pembaca akan dapat mengetahui gaya cerita novel tersebut. Metode yang digunakan pada penelitian ini adalah metode TF-IDF (Term Frequency Inverse Document Frequency) dan Naïve Bayes Classifier. Metode TF-IDF digunakan untuk mendapatkan bobot dari setiap kata yang terkandung dalam dokumen yang hasilnya digunakan dalam metode Naïve Bayes Classifier untuk mendapatkan hasil klasifikasi sinopsis kedalam bentuk genre. Berdasarkan evaluasi menggunakan confusion matrix dengan menggunakan 600 data latih dan 200 data uji didapatkan akurasi sebesar 80,5%.

Kata Kunci: *Klasifikasi, Genre, TF-IDF, Naïve Bayes Classifier*

Abstract

It is undeniable that technological progress is developing very quickly in the field of computers, now with computers the work that was originally done by humans can be taken over by computers to help human work itself, like case studi of this research is a system that can classification the text like synopsis into genre group. Genre is the style of story in a novel, there are many genres in the novel that are expected to be romantic, comedy, mystery, horror and others, by knowing the genre of the novel the reader will be able to know the story style of the novel. The method used in this research is TF-IDF (Term Frequency Inverse Document Frequency) and Naïve Bayes Classifier. The TF-IDF method is used to get the weight of each word contained in the resulting document is used in the Naïve Bayes Classifier method to get the synopsis classification results into genre. Based on the evaluation using a confusion matrix using 600 training data and 200 test data obtained an accuracy of 80.5%.

Keywords: *Classification, Genre, TF-IDF, Naïve Bayes Classifier*

1. Pendahuluan

Novel adalah sebuah karya prosa fiksi yang ditulis secara naratif, dalam kamus besar bahasa Indonesia novel adalah karangan panjang yang mengandung rangkaian cerita kehidupan sesesorang dengan orang disekitarnya. Beragam novel yang ada menyebabkan novel dapat dikelompokkan berdasarkan jenisnya diantaranya adalah berdasarkan kebenaran cerita yaitu fiksi dan nonfiksi, berdasarkan genre yaitu horor, romantis, misteri, komedi, inspiratif, dan lain-lain, berdasarkan isi dan tokoh yaitu *teenlit*, *chicklit*, *songlit*, dewasa [1], dan metropop [2].

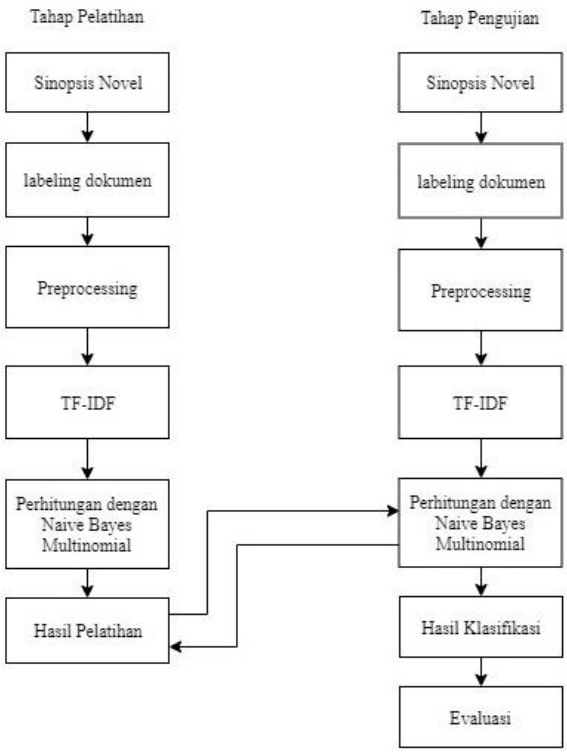
Pada setiap toko buku, diketahui bahwa tata letak buku akan disusun dengan sedemikian rupa agar memudahkan baik pengelola toko maupun pembeli untuk menemukan dan mengembalikan buku ke tempatnya, hal ini mudah dilakukan secara manual dengan buku-buku dengan sampul yang memuat banyak informasi mengenai buku tersebut sehingga mudah bagi pengelola membuat pengelompokkan, akan tetapi ada buku-buku yang tidak memiliki informasi yang cukup seperti novel yang secara umum pada bagian sampul hanya memiliki informasi judul, dan sinopsis. Dari permasalahan ini penulis melakukan penelitian tentang sebuah sistem yang dapat melakukan klasifikasi otomatis dengan menggunakan sinopsis pada novel dengan metode *Term Frequency Inverse Document Frequency* dan *Multinomial Naïve Bayes*

Naïve Bayes Classifier atau disingkat NBC adalah salah satu metode klasifikasi yang berakar pada teorema bayes dimana proses klasifikasi dilakukan dengan melakukan perhitungan probabilitas dari suatu data [3]. Pada kasus [4] metode NBC digunakan mengklasifikasikan artikel berita, data yang diteliti adalah data artikel berita berbahasa Indonesia yang diambil dari web portal. Pada kasus [5], NBC digunakan untuk mengklasifikasikan dokumen dengan konten *E-government* dengan data berupa dokumen dengan format HTML yang diubah menjadi TXT. Pada kasus [6], NBC digunakan untuk mengklasifikasikan kategori cerita pendek dengan data berupa kumpulan cerita-cerita pendek. Pada kasus [7], NBC digunakan untuk mengklasifikasikan dokumen berita olahraga dengan data yang didapatkan dari web portal berita olahraga.

Term Frequency Inverse Document Frequency atau TF-IDF adalah metode pembobotan kata yang bertujuan untuk memberikan bobot nilai pada setiap kata. Pada kasus [8], klasifikasi menggunakan metode *Naïve Bayes* menggunakan hasil dari TF-IDF memberikan hasil rata-rata akurasi yang lebih baik dibandingkan tanpa menggunkan TF-IDF.

Penelitian menggunakan metode NBC banyak ditemui pada klasifikasi dokumen berita [3] [4] [7], peneliti akan melakukan klasifikasi sinopsis novel menggunakan metode *Naïve Bayes Classifier* dan *Term Frequency Inverse Document Frequency* dengan harapan hasil yang didapatkan dapat membantu pengelola toko.

2. Metode Penelitian



Gambar 1. Tahapan Klasifikasi

Alat dan Bahan yang digunakan dalam pengembangan sistem ini adalah sebagai berikut:

1. Laptop dengan spesifikasi ASUS A456UQ, Processor Intel Core i7 7500 up to 2.90GHz, RAM 8GB DDR3.
2. Sistem Operasi Microsoft Windows 10 Pro (64-bit).
3. Bahasa Pemrograman Python3 dan *framework jupyter notebook*.

Gambar 1 menampilkan tahapan pengembangan sistem yang dilakukan pada penelitian ini terdiri dari pengumpulan data teks berupa sinopsis novel, labeling data, *preprocessing*, feature selection dengan TF-IDF, pelatihan data dengan *Naïve Bayes*, pengujian data dengan *Naïve Bayes*, klasifikasi, dan evaluasi.

3. Hasil Penelitian dan Pembahasan

Pada Gambar 1 diatas merupakan tahapan-tahapan yang akan dilakukan dalam pengembangan sistem diantaranya pengumpulan data teks berupa sinopsis novel, labeling data, *preprocessing*, feature selection dengan TF-IDF, pelatihan data dengan *Naïve Bayes*, pengujian data dengan *Naïve Bayes*, klasifikasi, dan evaluasi.

3.1. Dataset

Tahap awal mengumpulkan dataset, pada peneitian ini dataset yang digunakan adalah data teks berupa sinopsis novel, dataset dibagi menjadi dua yaitu data latih dan data uji, guna dari data latih adalah sebagai data yang digunakan sebagai bahan dari tahap pelatihan dan data uji sebagai bahan dari tahap pengujian yang akan diklasifikasikan oleh sistem. Dataset yang digunakan berjumlah 800 sinopsis, pembagian data dilakukan dengan perbandingan 0.75 dan 0.25 sehingga didapatkan data latih berjumlah 600 dan data uji berjumlah 200.

3.2. Labeling

Pada penelitian ini kelas dibagi menjadi empat diantaranya kelas romantis, komedi, misteri, dan horor, data sejumlah 800 teks sinopsis akan diberikan label sesuai kelasnya dengan pembagian sama rata yaitu 200 teks sinopsis untuk masing-masing kelas.

3.3. Preprocessing

Preprocessing dilakukan untuk mengubah bentuk data tekstual yang tidak terstruktur menjadi data yang terstruktur [8] untuk mempersiapkan data untuk diolah pada tahap selanjutnya, tahapan text preprocessing pada penelitian ini yaitu [9] :

a. Case Folding
Case folding merupakan suatu proses mengubah teks kapital secara keseluruhan menjadi huruf kecil.

b. Filtering
Filtering merupakan suatu proses penyaringan teks dengan menghilangkan spesial karakter dan kata-kata yang dianggap kurang penting (stopwords removal).

c. Stemming
Stemming merupakan suatu proses mengembalikan bentuk dasar dari kata berimbuhan, guna memperkecil ragam kata.

d. Tokenizing
Tokenizing merupakan suatu proses pemisahan sebuah teks yang semulanya adalah kalimat menjadi kata tanggal.

3.4. TF-IDF

TF-IDF adalah sebuah metode pembobotan kata yang digunakan untuk mengekstraksi ciri dari suatu teks, terdapat dua hal dalam perhitungan nilai bobot yaitu term frequency atau TF dan inverse document frequency atau IDF. TF digunakan untuk mencari nilai dari kemunculan kata pada suatu dokumen yang mana semakin banyak suatu kata yang muncul maka semakin tinggi nilai TF. Dan IDF adalah nilai kemunculan dari kata pada keseluruhan dokumen, nilai IDF berbanding terbalik dengan nilai TF, semakin banyak kata yang muncul maka nilai IDF akan semakin kecil [8].

Pada tahap ini TF-IDF dilakukan untuk mendapatkan nilai TF yang berupa frekuensi kemunculan kata dan IDF berupa nilai setiap kata dari keseluruhan dokumen, proses TF dan IDF akan dilakukan pada setiap kelas, yang hasilnya akan dikalikan dan menghasilkan nilai bobot dari kata.

Rumus TF-IDF yang digunakan ditunjukkan pada Persamaan 1.

$$W_{dt} = tf_{dt} x idf_t = tf_{dt} x log(\frac{N}{df_t}) \quad (1)$$

3.5. Naïve Bayes Classifier

Naïve Bayes adalah salah satu metode klasifikasi berbasis numeris dengan pendekatan probabilistic [13] yang berakar pada teorema bayes dimana proses klasifikasi dilakukan dengan

melakukan perhitungan probabilitas dari suatu data [4], *Naïve Bayes* merupakan salah satu metode supervised document classification yang artinya untuk melakukan klasifikasi dibutuhkan data latih [14], *Naïve Bayes* memiliki beberapa kelebihan yaitu sederhana, cepat, dan akurasi yang cukup tinggi. Pada Gambar 1 ditampilkan alur klasifikasi sinopsis dan dapat dilihat terdapat dua proses pada klasifikasi yaitu proses pelatihan dan proses pengujian.

Tahap pelatihan dilakukan guna mendapatkan nilai probabilitas dari setiap kata dari data latih yang akan digunakan sebagai bahan untuk tahap pengujian, kumpulan nilai probabilitas kata pelatihan ini dapat disebut dictionary, didalam tahap pelatihan dilakukan proses untuk mendapatkan probabilitas prior, dan probabilitas kata latih.

Tahap pengujian dilakukan guna mendapatkan nilai probailitas dari data uji yang mana digunakan untuk menentukan hasil klasifikasi.

Rumus *Naïve Bayes* yang digunakan seperti pada Persamaan 1.

$$P_{\frac{c}{term\ dokumen\ d}} = P(c)xP\frac{t_1}{C}xP\frac{t_2}{C}x \ldots P\frac{t_n}{C} \quad (2)$$

Keterangan:
P(c/term dokumen d) = probabilitas suatu dokumen termasuk kategori c.
P(c) = probabilitas dokumen kategori c.
t_n = probabilitas kata pada dokumen ke-n.
P(t_n/c) = probabilitas kata ke-n dari kategori c.

Rumus probabilitas prior kategori c dengan menggunkan Persamaan 3.

$$P_c = \frac{N_c}{N} \quad (3)$$

Keterangan:
N_c = Jumlah kategori c pada seluruh dokumen.
N = Jumlah seluruh dokumen.

Sementara rumus Multinomial yang digunakan dengan pembobotan kata TF-IDF pada Persamaan 4.

$$P = \frac{t_n}{C} = \frac{Wct + 1}{(\sum W'€VW'ct) + B'} \quad (4)$$

Keterangan:
Wct = Nilai pembobotan tfidf atau W dari kata t di kategori c.
$\sum W'€VW'ct$ = Jumlah total W dari keseluruhan kata yang berada di kategori c.
B' = Jumlah W kata unik (nilai idf tidak dikali dengan tf) pada seluruh dokumen.

Klasifikasi dilakukan setelah melewati tahap preprocessing dan perhitungan nilai tiap fitur yang hasilnya akan digunakan pada proses klasifikasi, Tahap-tahap perhitungan pada proses klasifikasi adalah sebagai berikut:
1. Hitung probabilitas prior setiap kelas dengan Persamaan 3.
2. Hitung probabilitas kata ke-n dengan Persamaan 4.
3. Hitung probabilitas dokumen menentukan kelas dengan Persamaan 2.
4. Penentuan kelas dokumen dengan memilih nilai probabilitas tertinggi.

3.6. Hasil Klasifikasi

Hasil klasifikasi yang didapatkan disajikan dalam Tabel 1 *confussion matrix* yang menampilkan informasi mengenai prediksi oleh sistem dari masing-masing kelas.

Tabel 1. Confussion Matrix

Real	Kelas				Total Kelas
	Romantis	Komedi	Misteri	Horor	
Romantis	43	4	1	2	50
Komedi	2	42	2	4	50
Misteri	4	0	33	13	50
Horor	3	1	3	43	50
Prediksi	52	47	39	62	

3.7. Evaluasi

Evaluasi dilakukan untuk menguji apakah penelitian yang dilakukan sudah berjalan sesuai dengan tujuan penelitian atau tidak, Evaluasi pada penelitian ini dilakukan dengan perhitungan *accuracy*, *precission*, dan *recall* dari hasil klasifikasi yang ditampilkan pada confusion matrix [8].

Tabel 2. Nilai Recall dan Precission

Recall	Precission	Kelas
86%	82.69%	Romantis
84%	89.36%	Komedi
66%	84.61%	Misteri
86%	69.35%	Horor

Tabel 2 menunjukkan bahwa nilai recall dan precission dari kelas misteri dan horor berbanding terbalik dikarenakan kemiripan dari teks penyusun antara keduanya sehingga menyebabkan kegagalan sistem meningkat.

4. Kesimpulan

Hasil klasifikasi yang didapatkan dan disajikan pada tabel 1 confussion matrix adalah berdasarkan hasil klasifikasi oleh sistem, dengan hasil accuracy 80.5%. Terjadinya error atau kegagalan dalam klasifikasi disebabkan adanya kemiripan antara kata-kata penyusun suatu kelas dengan kelas lainnya seperti halnya penulis mengamati susunan kata dari setiap data yang digunakan sebagai contoh pada kelas misteri dan horor dengan nilai *recall* dan *precission* yang berbanding terbalik disebabkan oleh kata-kata yang menyusun data teks kelas misteri dan horor sangat mirip yang menyebabkan klasifikasi gagal, dan jumlah dataset yang digunakan juga dapat mempengaruhi hasil klasifikasi, semakin banyak jumlah data yang digunakan memungkinkan untuk meningkatkan hasil klasifikasi.

Referensi

[1] T. Jatiningsih, K. Saddhono, and B. Waluyo, “Analisis Karakter Tokoh dan Nilai Pendidikan dalam Novel Ayahku (Bukan) Pembohong Karya Tere Liye Serta Kesesuaiannya sebagai Materi Pembelajaran Bahasa Indonesia di SMA (Tinjauan Psikologi Sastra),” pp. 8–55, 2015.
[2] M. C. Kustanti, “Tema Dan Pesan Dalam Fungsi Media Pada Novel Laskar Pelangi Karya Andrea Hirata (Analisis Wacana Pragmatik),” vol. 1, no. 2, pp. 186–195, 2016.
[3] D. N. Chandra, G. Indrawan, and I. N. Sukajaya, “Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram,” vol. 10, no. 1, pp. 11–19, 2016.
[4] I. A. Setiawan, T. H. P, and D. Nursantika, “Klasifikasi Artikel Berita Menggunakan Metode Text,” pp. 1–6, 2017.
[5] A. P. Wijaya, H. A. Santoso, J. T. Informatika, U. Dian, and N. Semarang, “Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government,” vol. 1, no. 1, pp. 48–55, 2016.
[6] O. Somantri and M. Khambali, “Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika,” J. Nas. Tek. Elektro dan Teknol. Inf., vol. 6, no. 3, 2017.
[7] Y. D. Pramudita, S. S. Putro, and N. Makhmud, “Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer,” J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 3, p. 269, 2018.
[8] A. Rahman, “Online News Classification Using Multinomial Naive Bayes,” vol. 6, no. 1, 2017.
[9] R. Saptono et al., “Text Classification Using Naive Bayes Updateable,” vol. 13, no. 02, pp. 123–133, 2016.

[10] V. Chandani, F. I. Komputer, and U. D. Nuswantoro, “Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film,” J. Intell. Syst., vol. 1, no. 1, pp. 56–60, 2015.
[11] H. Februariyanti, “Klasifikasi Dokumen Berita Teks Bahasa Indonesia menggunakan Ontologi,” vol. 17, no. 1, pp. 14–23, 2012.
[12] R. Saptono et al., “Text Classification Using Naive Bayes Updateable,” vol. 13, no. 02, pp. 123–133, 2016.
[13] A. Hamzah, “Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis,” Pros. Semin. Nas. Apl. Sains Teknol. Periode III, no. 2011, pp. 269–277, 2012.
[14] R. Efendi and R. F. Malik, “Klasifikasi dokumen berbahasa indonesia menggunakan naive bayes classifier,” vol. I, no. I, pp. 7–13, 2012.