Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan *Naïve Bayes* (NB)

Oman Somantri[#1]

[#]*Program Studi Teknik Informatika, Politeknik Harapan Bersama Tegal*
Jln. Mataram No.09 Pesurungan Lor Kota Tegal, Indonesia
[1]oman.somantri@poltektegal.ac.id

Abstract*— Determination of the category of a short story requires a slightly long process, in other way we must read a whole or at least a half of the contents of the short story to know the entire contents from the beginning to the end. These constraints require a solution to overcome by using Naïve Bayes algorithm (NB) to serve as the solution of the existing problems. Naïve Bayes, used as a model, resulted with accuracy of 78.59%. Evaluation was conducted by comparing the level of accuracy produced with other models of Support Vector Machine (SVM). The result of the research show that level of accuracy NB greater than Support Vector Machine (SVM) with accuracy level 64,36%. Based on the results of research conducted can be concluded that Naïve Bayes has a higher level of accuracy than the Support Vector Machine (SVM) for the short story category classification.

***Keywords*— Naïve Bayes, Support Vector Machine, short story, Model, accuracy**

Abstrak*— Penentuan kategori sebuah cerita pendek memerlukan sebuah proses yang lama. Kita harus membaca secara keseluruhan atau, minimal, setengah dari isi dari cerpen tersebut. Untuk mengetahui seluruh isi konten dari suatu cerpen adalah dengan membaca isi cerpen, mulai dari awal sampai akhir. Kendala ini memerlukan sebuah solusi untuk mengatasinya. Pada penelitian ini diusulkan sebuah model dengan menggunakan algoritme *Naïve bayes* (NB) untuk dijadikan sebagai solusi dari permasalahan yang ada. *Naïve Bayes* digunakan sebagai model dengan tingkat akurasi sebesar 78,59%. Evaluasi dilakukan dengan membandingkan tingkat akurasi yang dihasilkan dengan model lain, yaitu *Support Vector Machine* (SVM). Hasil penelitian memperlihatkan bahwa tingkat akurasi NB lebih besar dibandingkan dengan SVM, yaitu dengan tingkat akurasi 64,36%. Oleh karena itu, didapatkan kesimpulan pada penelitian ini bahwa *Naïve Bayes* mempunyai tingkat akurasi lebih tinggi dibandingkan dengan *Support Vector Machine* untuk klasifikasi kategori cerpen.

***Kata Kunci*— Naïve Bayes, Support Vector Machine, cerpen, model, akurasi**

I. PENDAHULUAN

Sebagai salah satu bagian dari kebudayaan Indonesia, cerpen merupakan karya sastra yang paling banyak diminati oleh banyak orang. Sebuah cerpen akan dapat diminati orang apabila isi dari cerpen tersebut menarik dan dapat membawa orang yang membacanya hanyut ke dalam isi dari cerita tersebut. Berbagai macam latar belakang pembaca cerpen saat ini, mulai dari remaja, anak-anak, dewasa, maupun para orang tua. Perbedaaan latar belakang inilah tentunya menjadikan sebuah cerpen memiliki segmentasi yang berbeda. Cerpen memiliki banyak kategori sesuai dengan isinya, seperti kategori cerpen anak, dongeng, fiksi, pendidikan, dewasa, romantis, dan sebagainya. Cerpen adalah cerita fiktif yang belum pasti kebenarannya. Ceritanya relatif pendek dan cerpen bukanlah suatu analisis argumentatif [1].

Untuk dapat menentukan sebuah cerpen masuk ke dalam kategori cerpen tertentu bukanlah hal yang mudah. Sudah tentu orang harus membaca keseluruhan atau minimal sebagian isi cerpen tersebut kemudian barulah dapat mengetahui cerpen tersebut masuk ke dalam kategori apa. Hal inilah yang menjadi kesulitan dalam menentukan sebuah cerpen masuk ke dalam kategori tertentu, sedangkan terkadang banyak orang yang tidak bisa membaca terlebih dahulu isi dari cerpen tersebut. Permasalahan kadang terjadi pada para orang tua yang ingin memberikan sebuah cerpen kepada anaknya. Karena belum mengetahui cerpen tersebut termasuk ke dalam kategori apa, maka ada kemungkinan isi cerpen tidak sesuai dengan umur usia anak. Hal ini merupakan salah satu contoh kasus yang sering terjadi. Berdasarkan permasalahan tersebut, maka perlu sebuah solusi untuk mengatasinya. Solusi ini dapat dijadikan sebagai pendukung keputusan dalam menentukan kategori sebuah cerpen.

Dalam bidang komputerisasi yang termasuk ke dalam *machine learning*, *Naïve Bayes* dan *Support Vector Machine* (SVM) merupakan metode yang digunakan untuk klasifikasi teks dalam *text mining*. Sebagai salah satu metode komputasi yang efisien dan mempunyai *performance preditictive* yang baik, *Naïve Bayes* merupakan salah satu metode klasifikasi teks yang popular [2]. *Naïve Bayes* merupakan algoritme yang sering digunakan dalam pengkategorian teks, di mana konsep dasarnya adalah menggabungkan probabilitas kata-kata dan kategori sebuah dokumen [3][8].

Penelitian terkait dengan klasifikasi teks menggunakan *Naïve Bayes* sudah dilakukan oleh para peneliti sebelumnya. S. A. Nurul (2016) melakukan penelitian untuk membandingkan *Naïve Bayes* dan *Support Vector Machine* (SVM) untuk klasifikasi emosi pada teks bahasa Indonesia [4]. A. Hamzah (2012) melakukan penelitian klasifikasi teks dengan *Naïve Bayes Classifier* (NBC) untuk pengelompokan teks berita dan abstrak akademis [5]. Selanjutnya, N. A. S. Winarsih dan C. Supriyanto (2016) meneliti untuk mengevaluasi metode klasifikasi deteksi emosi pada teks Indonesia [6]. Sedikit berbeda dengan yang dilakukan oleh N. Jamal, dkk. (2012),

meneliti klasifikasi puisi dengan menggunakan *Support Vector Machine* (SVM) [7].

Perbedaan penelitian ini dengan penelitian-penelitian yang telah dilakukan sebelumnya adalah pada proses prapemrosesan data dan metode yang digunakan untuk klasifikasi kategori cerpen. Penelitian ini mengusulkan *Naïve Bayes* sebagai metode yang diterapkan untuk pengklasifikasian jenis kategori cerita pendek. Hasil akhirnya adalah sebuah rekomendasi model yang tepat untuk menghasilkan tingkat akurasi terbaik untuk klasifikasi kategori cerita pendek.

II. KONTEN UTAMA

A. Naive Bayes

Naive Bayes merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana, metode ini mengklasifikasikan data berdasarkan probabilitas *P* atribut *x* dari setiap kelas *y* data [8]. *Naive Bayes* adalah metode yang digunakan dalam statistika untuk menghitung peluang dari suatu hipótesis, *Naive Bayes* menghitung peluang suatu kelas berdasarkan pada atribut yang dimiliki dan menentukan kelas yang memiliki propabilitas paling tinggi. *Naive Bayes* mengklasifikasikan kelas berdasarkan pada probabilitas sederhana dengan mengasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. Pada model probabilitas setiap kelas k dan jumlah atribut *a* yang dapat dituliskan seperti persamaan dibawah ini:

$$P = \left(y_1 \mid x_1, x_2, \ldots x_a \right) \tag{1}$$

Perhitungan *Naive Bayes*, yaitu probabilitas dari kemunculan dokumen X_a pada kategori kelas Y_k $P(x_a|y_k)$ dikali dengan probabilitas katgori kelas $P(y_k)$. Dari hasil kali tersebut kemudian dilakukan pembagian terhadap probabilitas kemunculan dokumen $P(x_a)$, sehingga didapatkan rumus perhitungan *Naive Bayes* dituliskan pada persamaan:

$$P\left(y_k \mid x_a \right) = \frac{P(y_k)P(x_a \mid y_k)}{P(x_a)} \tag{2}$$

Proses pemilihan kelas yang optimal dilakukan berdasarkan nilai peluang terbesar dari setiap probabilitas kelas yang ada. Didapatkan rumus untuk memilih nilai terbesar pada persamaan berikut:

$$y(x_i) = \arg\max P(y) \prod_{i=1}^{a} P(x_1 \mid y) \tag{3}$$

Pembobotan suatu atribut kelas dapat meningkatkan pengaruh prediksi. Dengan memperhitungkan bobot atribut terhadap kelas, maka yang menjadi dasar ketepatan klasifikasi adalah bukan hanya probabilitas, melainkan juga pada bobot setiap atribut kelas.

B. Metodologi Penelitian

1) Dataset Penelitian

Dataset yang digunakan dalam penelitian ini diambil dari *www.cerpenmu.com*. Data *online* ini adalah berupa teks yang berbentuk cerita pendek yang sudah ditentukan kategorinya, yaitu kategori cerpen anak dan kategori cerpen dongeng. *Dataset* adalah data yang dibuat antara tahun 2015 sampai dengan 2016 dengan jumlah data sebanyak 121 cerpen.

2) Prapemrosesan Data

Sebelum *dataset* dimasukkan ke dalam model yang diusulkan, terlebih dahulu dilakukan prapemrosesan data. Pada tahap ini dilakukan beberapa hal, di antaranya adalah *tokenized*, *transform cases*, *filter tokens*, *filter stopword* dan *Stem* [9].

- *Tokenized* merupakan proses untuk memisah-misahkan kata. Hasil dari pemisahan tersebut dinamakan token.
- *Transform case* merupakan proses untuk mengubah bentuk kata-kata. Pada proses ini karakter dijadikan huruf kecil atau *lower case* semua.
- *Filter tokens* merupakan proses pengambilan kata-kata yang penting dari token yang sudah dihasilkan berdasarkan jumlah karakter. Pada proses ini parameter yang digunakan adalah *min chars* = 3, dan *max chars* = 25.
- *Filter stopword* merupakan proses menghilangkan kata-kata yang sering muncul, namun tidak memiliki pengaruh apapun dalam ekstraksi klasifikasi teks. Pada proses ini kata yang termasuk adalah penunjuk waktu, kata tanya, dan kata sambung.
- *Stem* merupakan proses pengubahan bentuk kata menjadi kata dasar. Metode ini merupakan proses pengubahan bentuk kata menjadi kata dasar yang menyesuaikan dengan struktur yang digunakan dalam proses *stemming*.

3) Model Yang Diusulkan

Pada penelitian ini model yang diusulkan adalah *Naïve Bayes* sebagai algoritme pembelajaran (Gambar 1). Untuk mendapatkan tingkat akurasi yang sesuai, proses validasi dilakukan dengan menggunakan *K-Fold Cross Validation* dengan harapan hasil validasi ekperiment dapat menghasilkan hasil yang terbaik. Model yang diusulkan kemudian dievaluasi dengan cara membandingkan hasil tingkat akurasi yang didapatkan dengan model yang lain yaitu *Support Vector Machine* (SVM).

4) Validasi

Evaluasi model pada tahap ini menggunakan evaluasi *matrixs confusion* [10], seperti pada Tabel I.

Untuk menghitung tingkat akurasi digunakan persamaan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

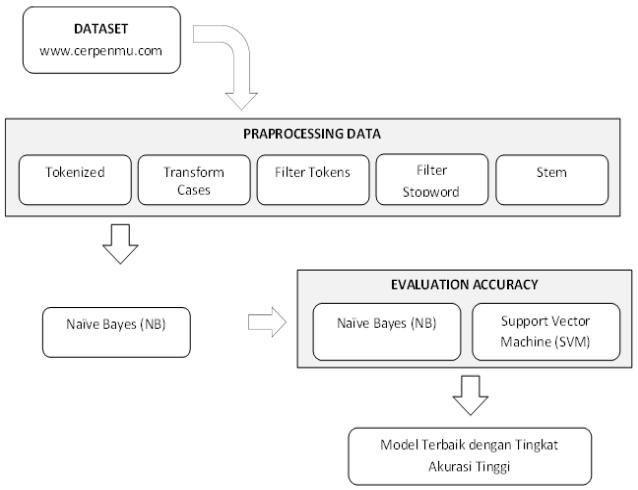
Keterangan:
- *True Positive* (*TP*)
- *False Positive* (*FP*)
- *False Negative* (*FN*)
- *True Negative* (*TN*)

III. HASIL DAN PEMBAHASAN

Penelitian menggunakan *tools* Rapid Miner 5.3 untuk analisis data. Komputer dengan spesifikasi CPU Intel Core i5 2,67 GHz, memori RAM 4 GB, sistem operasi Windows 7 profesional SP1 32-bit.

A. Hasil Ekperimen Naïve Bayes

Pada eksperimen terhadap model yang digunakan dilakukan dengan mengggunakan cerpen yang sudah ditetapkan sesuai dengan kategori cerpen, yaitu cerpen anak dan cerpen dongeng dengan jumlah *dataset* sebanyak 121 cerpen.



Gambar 1 Model yang diusulkan

TABEL I
CONFUSION MATRIX

	Hasil Prediksi	
	Positive	***Negative***
Positive	*True Positive*	*False Positive*
Negative	*False Negative*	*True Negative*

Berikut ini adalah contoh konten isi cerpen yang digunakan sebagai salah satu *dataset* penelitian, yaitu cerpen kategori dongeng.

```
Hai namaku Sabila Salwa Putri Wahyuhadi cukup
dipanggil Salwa

"Salwa bangun! cepat bangun terus shalat subuh"
bentak bunda. Aku bangun dan shalat subuh sehabis
shalat subuh aku tidur.
Dan paginya aku siap siap mandi, makan lalu
berangkat sekolah. Ini hari pertamaku masuk
sekolah kelas 3. aku masuk dan perkenalan setelah
itu pelajaran lalu istirahat.

Pas istirahat, aku kenalan sama teman-teman yang
lain yang paling aku sukai adalah aca, gadis
kecil memakai kerudung dengan pipi yang
menggelembung.

Sudah 2 minggu aku sekolah di sana rasanya sangat
menyenangkan apalagi aca, dia sudah menjadi
sahabatku sejak 6 hari yang lalu.

"Heh ca kita main bareng yuk!" ajakku pada aca.
Tapi, dia tidak menjawab pertanyaanku dan murung
dia sangat sedih. Aku juga tidak tahu mengapa dia
sedih. Dan aca pun langsung pergi. Aku
mengikutinya dan ternyata dia mengkhianatiku. Dia
bersahabat dengan silvia ayu musuh terbesarku.

Aku sedih dan langsung pulang ke rumah dan
akhirnya aku memutuskan untuk tidak bersahabat
lagi dengan aca. Walaupun sangat berat.
```

Hasil ekperimen terhadap *dataset* yang sudah didapatkan dengan menggunakan model *Naive Bayes* diperlihatkan pada Tabel II. Pada Tabel II diperlihatkan bahwa tingkat akurasi dari hasil ekperimen menunjukan bahwa *Naïve Bayes* menghasilkan tingkat akurasi sebesar 78,59%. Terjadinya kesalahan dalam klasifikasi kategori cerpen mengakibatkan tingkat akurasi yang dihasilkan menjadi kecil. Hal ini disebabkan oleh model yang diusulkan masih belum sesuai dengan yang diinginkan. Hal ini dapat juga terjadi akibat dari berbagai aspek lain, seperti perbedaaan model yang digunakan, proses prapemrosesan data, *setting* parameter model yang digunakan, dan aspek lain yang dianggap mempengaruhi tingkat akurasi klasifikasi yang dihasilkan.

Gambar 2 memperlihatkan grafik hasil eksperimen yang dengan menggunakan model algoritme *Naïve Bayes*.

TABEL II
HASIL EKPERIMEN MODEL NAIVE BAYES

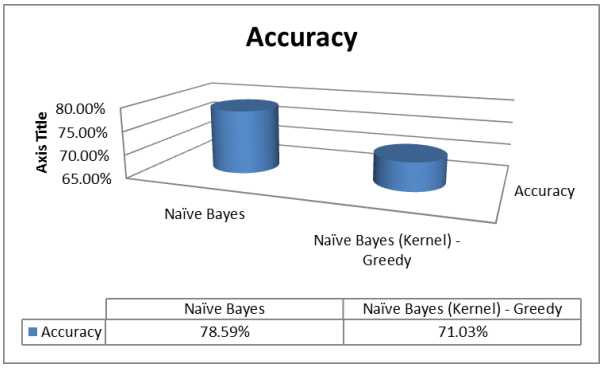
No	**Hasil Ekperimen**	
	Model	**Accuracy**
1	Naïve Bayes	78.59%
2	Naïve Bayes (Kernel) - Greedy	71.03%

B. Hasil Ekperimen Support Vector Machine

Pada model *Support Vector Machine* (SVM) ekperimen yang dilakukan menghasilkan seperti yang diperlihatkan pada Tabel III. Pada Tabel III terlihat bahwa tingkat akurasi yang paling tinggi pada SVM adalah dengan menggunakan *kernel type dot*, yaitu sebesar 64,36%. Apabila dibuatkan pada grafik, maka akan tampak seperti pada Gambar 3.

C. Evaluasi Model

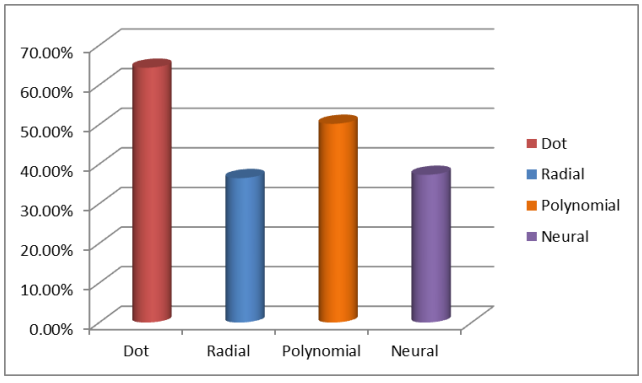
Hasil ekperimen yang diperoleh dengan model *Naive Bayes* dan *Support Vector Machine* diperlihatkan pada Tabel IV. Terlihat bahwa *Naive bayes* menghasilkan tingkat akurasi sebesar 78,59%, sedangkan *Support Vector Machine* (SVM) menghasilkan tingkat akurasi sebesar 64,34%.



Gambar 2 Hasil eksperimen Naïve Bayes

TABEL III
HASIL EKPERIMEN MODEL SUPPORT VECTOR MACHINE (SVM)

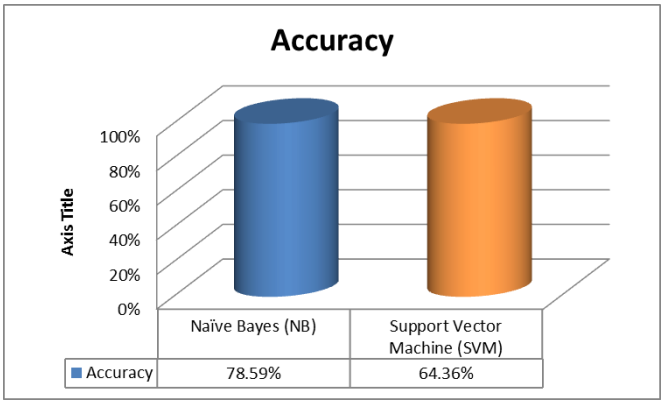
No.	Hasil Ekperimen SVM	
	Tipe Kernel	Akurasi
1	Dot	64,36%
2	Radial	36,47%
3	Polynomial	50,19%
4	Neural	37,31%



Gambar. 3 Hasil eksperimen *Support Vector Machine* (SVM)

TABEL IV
HASIL EKPERIMEN NAIVE BAYES & SVM

No.	Hasil Ekperimen Optimasi	
	Model	Akurasi
1	*Naïve Bayes* (NB)	78,59%
3	*Support Vector Machine* (SVM)	64,36%



Gambar. 4 Hasil eksperimen Naïve Bayes dan SVM

IV. KESIMPULAN

Hasil eksperimen yang dilakukan terhadap model yang diusulkan, yaitu *Naïve Bayes*, menghasilkan tingkat akurasi sebesar 78,59% lebih besar dibandingkan dengan model lain, yaitu *Support Vector Machine* (SVM). Dari hasil penelitian dapat disimpulkan bahwa *Naïve Bayes* memiliki tingkat akurasi yang lebih baik dibandingkan dengan SVM untuk pengklasifikasian kategori cerita pendek. Meskipun demikian hasil yang didapatkan belumlah sempurna, sehingga perlu adanya sebuah optimasi dalam model tersebut. Saran untuk penelitian selanjutnya adalah dilakukannya ekperimen dengan model-model lain dan *setting* parameter yang lebih tepat, sehingga didapatkan tingkat akurasi yang lebih baik.

DAFTAR REFERENSI

[1] J. Sumardjo dan Saini K.M. *Apresiasi Kesusastraan*. Jakarta: PT Gramedia, 1998.

[2] J. Chen, H. Huang, S. Tian, & Y. Qu. "Feature selection for text classification with Naïve Bayes." *Expert Systems with Applications*, 36, pp. 5432–5435, 2009.

[3] Zhang and F. Gao. "an Improvement to NB for Text Classification." *Procedia Engineering*, 15, pp. 2160–2164, 2011.

[4] S. A. Nurul. "Perbandingan Metode Naïve Bayes dan Support Vector Machine Untuk Klasifikasi Emosi pada Teks Bahasa Indonesia". Skripsi, Fakultas Ilmu Komputer, 2016.

[5] A. Hamzah. "Klasifikasi teks dengan naïve bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis," dalam *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*, 2012.

[6] N. A. S. Winarsih dan C. Supriyanto, "Evaluation of classification methods for Indonesian text emotion detection," in *International Seminar on Application for Technology of Information and Communication (ISemantic)*, 2016, pp. 130-133.

[7] N. Jamal, M. Mohd, dan S. A. Noah. "Poetry classification using support vector machines." *Journal of Computer Science*, *8*(9), pp.1441, 2012.

[8] A. McCallum dan K. Nigam, "a Comparison of event models for Naive Bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[9] C. C. Aggarwal and C. Zhai, (Eds.). *Mining text data*. Springer Science & Business Media. 2012.

[10] J. Davis and M. Goadrich. "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd Iinternational Conference on Machine Learning*. ACM, June 2006.

Oman Somantri, lahir pada tahun 1985 Sumedang, menerima gelar Sarjana Komputer (S.Kom) dari STMIK Sumedang jurusan Teknik Informatika pada tahun 2011 dan gelar Magister Komputer (M.Kom) dari Universitas Dian Nuswantoro (Udinus) Jurusan Teknik Informatika pada tahun 2015. Saat ini mengajar sebagai dosen di Politeknik Harapan Bersama Tegal. Minat penelitian adalah *Intelligent System*, *Machine Learning*, *Data Mining*, dan *Text Mining*.

Halaman kosong