

# NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINES UNTUK SENTIMENT ANALYSIS

**Ni Wayan Sumartini Saraswati**

Prodi S1 Teknik Informatika, STMIK Denpasar

Jl. Tukad Balian No 15, Denpasar, 80226

Telp : (0361) 8987039, Fax : -

E-mail : [geck\\_nik@yahoo.co.id](mailto:geck_nik@yahoo.co.id)

---

## **Abstrak**

*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.*

*Pada penelitian ini dibahas klasifikasi opini sebagai opini positif dan opini negatif pada data berbahasa Inggris dan data berbahasa Indonesia menggunakan metode Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM). Baik metode NBC maupun metode SVM memberikan unjuk kerja yang baik dalam sentiment analysis pengklasifikasian opini berbahasa Inggris dan berbahasa Indonesia pada penelitian ini. Hasil percobaan menunjukkan bahwa metode SVM memberikan unjuk kerja yang lebih baik daripada metode NBC untuk mengklasifikasikan opini berbahasa Inggris dan opini positif berbahasa Indonesia. Sedangkan NBC memberikan unjuk kerja yang lebih baik dalam mengklasifikasikan data uji opini negatif berbahasa Indonesia.*

**Kata kunci :** *text mining, sentiment analysis, opinion mining, NBC, SVM*

## **Abstract**

*Text mining refers to the process of deriving high-quality information from text. High-quality information is typically derived through the dividing of patterns and trends through means such as statistical pattern learning. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.*

*This research discussed the opinions classification as positive opinions and negative opinions on English and Indonesian language data using the Naive Bayes Classifier (NBC) and Support Vector Machine (SVM). In this study, both NBC and SVM method gives a good sentiment analysis performance in classifying opinion for English and Indonesia language. The experimental results showed that SVM method gives better performance than NBC method for classifying English opinions. Whether NBC method gives better performance for classifying negative opinions experimental data on Indonesian language.*

**Keywords :** *text mining, sentiment analysis, opinion mining, NBC, SVM*

## **1. PENDAHULUAN**

Web adalah tempat yang baik bagi orang - orang untuk mengekspresikan pendapat mereka, pada berbagai topik. Bahkan pemberi opini secara profesional, seperti reviewer film, memiliki blog dimana publik dapat mengomentari dan merespon apa yang mereka pikirkan. Kemampuan untuk mengekstrak pendapat tersebut dari baris-baris teks dapat menjadi sangat berguna, dan ini adalah area studi yang banyak dikaji, tidak diragukan karena kemungkinan nilai komersialnya (Ian Barber, 2010). Dikarenakan kebanyakan informasi (perkiraan umum mengatakan lebih dari 80%) saat ini disimpan sebagai teks, text mining diyakini memiliki potensi nilai komersial tinggi (Clara Bridge, 2011).

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 (yaitu, pembelajaran hubungan antara entitas bernama).

Sentiment Analysis atau opinion mining adalah studi komputasional dari opini-opini orang, appraisal dan emosi melalui entitas, event dan atribut yang dimiliki (Biu, L. 2010). Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/ tingkat aspek - apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas / aspek bersifat positif, negatif atau netral (Dehaff, M., 2010).

Pada hasil eksperimen (Wulandini, F. & Nugroho, A. N. 2009) untuk kategorisasi teks berbahasa Indonesia didapatkan bahwa SVM menunjukkan performansi yang sedikit lebih baik dengan akurasi 92,5% dibandingkan metode NBC dengan akurasi 90% padahal metode NBC adalah metode yang jauh lebih konvensional dan lebih sederhana. Sehingga pada penelitian ini ingin diketahui metode yang mana memiliki performansi yang lebih baik untuk diimplementasikan dalam sentiment analysis opini berbahasa Inggris dan berbahasa Indonesia.

## 2. KAJIAN PUSTAKA

Telah dilakukan penelitian dalam bidang *text mining* oleh peneliti di Indonesia maupun di luar negeri. Berikut adalah beberapa penelitian relevan beserta hasilnya. Sub bab selanjutnya akan membahas mengenai klasifikasi dokumen menggunakan NBC dan SVM.

### 2.1 State of Art Review

Ian Barber dalam Bayesian Opinion Mining, 2010 telah melakukan eksperimen untuk data review film dan menghasilkan tingkat akurasi 80% menggunakan metode NBC. Tingkat akurasi diujikan untuk ketepatan menentukan klas opini dengan 5000 record opini negatif dan 5000 record opini positif sebagai data latih. Data uji adalah 333 opini negatif. Ian Barber belum melakukan eksperimen untuk data uji sentiment positif dan pengaruh variasi jumlah data latih dan data uji terhadap performansi metode NBC. Ian Barber juga belum melakukan eksperimen menggunakan data tersebut untuk metode klasifikasi teks yang lain misalkan SVM.

Yudi Wibisono dalam Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier, 2006 telah melakukan klasifikasi pada 582 dokumen berbahasa Indonesia menggunakan metode NBC dan memperoleh hasil eksperimen seperti pada tabel 1 :

Tabel 1 Hasil Eksperimen Yudi Wibisono pada klasifikasi dokumen berbahasa Indonesia

Jumlah Dokumen Contoh	Jumlah Dokumen Uji Coba	Akurasi (%)
524 (90%)	58 (10%)	89,47
407 (70%)	175 (30%)	90,23
291 (50%)	291 (50%)	86,90
175 (30%)	407 (70%)	85,47
58 (10%)	524 (90%)	68,64

Dari tabel 1 terlihat bahwa nilai akurasi NBC tinggi, terutama jika dokumen contoh yang digunakan besar ( $\geq 400$  dokumen). Hal yang menarik adalah akurasi tidak menunjukkan peningkatan yang signifikan walaupun dokumen contoh telah meningkat banyak dari 70% menjadi 90% serta akurasi masih relative tinggi walaupun dokumen contoh secara ekstrim dikurangi hanya 58 dokumen (10%).

Fatimah Wulandini dan Anto Satriyo Nugroho dalam (Text Classification Using Support Vector Machine for Webmining Based Spation Temporal Analysis of the Spread of Tropical Diseases, 2009) mendapatkan hasil bahwa metode SVM menunjukkan hasil paling baik pada kategorisasi teks berbahasa Indonesia. Eksperimen dilakukan pada 3713 features dan 360 instances. Data tersebut dibagi menjadi 120 instances sebagai data uji dan 240 instances sebagai data latih. Hasilnya seperti ditunjukkan oleh tabel 2 :

Tabel 2 Hasil Eksperimen Fatimah Wulandini dan Anto Satriyo Nugroho

Metode	Akurasi
SVM	92,5 %
K- Nearest Neighbor	29,17 %
Naïve Bayes Classifier	90 %
C45	77,5 %

Tabel 2 tersebut menunjukkan performansi yang tidak berbeda jauh antara metode SVM dan NBC walaupun metode NBC adalah metode yang lebih konvensional dan lebih sederhana.

Fabrice Colas & Pavel Brazdil dalam penelitiannya berjudul *Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks* mendapatkan bahwa metode NBC memiliki performansi yang lebih baik dibandingkan KNN dan SVM untuk menyelesaikan binary classification pada dokumen berbahasa Inggris. Hasil penelitiannya juga menyebutkan waktu komputasi yang jauh lebih pendek oleh metode NBC dan KNN. Waktu komputasi SVM berkembang secara kuadratik seiring dengan perkembangan jumlah data latih.

Namun penelitian oleh Jason D. M. Rennie & Ryan Rifkin yang berjudul *Improving Multiclass Text Classification with the Support Vector Machine* menunjukkan hasil bahwa SVM menghasilkan performansi yang lebih baik dalam menyelesaikan klasifikasi teks multi kelas dibandingkan metode NBC. Hal ini sesuai dengan pernyataan Fabrice Colas & Pavel Brazdil bahwa SVM unggul dalam klasifikasi multiclass.

## 2.2 Naïve Bayes Classifier untuk Klasifikasi Dokumen

Model probabilitas untuk classifier adalah model kondisional

$$p(C|F_1, \dots, F_n) \quad (1)$$

terhadap variabel kelas dependen  $C$  dengan sejumlah kecil hasil atau kelas, tergantung pada beberapa variabel fitur  $F_1$  sampai  $F_n$ . Masalahnya adalah bahwa jika jumlah fitur  $n$  besar atau bila fitur bisa mengambil sejumlah besar nilai, maka membuat sebuah model pada tabel probabilitas adalah tidak mungkin. Oleh karena itu kita mereformulasi model untuk membuatnya lebih fleksibel.

Menggunakan teorema Bayes, kita menulis

$$p(C|F_1, \dots, F_n) = (p(C)p(F_1, \dots, F_n|C)) / (p(F_1, \dots, F_n)) \quad (2)$$

Dalam bahasa Inggris persamaan di atas dapat ditulis sebagai

$$\text{posterior} = (\text{prior} \times \text{likelihood}) / \text{evidence} \quad (3)$$

Dalam prakteknya kita hanya tertarik pada pembilang dari persamaan tersebut, karena penyebut tidak tergantung pada  $C$  dan nilai-nilai fitur  $F_i$  diberikan, sehingga penyebut secara efektif konstan.

Berikut ini adalah sebuah contoh dari NBC untuk permasalahan klasifikasi dokumen. Masalah mengklasifikasikan dokumen adalah berdasarkan konten, misalnya spam dan non-spam e-mail. Bayangkan bahwa dokumen yang diambil dari beberapa kelas dokumen yang dapat dimodelkan sebagai set kata-kata dimana probabilitas (independen) bahwa kata ke- $i$  dari suatu dokumen tertentu terjadi dalam dokumen dari kelas  $C$  dapat ditulis sebagai

$$p(w_i|C) \quad (4)$$

Untuk perlakuan ini, kita menyederhanakan hal-hal lebih lanjut dengan mengasumsikan bahwa kata-kata secara acak terdistribusi dalam dokumen - yaitu, kata-kata tidak tergantung pada panjang dokumen, posisi dalam dokumen, dengan hubungannya dengan kata lain, atau dokumen-konteks yang lain.

Maka probabilitas suatu dokumen  $D$ , kelas  $C$ , adalah

$$p(D|C) = \prod p(w_i | C) \quad (5)$$

Pertanyaan yang ingin dijawab adalah: "berapa probabilitas bahwa dokumen  $D$  adalah milik kelas  $C$ ?" Dengan kata lain, berapa  $p(C|D)$ ?

Sekarang menurut definisi

$$p(D|C) = \frac{p(D \cap C)}{p(C)} \quad (6)$$

dan

$$p(C|D) = \frac{p(D \cap C)}{p(D)} \quad (7)$$

Teorema Bayes memanipulasi ini ke dalam pernyataan dari probabilitas dalam bentuk *likelihood*.

$$p(C|D) = \frac{p(C)}{p(D)} p(D|C) \quad (8)$$

Asumsikan untuk saat ini bahwa hanya ada dua kelas yang saling eksklusif,  $S$  dan  $\neg S$  (misalnya spam dan bukan spam). Menggunakan hasil Bayesian di atas, kita bisa menulis

$$p(S/D) = \frac{p(S)}{p(D)} \prod_i p(w_i | S) \quad (9)$$

$$p(\neg S/D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i | \neg S) \quad (10)$$

Membagi satu dengan yang lain memberikan:

$$\frac{p(S/D)}{p(\neg S/D)} = \frac{p(S) \prod_i p(w_i | S)}{p(\neg S) \prod_i p(w_i | \neg S)} \quad (11)$$

Yang bisa difaktorisasi ulang sebagai :

$$\frac{p(S/D)}{p(\neg S/D)} = \frac{p(S)}{p(\neg S)} \prod_i \frac{p(w_i | S)}{p(w_i | \neg S)} \quad (12)$$

Dengan demikian, rasio probabilitas  $p(S/D) / p(\neg S/D)$  dapat dinyatakan dalam serangkaian rasio kemungkinan.  $p$  probabilitas aktual ( $S/D$ ) dapat dengan mudah dihitung dari  $\log(p(S/D) / p(\neg S/D))$  berdasarkan pengamatan dimana  $p(S/D) + p(\neg S/D) = 1$ .

Mengambil logaritma dari semua rasio, kita memiliki:

$$\ln \frac{p(S/D)}{p(\neg S/D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i | S)}{p(w_i | \neg S)} \quad (13)$$

Akhirnya, dokumen dapat diklasifikasikan sebagai berikut. Dikategorikan spam jika  $p(S/D) > p(\neg S/D)$  (yaitu  $\ln \frac{p(S/D)}{p(\neg S/D)} > 0$ ), jika tidak memenuhi maka bukan spam.

### 2.3 Support Vector Machines untuk pengolahan teks

Proses klasifikasi menggunakan SVM dimulai dengan mengubah text menjadi data vector. Vector dalam penelitian ini memiliki dua komponen yaitu dimensi (*word id*) dan bobot. Bobot ini sering dikombinasikan ke dalam sebuah nilai tf-idf, secara sederhana dengan mengalikan mereka bersama-sama. Ada banyak variasi pada gagasan dasar tf-idf, tetapi implementasi langsung akan terlihat seperti:

```
<?php
    $tfidf = $term_frequency * // tf
    log( $total_document_count / $documents_with_term, 2);/idf
?>
```

idf adalah jumlah total dokumen atas hitungan yang berisi istilah tersebut. Jadi, jika ada 50 dokumen dalam koleksi, dan dua di antaranya terdapat istilah yang menjadi query, IDF akan menjadi  $50 / 2 = 25$ . Untuk menjadi akurat, kita harus memasukkan query dalam perhitungan IDF, jadi jika dalam koleksi ada 50 dokumen, dan 2 berisi istilah dari query, perhitungan yang sebenarnya akan  $(50 + 1) / (2 + 1) = 51 / 3$ .

Diambil log dari IDF untuk memberikan beberapa penghalusan. Jika sebuah istilah A direpresentasikan dalam  $x$  buah dokumen, dan istilah B sejumlah  $2x$  kali, maka istilah A adalah istilah yang lebih spesifik yang harus memberikan hasil yang lebih baik, tetapi belum tentu dua kali lebih baik. Kelembutan dari log adalah pemecahan perbedaan-perbedaan ini. Dokumen dapat dinyatakan sebagai list dari term. Kita kemudian menormalisasi tiap komponen dengan panjang dari vector sehingga bobot tersebut dinyatakan dalam 1 unit panjang.

Contoh format data input untuk klasifikasi SVM dalam penelitian ini adalah +1 1:0.049 45:0.0294. Dengan masukan yang pertama +1 atau -1 menyatakan dua kelas (atau 0 untuk data yang akan diklasifikasi). Angka kedua menyatakan dimensi (*row id*) dan angka ketiga (setelah karakter “:”) menyatakan bobot dari term tersebut, tiap term dalam sebuah dokumen dipisahkan dengan spasi.

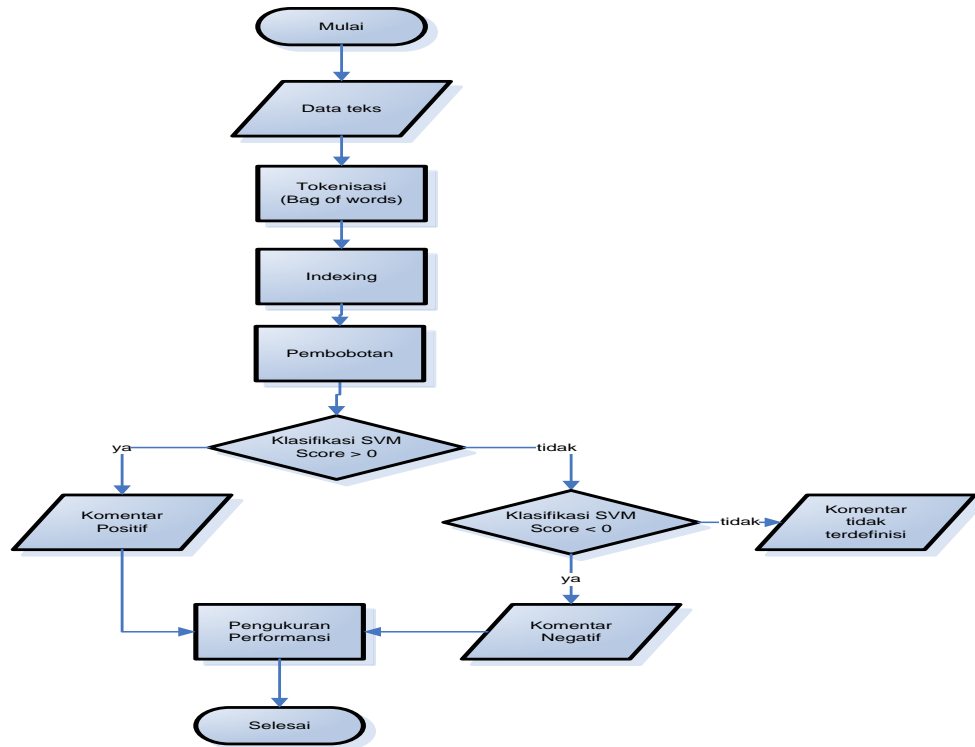
SVM adalah sistem untuk melakukan klasifikasi, tetapi mereka hanya peduli tentang titik dalam ruang, daripada email atau dokumen. Untuk tujuan ini model ruang vektor digunakan untuk memberikan setiap kata dalam dokumen sebuah ID (dimensi) dan sebuah bobot berdasarkan seberapa penting keberadaannya dalam dokumen (posisi dokumen dalam dimensi itu). SVM mencoba untuk menemukan garis yang terbaik membagi dua kelas, dan kemudian mengklasifikasikan dokumen uji berdasarkan di sisi mana dari garis tersebut mereka muncul.

Intuisi yang mendorong SVM sebagai penentu garis terbaik yang memisahkan kedua kelas adalah yang memiliki margin terbesar diantaranya dan titik pelatihan *sample* terdekat di kedua sisinya. Oleh karena itu, *support vector* adalah vektor yang menentukan margin tersebut dimana vektor ini yang paling dekat dengan

*dividing lines*. *Dividing lines* merupakan kombinasi dari vektor yang memberikan keputusan fungsi (kelas atau bukan kelas) untuk *classifier SVM*.

Penilaian kemudian dibuat dengan menilai *score* apakah positif atau negatif yang merepresentasikan di sisi mana dari garis pemisah dokumen berada. Sejauh ini fungsi kernel dapat diasumsikan sebagai *dot product* antara dua vektor.

Diagram alir proses klasifikasi dengan SVM ditunjukkan oleh gambar 1.



Gambar 1. Diagram Alir Klasifikasi dengan SVM

## 1. Percobaan dan Analisis

Data opini berbahasa Inggris diambil dari review film sedangkan data opini berbahasa Indonesia diambil dari rubrik Bali Terkini yang dimuat di harian Bali Post dari bulan Januari 2010 sampai Februari 2011. Jumlah kalimat opini adalah 5000 record masing-masing untuk opini negatif maupun positif. Data tersebut dibagi menjadi data latih dan data uji. Pada tahap pertama akan dibentuk pengetahuan berdasarkan masukan data latih terhadap masing-masing metode. Tahap selanjutnya dilakukan dengan memasukkan data uji ke dalam software yang telah dibekali pengetahuan tersebut. Akurasi dihitung dengan rumus :

$$\text{Akurasi} = \frac{\text{jumlah kalimat uji dengan klasifikasi benar}}{\text{jumlah semua kalimat yang diuji}} \times 100\% \quad (14)$$

Berikut adalah rata-rata nilai akurasi yang didapat dengan variasi jumlah data uji dan data latih yang berbeda.

Tabel 3. Rata-rata akurasi hasil percobaan

Opini	NBC	SVM
Berbahasa Inggris	78,38%	88,51%
Berbahasa Indonesia	74,39%	75,06%

Dari tabel 3 kita melihat hasil percobaan menunjukkan bahwa metode SVM memberikan unjuk kerja yang lebih baik daripada metode NBC untuk mengklasifikasikan opini berbahasa Inggris dan opini berbahasa Indonesia. Perbedaan akurasi tidak terlalu besar yang menunjukkan NBC masih merupakan metode yang baik dalam pengklasifikasian teks. NBC dapat menjadi pilihan mengingat sifat komputasinya yang lebih ringan.

Dari tabel 3 pula kita melihat kecenderungan hasil akurasi lebih baik pada opini berbahasa Inggris. Berdasarkan analisis sifat kosakata dalam bahasa Inggris dimana kata negatif maupun kata positif berdiri secara independen serta cara penulisan opini yang minim menggunakan kata negasi seperti 'NOT' maka metode klasifikasi menunjukkan hasil yang lebih baik dibandingkan klasifikasi dalam kalimat berbahasa Indonesia. Opini negatif

dalam bahasa Indonesia seringkali hanya dibedakan dengan kata negasi berupa kata ‘TIDAK’ dan melibatkan kosakata yang sama dengan opini positifnya.

Untuk data uji yang semakin besar maka data latih akan semakin kecil. Berikut adalah hasil pengujian secara lengkap.

Tabel 4. Hasil Pengujian

Data Berbahasa Inggris			Data Berbahasa Indonesia		
Variasi data uji	NBC	SVM	Variasi data uji	NBC	SVM
30 data uji positif	80,18%	79,72%	30 data uji positif	74,29%	78,2%
100 data uji positif	78,29%	80,15%	100 data uji positif	68,57%	75,71%
500 data uji positif	76,60%	78,51%	500 data uji positif	63,14%	71,88%
1000 data uji positif	76,03%	77,85%	1000 data uji positif	59,44%	70,67%
2000 data uji positif	75,26%	77,99%	2000 data uji positif	57,65%	75,09%
3000 data uji positif	74,40%	76,13%	3000 data uji positif	60,26%	65,93%
30 data uji negatif	83,86%	97,69%	30 data uji negatif	87,14%	87,43%
100 data uji negatif	82,65%	98,36%	100 data uji negatif	85%	80,52%
500 data uji negatif	80,91%	98,83%	500 data uji negatif	85,42%	75,24%
1000 data uji negatif	78,86%	98,95%	1000 data uji negatif	86,68%	75,14%
2000 data uji negatif	77,50%	98,90%	2000 data uji negatif	83,07%	72,85%
3000 data uji negatif	75,98%	98,98%	3000 data uji negatif	82,06%	72,02%

Dari tabel 4 kita melihat semakin kecil data uji diiringi semakin besar data latih menunjukkan kecenderungan peningkatan akurasi. Variasi 30 data uji menyatakan bahwa percobaan dilakukan dengan 30 kalimat yang diuji dan 4970 kalimat yang digunakan sebagai data latih yang diperoleh dari semua kalimat dalam data dikurangi kalimat yang akan diuji.

#### 4. Simpulan dan Saran

Berikut adalah simpulan dan saran yang didapatkan dari hasil penelitian.

##### 4.1 Simpulan

Hasil percobaan menunjukkan bahwa metode SVM memberikan unjuk kerja yang lebih baik daripada metode NBC untuk mengklasifikasikan opini berbahasa Inggris dan opini berbahasa Indonesia. Semakin kecil data uji diiringi semakin besar data latih menunjukkan kecenderungan peningkatan akurasi. Selama ini belum ditemukan klasifikasi opini berbahasa Indonesia dengan metode SVM. Metode SVM ternyata dapat digunakan untuk mengklasifikasikan opini berbahasa Indonesia seperti halnya opini berbahasa Inggris. Hasil akurasi lebih baik pada klasifikasi opini berbahasa Inggris dibandingkan opini berbahasa Indonesia yang disebabkan sifat kosakatanya.

##### 4.2 Saran

Pada penelitian ini kalimat yang akan diklasifikasikan dipandang sebagai *bag of words* atau sekumpulan kata-kata. Faktor yang berpengaruh adalah frekuensi kemunculan kata pada kalimat tersebut. Kedepannya diharapkan dapat diteliti pengklasifikasian kalimat yang juga memperhitungkan penanganan frase sehingga membentuk sebuah *sentence processor*.

#### 5. DAFTAR RUJUKAN

- [1] Barber, I. 2010. *Bayesian Opinion Mining*. [Online]. Tersedia di: <http://phpir.com/bayesian-opinion-mining> [diunduh: 10 Nov 2010].
- [2] Dehaff, M. 2010. *Sentiment Analysis, Hard But Worth It!*. [Online]. Tersedia di: <http://www.customerthink.com/blog/sentiment-analysis-hard-but-worth-it> [diunduh : 5 Nov 2010].
- [3] Bridge, C. 2011. *Unstructured Data and the 80 Percent Rule*. [Online]. Tersedia di: <http://www.clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551> [diunduh : 5 Nov 2010].
- [4] Colas, F. & Brazdil, P. 2005. *Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks*.

- [5] Tan, P. N., Steinbach, M. & Kumar, V. 2006. *Introduction to Data Mining*. Boston : Pearson Addison Wesley.
- [6] Wulandini, F. & Nugroho, A. N. 2009. *Text Classification Using Support Vector Machine for Webmining Based Spation Temporal Analysis of the Spread of Tropical Diseases. International Conference on Rural Information and Communication Technology 2009*. [Online] 189-192. Tersedia di: [http://asnugroho.net/papers/rict2009\\_textclassification.pdf](http://asnugroho.net/papers/rict2009_textclassification.pdf) [diunduh:5 Nov 2010].
- [7] Wibisono, Y. 2005. Klasifikasi Berita Berbahasa Indonesia menggunakan *Naïve Bayes Classifier*. [Online]. Tersedia di: [http://fpmipa.upi.edu/staff/yudi/yudi\\_0805.pdf](http://fpmipa.upi.edu/staff/yudi/yudi_0805.pdf) [diunduh: 1 Nov 2010]
- [8] Wibisono, Y. 2005. *Clustering Berita Berbahasa Indonesia*. [Online]. Tersedia di: [http://fpmipa.upi.edu/staff/yudi/KNSI\\_Clustering\\_yudi\\_masayu.pdf](http://fpmipa.upi.edu/staff/yudi/KNSI_Clustering_yudi_masayu.pdf) [diunduh : 1 Nov 2010].