

Penggunaan Metode *Naïve Bayes Classifier* pada Analisis Sentimen *Facebook* Berbahasa Indonesia

Ni Putu Sri Merta Suryani¹, Linawati², Komang Oka Saputra³

[Submission: 26-03-2019, Accepted: 15-04-2019]

Abstract— Sentiment analysis is a field that is currently in great demand by various groups. Sentiment analysis can be done using documents and opinions from social media. One social media that is usually used as a means of opinion is Facebook social media. Before a text is classified, it is necessary to do POS Tagging which is the word labeling stage where the purpose is to determine the words which include opinions and non opinions. For labeling words can use the Hidden Markov Model or Rule Based. The method commonly used in sentiment analysis is the Naïve Bayes Classifier method. This method simply classifies probabilities. Naïve Bayes Classifier can be used to classify opinions into positive and negative opinions. In addition, this method uses training data in the classification process. The classification produced from the Naïve Bayes Classifier method is quite good. To test the accuracy of the system in classifying opinions, the classification results are tested. From the test results obtained an average accuracy of 87.1%. The more training data that is similar to testing data, the better the classification results.

Intisari— Analisis sentimen merupakan bidang yang saat ini sangat diminati oleh berbagai kalangan. Analisis sentimen dapat dilakukan dengan menggunakan dokumen maupun opini dari media sosial. Salah satu media sosial yang biasanya digunakan sebagai sarana opini adalah media sosial Facebook. Sebelum sebuah teks diklasifikasikan, perlu dilakukan POS Tagging yang merupakan tahap pelabelan kata dimana tujuannya adalah untuk menentukan kata yang termasuk opini dan non opini. Untuk pelabelan kata dapat menggunakan Hidden Markov Model maupun Rule Based. Adapun metode yang biasa digunakan dalam analisis sentimen adalah metode *Naïve Bayes Classifier*. Metode ini mengklasifikasikan probabilitas dengan sederhana. *Naïve Bayes Classifier* dapat digunakan untuk mengklasifikasikan opini ke dalam opini positif maupun negatif. Selain itu metode ini menggunakan data latih dalam proses klasifikasinya. Klasifikasi yang dihasilkan dari metode *Naïve Bayes Classifier* tergolong cukup baik. Untuk menguji tingkat keakuratan sistem dalam mengklasifikasikan opini, maka dilakukan pengujian hasil klasifikasi. Dari hasil pengujian diperoleh rata-rata akurasi sebesar 87,1 %. Semakin banyak data latih yang mirip dengan data testing maka hasil klasifikasi akan semakin bagus.

Kata Kunci— Facebook, Analisis Sentimen, POS Tagging, *Naïve Bayes Classifier*, Akurasi

I. PENDAHULUAN

¹Mahasiswa Pascasarjana, Teknik Elektro Fakultas Teknik Universitas Udayana, Jln Kampus Bukit Jimbaran 80361 INDONESIA (tel: 0361-703315; fax: 0361-703315; e-mail: indertasuryani@gmail.com)

^{2, 3}Dosen, Jurusan Teknik Elektro dan Komputer Fakultas Teknik Universitas Udayana, Jln. Jalan Kampus Bukit Jimbaran 80361 INDONESIA (tel: 0361-703315; fax: 0361-703315; e-mail: linawati@unud.ac.id, okasaputra@unud.ac.id)

Di era digital ini peranan teknologi informasi sangat penting. Salah satunya adalah adanya teknologi *Big Data* yang saat ini menjadi sebuah trend. *Big Data* salah satunya menyimpan informasi dalam bentuk teks yang jumlahnya sangat besar. Informasi yang berbentuk teks adalah sebuah informasi yang dapat diperoleh baik dari buku, literatur maupun media sosial. Teks adalah sebuah hamparan bahasa, baik secara lisan maupun tulisan yang memiliki sebuah makna tertentu, bersifat praktis dan bermanfaat untuk umum serta berhubungan dengan dunia nyata [1].

Informasi yang diambil pada sebuah teks atau yang disebut dengan *text mining* seperti analisis sentimen (*sentiment analysis*). Analisis sentimen (*sentiment analysis*) merupakan sebuah studi dalam bidang komputasi dengan menggunakan opini, pendapat dan emosi yang di representasikan ke dalam bentuk teks [2]. Analisis sentimen dapat mengelompokkan teks yang terdapat kalimat, opini, pendapat maupun dokumen. Sentimen tersebut dapat berupa aspek positif, negatif atau netral [3]. Analisis sentimen bertujuan untuk mengetahui pendapat terhadap suatu masalah apakah termasuk opini positif atau negatif.

Saat ini banyak penelitian yang berkaitan dengan analisis sentimen. Salah satunya yang menjadi trend saat ini adalah analisis sentimen pada opini yang terdapat di media sosial. Media sosial tersebut diantaranya adalah Facebook, Twitter, Path, Instagram dan media sosial lainnya. Analisis sentimen penting dilakukan karena dengan itu kita dapat mengetahui dengan cepat sentimen pada opini khususnya pada media sosial Facebook [4].

Ada berbagai metode yang digunakan dalam analisis sentimen, salah satunya adalah *Naïve Bayes Classifier*. *Naïve Bayes Classifier* adalah sebuah metode yang berbasis probabilistik. Metode *Naïve Bayes Classifier* adalah metode sederhana tetapi metode ini memiliki nilai akurasi dan performansi yang tinggi dalam mengklasifikasikan sebuah teks [5]. *Naïve Bayes Classifier* dapat digunakan untuk mengklasifikasikan sebuah opini ke dalam opini positif maupun negatif [6]. Terdapat hasil penelitian yang menunjukkan bahwa metode *Naïve Bayes Classifier* merupakan salah satu metode yang terbaik untuk pelatihan domain-domain dan hasil klasifikasinya memiliki tingkat akurasi yang tinggi.

Selain itu dalam analisis sentimen juga melibatkan pembelajaran bahasa dalam hal ini menggunakan metode *Part of Speech (POS) Tagging*. *POS Tagging* bertujuan untuk memberikan kelas atau label pada kata dalam suatu kalimat [7]. *POS Tagging* dapat menggunakan *Hidden Markov Model* dan *Rule Based*. *POS Tagging* dilakukan agar dapat diketahui kalimat yang mengandung opini dan non opini sebelum diklasifikasikan.



Pada penelitian ini dilakukan sebuah analisis sentimen pada opini yang terdapat pada media sosial *Facebook*. Opini yang digunakan merupakan opini dalam bahasa Indonesia. Tahapan dari analisis sentimen adalah pengambilan opini yang berbahasa Indonesia pada media sosial *Facebook*. Opini yang terkumpul kemudian di *preprocessing*. Pada proses *preprocessing* juga dilakukan proses *stemming*. Proses *stemming* merupakan proses merubah kata menjadi kata dasarnya. Setelah itu dilakukan pelabelan kata (*POS Tagging*) dengan *Rule Based*. Tujuannya untuk mengetahui kata yang termasuk opini dan non opini. Kalimat-kalimat yang termasuk opini kemudian akan diklasifikasikan ke dalam empat kelas dengan menggunakan metode *Naïve Bayes Classifier*. Selanjutnya dilakukan evaluasi dengan melakukan pengujian untuk mengetahui tingkat akurasi dari hasil klasifikasi metode *Naïve Bayes Classifier*.

II. TINJAUAN NAÏVE BAYES CLASSIFIER DALAM ANALISIS SENTIMEN

Sentimen analisis merupakan sebuah teknik menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi mengenai layanan, produk, individu atau kegiatan tertentu. Tujuan dari analisis sentimen untuk menentukan opini dari seseorang dengan memperhatikan topik tertentu. Langkah yang dilakukan dalam analisis sentimen adalah mendefinisikan domain dataset, *preprocessing*, *featured selection*, pelabelan, klasifikasi dan evaluasi.

Part of Speech (POS) Tagging merupakan sebuah proses yang bertujuan untuk memberikan label kata secara otomatis pada kalimat. Ada dua metode yang biasa digunakan dalam *POS Tagging* adalah *Hidden Markov Model* dan *Rule Based*. *Hidden Markov Model* digambarkan dalam satu kalimat yang dapat diberikan alur tag kata yang paling tepat. Sedangkan *Rule Based* menggunakan kamus untuk menandai kata dengan kelas kata (*leksikon*).

Metode *Naïve Bayes Classifier* dapat digunakan dalam mengklasifikasikan sebuah teks. Metode *Naïve Bayes Classifier* merupakan sebuah teknik dalam klasifikasi probabilistik yang sederhana dimana dalam teknik ini dilakukan perhitungan probabilitas dengan cara melakukan penjumlahan terhadap frekuensi dan kombinasi nilai dari sebuah dataset [8]. Adapun keuntungan dengan menggunakan metode *Naïve Bayes Classifier* karena metode *Naïve Bayes Classifier* memerlukan data latih yang sedikit dalam proses klasifikasi teks. Pada persamaan (1) merupakan persamaan *Naïve Bayes Classifier*.

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

Ket :

- X : Data dengan kelas yang tidak diketahui.
- H : Hipotesis data adalah kelas yang spesifik.
- P(H|X) : Probabilitas H berdasarkan kondisi X.
- P(H) : Probabilitas H.
- P(X|H) : Probabilitas X berdasarkan kondisi dari H.

P(X) : Probabilitas X.

Dalam melakukan proses klasifikasi, metode *Naïve Bayes Classifier* membutuhkan beberapa petunjuk dalam menentukan kelas yang cocok digunakan dalam menganalisis sebuah sampel. Sehingga persamaan (1) disesuaikan berdasarkan persamaan (2).

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)} \quad (2)$$

Pada persamaan (2), variabel C merupakan sebuah kelas, sedangkan $F_1 \dots F_n$ merupakan karakteristik dari petunjuk yang diperlukan dalam proses klasifikasi. Peluang dari masuknya sampel dengan karakteristik-karakteristik tertentu pada kelas C (*posterior*) merupakan sebuah peluang munculnya kelas C (sebelum masuknya sampel disebut *prior*), selanjutnya dikalikan dengan peluang munculnya karakteristik dari sebuah sampel pada kelas C (*likelihood*), dibagi dengan peluang munculnya karakteristik sampel secara global (*evidence*). Pada persamaan (3) merupakan rumus metode *Naïve Bayes Classifier* sederhana.

$$\text{Posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \quad (3)$$

Dalam satu sampel terdapat sebuah kelas dimana setiap kelas tersebut memiliki nilai *evidence* yang selalu tetap. Nilai dari *posterior* kemudian dibandingkan dengan nilai *posterior* yang ada pada setiap kelas. Tujuannya adalah untuk menentukan ke kelas apa sebuah sampel akan diklasifikasikan. Persamaan (4) menjabarkan rumus *Naïve Bayes Classifier* berdasarkan aturan perkalian.

$$\begin{aligned} P(C|F_1, \dots, F_n) &= P(C)P(F_1, \dots, F_n|C) \\ &= P(C)P(F_1|C)P(F_2, \dots, F_n|C, F_1) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n|C, F_1, F_2) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \dots, F_n|C, F_1, F_2, F_3) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots \\ &\quad P(F_n|C, F_1, F_2, F_3, \dots, F_n - 1) \quad (4) \end{aligned}$$

Hasil penjabaran pada persamaan (4) menyebabkan semakin banyak faktor yang mempengaruhi nilai probabilitas maka menyulitkan untuk menganalisis jika harus dilakukan analisa satu per satu. Untuk mengatasi hal tersebut, maka dilakukan asumsi independensi yang sangat tinggi (*naïf*), bahwa setiap petunjuk (F_1, F_2, \dots, F_n) bersifat independen antara satu sama lain. Pada persamaan (5) merupakan asumsi independensi.

$$P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i) \quad (5)$$

Untuk $i \neq j$, maka

$$P(F_i|C, F_j) = P(F_i|C)$$

Persamaan (5) adalah model teorema *Naïve Bayes Classifier* yang nantinya digunakan dalam melakukan proses klasifikasi [9].

Hasil dari proses klasifikasi kemudian diuji untuk menentukan tingkat akurasi sistem dalam melakukan proses klasifikasi. Pengujian untuk algoritma *Naïve Bayes Classifier* sebagai hasil kasifikasi dokumen ditunjukkan pada persamaan (6) dimana pengujian ini nantinya akan dilakukan dengan cara menghitung data-data yang sudah memiliki label secara manual dengan hasil yang diberikan oleh sistem [10].

$$P(P(NA)) = \frac{Data}{X} \times 100\% \dots\dots\dots(6)$$

Dimana:

- $P(NA)$: Persentase nilai akurasi
- $Data$: jumlah data uji yang bernilai benar
- X : total jumlah sampel

III. HASIL DAN PEMBAHASAN

A. Implementasi POS Tagging

Pada tahap *POS Tagging* menggunakan metode *Rule Based*. *Rule* yang digunakan untuk menentukan suatu kalimat apakah merupakan sebuah opini atau non opini. Setelah diketahui data yang menjadi opini kemudian menggunakan *rule* kembali untuk mengetahui opini yang berkaitan dengan bantuan sosial. Proses *POS Tagging* adalah membandingkan jumlah bobot pada kalimat yang baru di import, dengan setiap kalimat pada *Rule Based*. Pada Tabel I terdapat contoh tabel *Rule Based* sudah memiliki index dan bobotnya masing-masing. Kalimat baru yang di import akan disandingkan dengan bobot *Rule Based*, bobot yang paling dekat akan menjadi tag dari kalimat tersebut.

TABEL I
 CONTOH *RULE BASED* DAN PERBANDINGAN

Kalimat Baru	Bobot	Kalimat Rule Based	Opini/non opini	Bobot	Hasil Tag
Saya puas dengan kebijakan pemerintah	0,3	Saya tidak puas	opini	0,5	-
Saya puas dengan kebijakan pemerintah	0,3	Saya puas	non opini	0	-
Saya puas dengan kebijakan pemerintah	0,3	Saya menyukai pemerintah	opini	0,3	opini
Saya puas dengan kebijakan pemerintah	0,3	Saya tidak suka	non opini	0	-

Karena perbandingan pada baris ke-3 memiliki nilai kemiripan paling tinggi, maka tag dari kalimat “saya puas dengan kebijakan pemerintah” adalah opini.

B. Implementasi Naïve Bayes Classifier

Klasifikasi dilakukan dengan menggunakan metode *Naïve Bayes Classifier*. Dalam melakukan klasifikasi akan memerlukan data latih sebagai proses mesin pembelajaran. Penelitian ini menggunakan data latih yang bersumber dari Facebook. Pada Tabel II terdapat hasil klasifikasi dengan menggunakan metode *Naïve Bayes Classifier*.

TABEL II
 HASIL KLASIFIKASI *NAÏVE BAYES CLASSIFIER*

SENTIMEN	HASIL KLASIFIKASI SISTEM
Positif	294
Negatif	130
Netral	52
Undefined	3

Pada Tabel II dapat dilihat hasil dari klasifikasi sistem menunjukkan bahwa dari 479 opini masyarakat terhadap program bantuan sosial lebih ke sentimen positif sebanyak 294 opini. Sedangkan opini masyarakat ke arah sentimen negatif sebanyak 130 opini, sentimen netral sebanyak 52 opini dan sentimen *undefined* sebanyak 3 opini.

C. Pengujian Akurasi

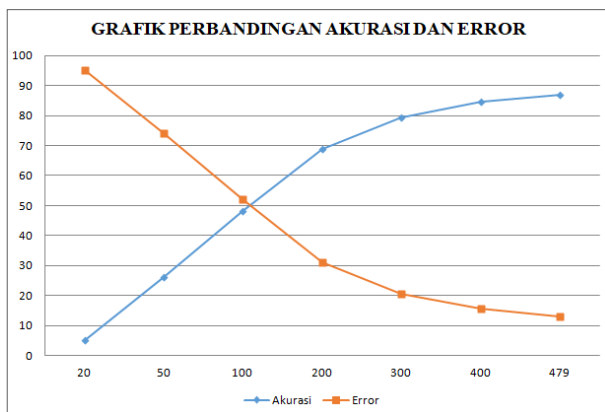
Akurasi dan error pada penelitian ini dihitung berdasarkan ketepatan dan kesalahan sistem dalam melakukan klasifikasi dengan menggunakan persamaan 8. Pengujian akurasi dilakukan untuk mengetahui tingkat akurasi klasifikasi opini yang dilakukan secara manual dengan klasifikasi opini yang dilakukan oleh sistem. Pengujian dilakukan dengan menggunakan 20 data hingga 479 data yang diambil secara acak dan sudah diberi label. Berikut pada Tabel III merupakan persentase nilai akurasi dari jumlah data uji.

TABEL III
 PERSENTASE AKURASI

No	Jumlah Data Uji	Akurasi	Error
1	20 data uji	5%	95%
2	50 data uji	26%	74%
3	100 data uji	48%	52%
4	200 data uji	69%	31%
5	300 data uji	79.3%	20.6%
6	400 data uji	84.5%	15.5%
7	479 data uji	87.1%	12.9%

Pada Tabel III, dari 20 data uji diperoleh nilai akurasi sebesar 5%. Sedangkan pada 50 data uji diperoleh akurasi sebesar 26%. Pada 100 data diperoleh nilai akurasi sebesar 48%. Untuk data uji 200 diperoleh akurasi sebesar 69%. Pada 300 data uji diperoleh akurasi sebesar 79,3%. Untuk 400 data uji diperoleh nilai akurasi sebesar 84,5%. Sedangkan dengan 479 data uji diperoleh nilai akurasi sebedar 87,1%. Berdasarkan pada Tabel III terdapat penurunan tingkat akurasi yang tidak signifikan. Hal tersebut karena ada ketidakcocokan data uji dengan data latih yang ada dalam sistem. Semakin banyak data uji yang memiliki kecocokan dengan data latih maka tingkat akurasinya akan semakin tinggi. Grafik hasil pengujian akurasi dan error dapat dilihat pada Gambar 1.





Gambar 1. Grafik Pengujian Akurasi dan Error

Berdasarkan pada Gambar 1, dapat dilihat bahwa nilai akurasi pada masing-masing data uji menghasilkan nilai yang berbeda. Dari 479 data yang diujikan diperoleh akurasi sebesar 87,1% dan error 12,9%. Sedangkan dari 20 data uji diperoleh akurasi sebesar 5% dan error 95%. Hal ini karena data uji memiliki kemiripan dengan data latih pada sistem. Semakin banyak data latih yang digunakan pada sistem akan memberikan pengaruh terhadap kinerja sistem dalam melakukan proses klasifikasi.

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dalam analisis sentimen penggunaan metode *Naïve Bayes Classifier* akan menimbulkan tingkat akurasi yang tinggi. Semakin banyak data yang diujikan maka tingkat akurasi dari hasil klasifikasi akan semakin tinggi. Hal ini juga dipengaruhi oleh jumlah data latih yang digunakan pada metode *Naïve Bayes Classifier*. Kedepannya diharapkan dalam penelitian ini dapat menggunakan metode lain sebagai pembandingan dari metode *Naïve Bayes Classifier*.

REFERENSI

- [1] Bolshakov and Gelbukh, *Computational Linguistics*. 1st edition. 2004
- [2] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers, May 2012.
- [3] A. Ortigosa, J. M. Martin, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning", *Computers in Human Behavior*, vol. 31, pp. 527–541, 2014.
- [4] Surroop, K., Canoo, K., & Pudaruth, S., "A Novel Position-based Sentiment Classification Algorithm for Facebook Comments", *International Journal of Advanced Computer Science and Applications*, Vol 7, issue 10, 261–268, 2017.
- [5] Routray, Swain, and Mishra, "A Survey on Sentiment Analysis", *International Journal of Computer Applications*, August, Vol 70, Issue 10, 2013.
- [6] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of Facebook statuses using Naïve Bayes classifier for language learning", *IEEE*, 2013.
- [7] K. Widhiyanti dan A. Harjoko, "POS Tagging Bahasa Indonesia dengan HMM dan Rule Based", *Jurnal Teknologi Komputer dan Informatika*, Vol 8, No 2, Nopember 2012.
- [8] P. B. Matharasi dan A. Senthilrajan, "Sentiment Analysis of Twitter Data Using Naïve Bayes with Unigram Apparoch", *International Journal of Scientific and Research Publication*, vol 7, issue 5, May 2017.
- [9] (2017) Informatikalogi website. [Online]. Available: <https://informatikalogi.com/algortima-naive-bayes/>
- [10] Astaridewi, "Sistem Klasifikasi Reporting Berita Menggunakan Metode Naive Bayes (Studi Kasus Situs Resmi Pemerintahan). Jimbaran: Program Studi Teknik Informatika Jurusan Ilmu Komputer

Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Udayana". Skripsi, Universitas Udayana, Indonesia, 2015.