

Analisis sentimen pada twitter dengan menggunakan metode Naive Bayes Classifier

Sigit Suryono¹ dan Emha Taufiq Luthfi²

- 1 Program Magister Teknik Informatika Program Pascasarjana Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta 55283
sigit.suryono@students.amikom.ac.id
- 2 Program Magister Teknik Informatika Program Pascasarjana Universitas Amikom Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta 55283
emhataufiq_luthfi@amikom.ac.id

Abstrak

Analisis Sentimen merupakan salah satu cabang dari bidang ilmu *text mining*. Analisis sentimen merupakan sumber penting dalam melakukan evaluasi dan pengambilan keputusan terhadap sebuah topik permasalahan. Tujuan utama dari analisis sentimen adalah untuk mengetahui polaritas dari kumpulan sentimen yakni apakah bernilai positif, negatif ataupun netral. Data sentimen paling umum bisa didapatkan dari media sosial, salah satunya adalah Twitter. Dalam penelitian ini, tweet-tweet yang berhubungan dengan suatu kata kunci tertentu, dicari dan dikumpulkan dari Twitter dengan menggunakan *application programming interface* menjadi data mentah. Data mentah tersebut diolah dengan *Natural Language Toolkit* menggunakan bahasa pemrograman Python untuk diproses lebih lanjut. Klasifikasi menggunakan metode *Naive Bayes Classifier* untuk mengetahui tingkat akurasi dari proses klasifikasi dilakukan dalam penelitian ini. Program *RapidMiner* digunakan untuk melakukan proses klasifikasi. Dari hasil uji coba sebanyak empat kali, didapatkan hasil tingkat akurasi pada percobaan pertama sebesar 62.98%, percobaan kedua sebesar 64.95%, percobaan ketiga sebesar 66.36%, dan percobaan keempat sebesar 66.79%. Dari hasil klasifikasi didapat tingkat persentase sentimen positif sebesar 28%, sentimen negatif sebesar 20% dan sentimen netral sebesar 52%.

Kata Kunci analisis sentimen, naive bayes, klasifikasi, natural language, opinion mining

1 Pendahuluan

Analisis sentimen merupakan salah satu cara untuk mengumpulkan pendapat orang banyak terhadap sesuatu seperti layanan publik, isu, kinerja pemerintahan atau hal lain yang berkaitan dengan sudut pandang orang terhadap suatu permasalahan. Analisis sentimen dapat digunakan sebagai salah satu cara untuk melakukan evaluasi terhadap layanan yang telah diberikan. Analisis sentimen dapat dilakukan melalui berbagai cara, salah satunya adalah dengan mengumpulkan pendapat orang banyak melalui media sosial.

Media sosial di Indonesia tidak hanya digunakan sebagai media untuk mengungkapkan apa yang sedang dirasakan tetapi juga digunakan sebagai media untuk memberikan saran atau kritik terhadap layanan yang telah diberikan oleh pemerintah. Melalui media sosial khususnya Twitter banyak orang yang memuji, menyalahkan atau tidak keduanya terhadap apa yang dilakukan oleh pemerintahan sekarang dalam hal ini pemerintahan presiden Joko Widodo.

Tweet yang berkaitan dengan pemerintahan presiden Joko Widodo dapat dikumpulkan lalu dilakukan analisis sentimen dengan menggunakan *Naive Bayes Classifier* (NBC). Pengumpulan data dilakukan dengan menggunakan *application programming interface* (API)



© Sigit Suryono dan Emha T. Luthfi;
licensed under Creative Commons License CC-BY
Jurnal Open Access

Jnanaloka Yayasan Lentera Dua Indonesia

yang telah disediakan oleh Twitter. Pengumpulan data dengan menggunakan API Twitter diimplementasikan ke dalam bahasa pemrograman Python. Pelabelan data tweet yang telah didapatkan dilakukan dengan menggunakan *sentistrength* yang diimplementasikan juga dengan menggunakan bahasa pemrograman Python. Sebanyak 40% data yang telah diberikan label digunakan sebagai dasar untuk menentukan sentimen terhadap tweet yang belum diberikan label dengan menggunakan metode NBC. Pengukuran tingkat akurasi metode NBC dengan menggunakan RapidMiner serta menghitung persentase untuk masing-masing sentimen berdasarkan tweet yang telah didapatkan sebelumnya.

Ramadhan WP, Astri Novianty dan Casi Setianingsih melakukan sebuah analisis sentimen dengan menggunakan *Support Vector Machine* (SVM) dan *Maximum Entropy* [1]. Dalam penelitian tersebut analisis sentimen dilakukan terhadap masing-masing calon dalam pemilihan kepala daerah Daerah Khusus Ibukota (DKI) Jakarta dan bertujuan untuk mencari tahu metode serta kernel yang memiliki tingkat akurasi yang baik. Penelitian memberi hasil bahwa metode SVM dan kernel linear memiliki tingkat akurasi yang tinggi. Dalam penelitian yang dilakukan oleh Ramadhan WP, Astri Novianty, Casi Setianingsih tersebut tidak memberikan gambaran berapa hasil persentase untuk masing-masing sentimen yang ada.

Fiktor Imanuel Tanesab, Irwan Sembiring and Hindriyanto Dwi Purnomo melakukan analisis sentimen terhadap komentar yang ada pada Youtube dengan menggunakan metode SVM [2]. Komentar yang dikumpulkan adalah tentang mantan gubernur DKI Jakarta Basuki Tjahaya Purnama. Penelitian tersebut bertujuan untuk mencari tingkat akurasi, presisi, *recall*, *true positive* (TP) dan *true negative* (TN) dari metode yang diajukan. Hasil dari penelitian tersebut memiliki tingkat akurasi 84%, presisi 91%, recall 80%, TP 91.1% dan TN 44.8%. Penelitian yang dilakukan oleh Fiktor Imanuel Tanesab, Irwan Sembiring dan Hindriyanto Dwi Purnomo tersebut belum memberikan gambaran berapa hasil persentase untuk masing-masing sentimen berdasarkan komentar yang telah diberikan sentimen sebelumnya.

Dalam penelitian ini dilakukan analisis sentimen pada Twitter dengan menggunakan metode NBC terhadap tweet- tweet yang berkaitan dengan pemerintahan presiden Joko Widodo. Hasil dari penelitian ini adalah tingkat akurasi dari metode yang digunakan serta persentase untuk masing-masing sentimen berdasarkan tweet-tweet yang telah didapat sebelumnya.

2 Metodologi

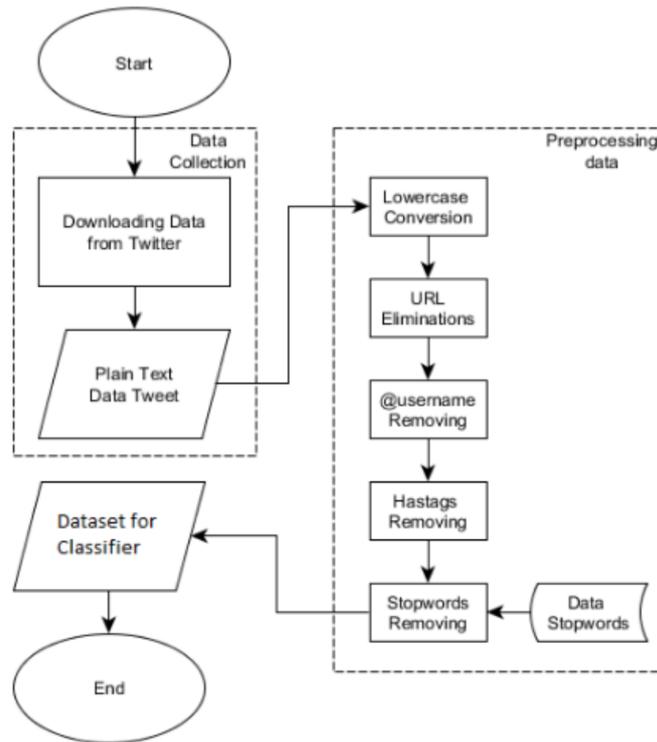
Analisis sentimen atau pada umumnya dikenal dengan *opinion mining* merupakan salah satu cabang studi tentang analisa pendapat ataupun pendapat seseorang sesuatu seperti layanan, produk, organisasi, dan lain sebagainya [3]. Analisis sentimen merupakan salah satu penelitian yang cukup kompleks. Karakteristik dari analisis sentimen adalah dapat terbagai sebagai berikut ini: [4]

1. pengkategorian sentimen yang akan membedakan antara kalimat subjektif dan objektif,
2. tingkatan analisis dibagi menjadi 3 bagian yaitu *message level*, *sentence level* dan *aspect level*,
3. pendapat yang memberikan perbandingan terhadap sesuatu serta pendapat yang hanya sekedar pendapat, dan
4. pembagian pendapat menjadi eksplisit dan implisit.

NBC merupakan salah satu metode atau model yang dapat melakukan proses klasifikasi secara baik [5]. Metode *Naïve Bayes* melakukan klasifikasi dengan menggunakan dua buah proses yang membagi data menjadi data *training* serta data *testing*. Proses klasifikasi

dalam *Naïve Bayes* terhadap data dilakukan dengan merepresentasikan setiap data kedalam “ $X_1, X_2, X_3, \dots, X_n$ ”.

Sebanyak 3485 data digunakan pada penelitian ini, data tersebut melalui beberapa proses untuk dapat menghasilkan data yang siap digunakan dalam proses klasifikasi. Proses pengumpulan, *preprocessing* data hingga menjadi data siap digunakan seperti tertampil pada Gambar 1.

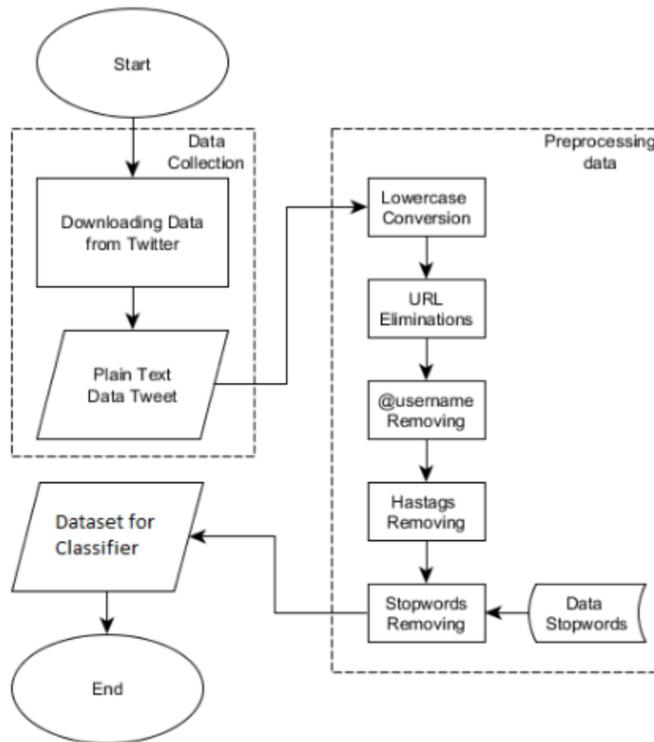


■ **Gambar 1** Proses pengumpulan dan *preprocessing* data

Gambar 1 dapat dijelaskan proses pertama yang dilakukan adalah pengumpulan data. Pengumpulan data dilakukan dengan cara mengunduh data tweet dari Twitter dan menghasilkan apa yang disebut dengan data mentah. Data mentah yang telah didapatkan tersebut perlu melalui tahapan *preprocessing* data terlebih dahulu sebelum dilakukan proses berikutnya. Tahapan *preprocessing* data terdapat lima sub proses yakni *lowercase conversion*, *URL eliminations*, *@username removing*, *hashtag removing* dan *stop words removing*. Proses *preprocessing* ini dilakukan dengan menggunakan *Natural language toolkit* (NLTK). NLTK merupakan sebuah *tools* yang dikembangkan khusus untuk bahasa pemrograman Python dan digunakan dalam proses yang berhubungan pemrosesan bahasa alamiah [6].

Setelah melalui *preprocessing* data maka dihasilkan dataset yang siap digunakan untuk proses klasifikasi. Proses klasifikasi dalam penelitian ini yang menggunakan metode NBC ditunjukkan pada Gambar 2.

Pada Gambar 2 dapat dijelaskan bahwa data yang telah siap diolah (diklasifikasi) diberikan label terlebih dahulu dengan menggunakan *sentistrength*. Data yang dilabeli diambil sebesar 40% dari dataset. Data yang telah diberikan label dijadikan sebagai acuan untuk menentukan klasifikasi sentimen pada data tweet yang belum memiliki label dengan menggunakan metode



■ **Gambar 2** Proses pengumpulan dan preprocessing data

NBC. Data yang telah memiliki label diukur tingkat akurasi dari metode NBC menggunakan aplikasi *RapidMiner*.

3 Hasil dan pembahasan

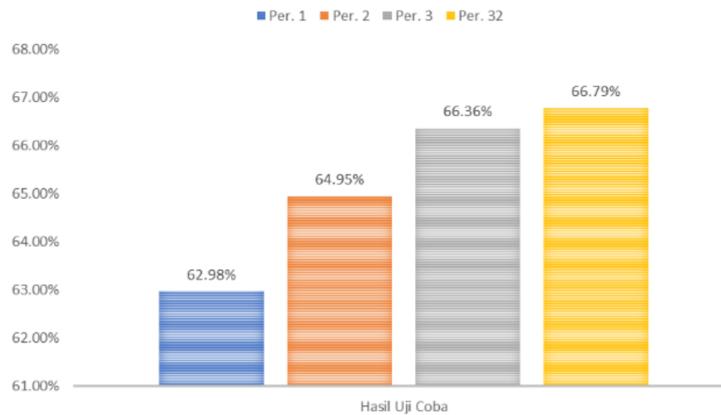
Data yang telah didapatkan yaitu sebanyak 3485 tweet dan 40% dari data tersebut telah dilabeli dengan *sentistrength* dan sisanya dilabeli dengan NBC. Skenario ujicoba untuk mengukur tingkat akurasi dari metode NBC pada penelitian ini ditunjukkan pada Tabel 1.

■ **Tabel 1** Skenario ujicoba

jumlah percobaan	data <i>training</i> (%)	data <i>testing</i> (%)
1	20	80
2	40	60
3	50	50
4	60	40

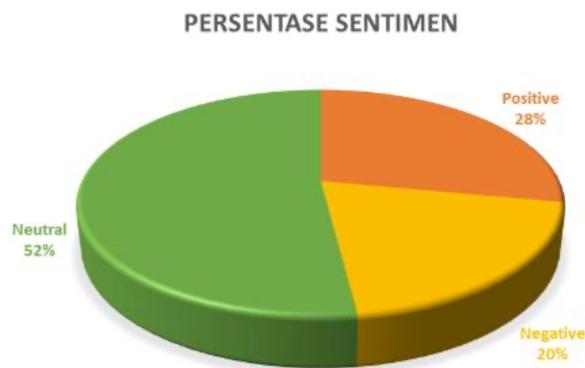
Berdasarkan hasil percobaan yang telah dilakukan sesuai dengan Table 1, adapun hasil tingkat akurasi dari keempat percobaan akan ditunjukkan pada Gambar 3.

Pada Gambar 3 dapat dilihat bahwa tingkat akurasi tertinggi didapat pada percobaan keempat dengan pembagian data 60:40. Pada percobaan keempat tersebut didapatkan



■ **Gambar 3** Tingkat akurasi dari 4 percobaan.

tingkat akurasi sebesar 66.79 %. Tingkat akurasi terendah didapat pada percobaan pertama dengan pembagian data 20:80. Pada percobaan pertama tingkat akurasi yang diperoleh adalah sebesar 62.98%. Berdasarkan hasil uji coba yang telah dilakukan dan hasil yang ditunjukkan, semakin besar data yang ditraining akan semakin meningkatkan akurasi. Hasil ini belum multak dikarenakan uji coba hanya dilakukan sebanyak empat kali. Berdasarkan hasil klasifikasi yang telah dilakukan adapun persentase sentimen dari dataset yang telah diperoleh tertampil pada Gambar 4.



■ **Gambar 4** Presentase sentimen

Pada Gambar 4 nampak bahwa persentase terbesar diperoleh oleh sentimen netral sebesar 52%. Ditempat kedua diperoleh oleh sentimen positif dengan 28% dan diurutan terakhir diperoleh oleh sentimen negatif sebesar 20%. Berdasarkan hasil tersebut maka sentimen yang dominan dalam data set yang telah diperoleh adalah sentimen netral. Ini berarti banyak para pengguna Twitter yang tidak memuji apa yang telah dilakukan oleh presiden Joko Widodo dan tidak juga menyalahkan apa yang telah dilakukan oleh presiden Joko Widodo.

4 Kesimpulan dan saran

Didasarkan pada hasil uji coba tingkat akurasi klasifikasi yang dilakukan oleh metode NBC memiliki nilai tertinggi pada percobaan keempat dengan 66.79% dengan perbandingan pembagian data yaitu 60% untuk data *training* dan 40% untuk data *testing*. Tingkat akurasi terendah diperoleh pada percobaan pertama dengan 62.98% dengan perbandingan pembagian data yaitu 20% untuk data *training* dan 80% untuk data *testing*. Berdasarkan hasil percobaan didapatkan fakta bahwa semakin banyak porsi data untuk data training maka tingkat akurasinya juga akan meningkat. Hasil lain yang diperoleh yaitu tingkat persentase sentimen. Persentase sentimen tertinggi didapat oleh sentimen netral dengan 52% lalu sentimen positif dengan 28% dan yang terakhir adalah sentimen negatif dengan 20%. Berdasarkan hasil persentase sentimen nampak bahwa pengguna Twitter didominasi oleh sentimen netral.

Untuk penelitian selanjutnya dapat menambahkan proses pada *preprocessing* data seperti proses untuk menangani singkatan-singkatan yang ada pada Tweet. Selain itu, penelitian ini dapat diterapkan terhadap metode pengklasifikasian data yang lainnya.

Pustaka

- 1 W. Ramadhan, S. A. Novianty, and S. C. Setianingsih, "Sentiment analysis using multinomial logistic regression," in *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*. IEEE, 2017, pp. 46–49.
- 2 F. I. Tanesab, I. Sembiring, and H. D. Purnomo, "Sentiment analysis model based on youtube comment using support vector machine," *International Journal of Computer Science and Software Engineering*, vol. 6, no. 8, p. 180, 2017.
- 3 B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- 4 F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, *Sentiment analysis in social networks*. Morgan Kaufmann, 2016.
- 5 C. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.
- 6 E. Loper and S. Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.