

## METODE NAÏVE BAYES CLASSIFIER DENGAN TEXTBLOB UNTUK ANALISIS SENTIMEN TERHADAP PELAYANAN INDIHOME DAN FIRST MEDIA

Navi Atri Lestari, Tubagus Mohammad Akhriza dan Eka Yuniar  
STMIK PPKIA Pradnya Paramita (STIMATA)  
Jl. Laksda Adi Sucipto 249A, Malang, Jawa Timur  
alnavi56@gmail.com

### ABSTRAK

*Media sosial saat ini menjadi media online yang banyak diakses oleh masyarakat. Salah satu jejaring sosial yang banyak digunakan adalah Twitter, yang memungkinkan pengguna untuk menulis tentang berbagai ulasan terhadap suatu produk, jasa layanan, tokoh publik dan sebagainya. Hal tersebut dapat dimanfaatkan oleh suatu perusahaan untuk melihat opini atau sentimen masyarakat terhadap produk/layanan mereka, yang kemudian dapat dimanfaatkan sebagai bahan evaluasi untuk meningkatkan layanan/produk perusahaan yang bersangkutan. Memanfaatkan data tweet dari Tweeter, penelitian ini diselenggarakan untuk mengetahui sentimen publik mengenai layanan yang diberikan oleh dua vendor Internet Service Provider (ISP) di Indonesia yaitu Indihome dan First Media. Sentimen diklasifikasi menjadi dua kelas yaitu positif dan negatif menggunakan metode Naive bayes classifier, dimana text-preprocessing diterapkan dalam menangani penulisan tweet yang tidak baku. Secara khusus, pustaka Textblob untuk Python digunakan untuk melakukan proses klasifikasi sentimen. Dari hasil penelitian dapat diketahui bahwa sentimen negatif paling banyak ditemukan dalam menanggapi pelayanan yang diberikan oleh kedua ISP. Hasil dari metode naïve bayes dalam penelitian ini memperoleh nilai Recall 66 %, Precision 83 % dan Accuracy 74 %.*

**Kata Kunci :** analisis sentimen, klasifikasi teks, naïve bayes, twitter

### PENDAHULUAN

Menurut data yang dirilis Twitter, pada tahun 2019 Indonesia menjadi negara yang pertumbuhan pengguna aktif harian twitter-nya paling besar [1]. Twitter memungkinkan pengguna untuk berbagi pesan menggunakan teks pendek disebut tweet. Terdapat sebuah ruang untuk melakukan penelitian dengan cara menganalisis opini dan sentimen secara komputatif [2]. Analisis mengenai tweet emosi para pengguna twitter disebut sebagai analisis opini atau sentimen (opinion analysis atau sentiment analysis). Analisis sentimen dilakukan untuk melihat kecenderungan pendapat atau opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung beropini negatif atau positif.

Sebagai dua vendor besar penyedia layanan internet dan TV kabel, Indihome dan First Media menyediakan channel twitter untuk menyampaikan pendapat mengenai layanan yang mereka berikan. Dengan beragamnya pendapat yang muncul tersebut, apabila diteliti lebih lanjut maka

akan dihasilkan sebuah sentimen yang dapat berguna untuk mengetahui tingkat kepuasan pelanggan terhadap suatu layanan atau produk, sehingga dapat menjadi bahan evaluasi terhadap layanan Indihome dan First Media agar dapat meningkatkan kualitas pelayanannya.

Dengan cara manual, analisa sentimen bisa saja dilakukan. Misalnya memonitor berita-berita di media massa. Akan tetapi untuk data tweet, cara manual tidak mungkin bisa dilakukan karena jumlah datanya yang sangat besar dan terus mengalir. Disinilah peranan text mining, yang secara otomatis dapat mengolah kata [3]. Setelah melakukan text mining, diperlukan normalisasi. Setelah melakukan normalisasi, dilakukan penggolongan atau clustering pada setiap tweet, salah satunya menggunakan Naïve Bayes.

Naïve Bayes Classifier (NBC) merupakan teknik prediksi kelas berbasis probabilitas sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif) [4].

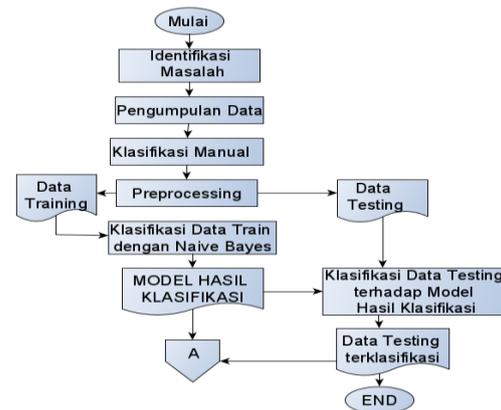
NBC telah banyak dipakai untuk melakukan Analisis Sentimen, seperti Analisis Penilaian Pelayanan Situs Belanja Online [5] membahas tentang opini yang ada di media sosial *twitter* terhadap beberapa situs belanja online di Indonesia. Dari hasil penelitian tersebut didapatkan rata-rata akurasi sebesar 93.33%. Analisis Sentimen pada Opini Produk [6] tentang pembuatan sebuah aplikasi berbasis website yang berhasil mengklasifikasikan sebuah sentimen kedalam tiga kelas yaitu: positif, negative, dan netral dengan tingkat akurasi 70.43% - 88,03%.

Memperhatikan efektifitas NBC dalam melakukan klasifikasi sentimen, artikel ini juga mengusulkan menggunakan NBC untuk melakukan klasifikasi sentimen masyarakat terhadap layanan di Indihome dan First Media yang disampaikan melalui Twitter.

#### METODE PENELITIAN

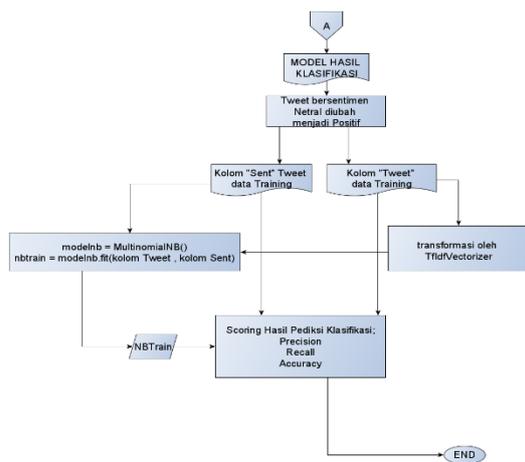
Klasifikasi teks dapat didefinisikan sebagai proses untuk menentukan suatu dokumen teks ke dalam suatu kelas tertentu (Siregar A. , 2018). Klasifikasi sentimen adalah proses menentukan suatu teks ke dalam kelas positif, netral, negatif. Menurut [8] Naïve Bayes adalah algoritma yang sering digunakan dalam pengkategorian teks. Ide dasarnya adalah menggabungkan probabilitas kata-kata dan kategori untuk memperkirakan probabilitas dari kategori sebuah dokumen. Textblob [9] adalah *lib python* untuk memproses data tekstual. Textblob memiliki fitur-fitur dasar Natural Language Processing (NLP), dalam penelitian kali ini fitur analisis sentimen digunakan untuk melakukan klasifikasi terhadap data testing .Data training merupakan inputan(masukan) yang akan diklasifikasi oleh *textblob*, sedangkan data testing merupakan inputan baru yang berbeda dari training yang akan diklasifikasi dengan data training [10].

Tahapan yang dilakukan pada penelitian kali ini adalah sebagai berikut.



Gambar 1. Kerangka kerja penelitian bagian 1

Pertama, proses pengumpulan data. Data yang akan dianalisa didapat dari twitter. Data diperoleh dengan menggunakan *lib tweepy*. Data yang sudah terkumpul nantinya akan dilakukan klasifikasi manual dan selanjutnya akan melalui *tahapan preprocessing*. Dalam penelitian ini juga memakai data kata baku yang diperoleh dari github [11]. Kedua, klasifikasi manual. Proses klasifikasi manual dilakukan oleh 3 user, yang masing-masing user memberikan klasifikasi pada setiap data tweet hasil dari proses crawling. Ketiga, *text prprocessing*. Pada tahap ini data tweet diolah dengan dibersihkan dari tanda baca, angka, *link url*, emoji, diubah menjadi *case folding*, *tokenizing*, dan menormalkan kata-kata slang agar menjadi kata baku. Keempat, klasifikasi data training dengan naïve bayes. Proses pengolahan data training dilakukan dengan menggunakan input data training yang telah dibuat pada tahap sebelumnya. Pada tahap ini digunakan *lib TextBlob* untuk melakukan klasifikasi data training. Keluaran dari proses ini adalah sebuah model data training yang selanjutnya akan diujikan pada data testing. Model yang terbentuk disimpan dalam format csv. Kelima, klasifikasi data testing dengan model training. Proses klasifikasi data testing dilakukan dengan menggunakan input model data training yang dihasilkan pada tahap sebelumnya. Maka sistem akan melakukan klasifikasi berdasarkan pola model yang terbentuk. Hasil dari proses ini adalah kategorisasi data testing kedalam sentimen positif dan negatif.



Gambar 2. Kerangka kerja Penelitian bagian 2

Keenam, perhitungan akurasi. Model data training yang didapat dari proses sebelumnya masih memiliki tweet yang bersentimen netral, maka dari itu ditahap ini tweet yang bersentimen netral diubah menjadi tweet positif. Data training kolom tweet dan kolom sentimen diubah menjadi list terlebih dahulu, lalu di latih oleh *MultinomialNB*, perhitungan akurasi dilakukan dengan *lib* yang dimiliki *sklearn*. Keluaran dari proses ini adalah prosentase akurasi dari data testing yang diklasifikasi oleh model training.

## HASIL DAN PEMBAHASAN

Data yang dikumpulkan merupakan data teks berupa tweet masyarakat kepada akun twitter @indihome dan @firstmediacares. Data tersebut dikumpulkan dengan sistem crawling yang dibangun dengan bahasa pemrograman python dengan memanfaatkan twitter API. Data yang dikumpulkan berjumlah 1400 tweet. Ditambahkan 157 data yang sudah terklasifikasi, hasil penelitian [6]. Tabel 1 menunjukkan data yang digunakan dalam penelitian ini.

Table 1. Rincian data yang digunakan

Jml Data	Sumber Data	Label		
		Pos	Neg	Net
356	@Firstmedia cares	105	166	85
660	@indihome	200	202	260
157	Dataset Hasil	135	3	19

	Penelitian [6]			
200	@indihome (untuk Testing set)	?	?	
200	@firstmedia cares (untuk Testing set)	?	?	

Dalam penelitian ini akan mencari sentimen positif dan negatif dari 200 data @indihome dan 200 data @firstmediacares, yang ditandai dengan tanda tanya (?) seperti tabel 1.

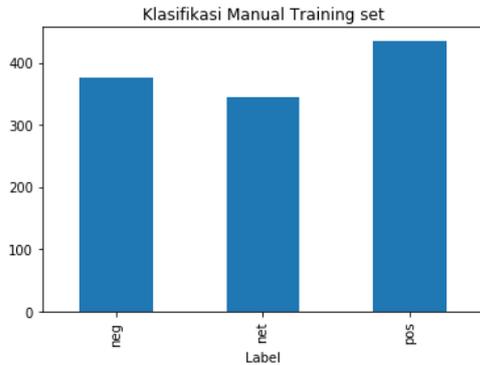
Data twitter yang dikumpulkan merupakan tweet dari pelanggan yang dipublish tanggal 10 Juli 2020, karena penelitian dimulai sekitar awal Juli dan tidak menggunakan akun premium twitter developer, maka crawling data hanya bisa mengambil tweet yang dipublish 7 sebelumnya dari tanggal dilakukannya proses crawling.

Tahapan didalam *text preprocessing* adalah mengubah bentuk tulisan menjadi *lower case*, menghapus kata "RT", menghapus username, *link url*, tanda baca, emoticon dan karakter-karakter yang tidak penting. Berikut contoh hasil *text preprocessing*.

Table 2. Hasil text preprocessing

Tweet sebelum pra proses teks	Tweet setelah pra proses teks
RT @AdamMaulud: Provider internet paling jelek menurut kalian apa guys? Kalo gw sih @IndiHome ,Kenapa? \r1.lemotnya gaada obat\r 2.keseringan...	provider internet paling jelek menurut kalian apa teman-teman kalau saya sih kenapa tidak ada obat
@FirstMediaCares hi admin, sepertinya skrng sudah ok <a href="https://t.co/Ws2wOyU5kz">https://t.co/Ws2wOyU5kz</a>	hai admin sepertinya sekarang sudah ok

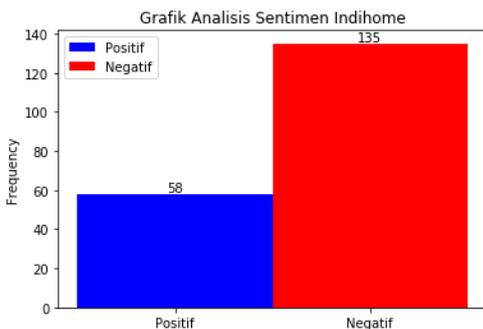
Setelah tahap text preprocessing, data yang dihasilkan dibagi menjadi 2 bagian, 1157 tweet untuk training dataset, 200 tweet @indihome dan 200 tweet @firstmediacares untuk testing dataset. Berikut kelas yang terbentuk didalam training dataset.



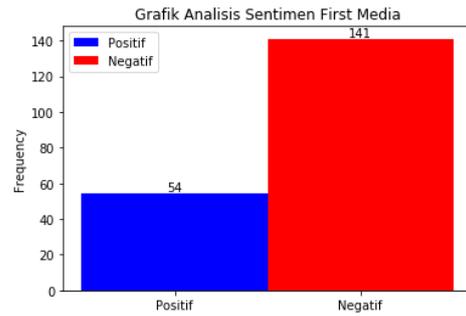
Gambar 3. Hasil klasifikasi manual

Data training tersebut kemudian di klasifikasi dengan metode *naïve bayes classifier*. *Lib Textblob* digunakan untuk melakukan klasifikasi naïve bayes. Data training yang terbentuk kemudian dijadikan model untuk analisis sentimen Indihome dan First Media.

Dari 200 dataset testing yang di lakukan analisis, diperoleh grafik sentimen masyarakat seperti gambar 4.

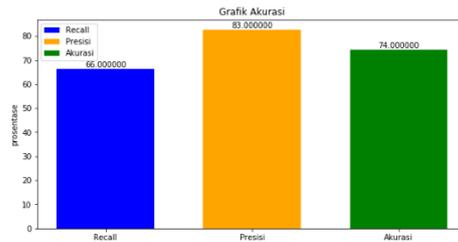


Gambar 4. Grafik sentimen terhadap Indihome



Gambar 5. Grafik sentimen terhadap First Media

Dari gambar 4 dan gambar 5 hasil analisis nya menunjukkan lebih banyak sentimen negatif dari masyarakat kepada kedua vendor. Analisis dilakukan dengan model yang telah terbentuk pada tahap sebelumnya. Berikut tingkat akurasi yang dihasil model untuk melakukan klasifikasi pada dataset Indihome dan First Media.



Gambar 6. Grafik tingkat akurasi, presisi, recall

Step *stop word removal* tidak diterapkan saat tahap *text preprocessing*, sehingga dihasilkan sentimen seperti gambar 4 dan gambar 5, dengan akurasi 74% seperti gambar 6.

Penambahan data training dilakukan untuk melihat perbandingan hasil akurasi training model dalam melakukan klasifikasi testing set. Tabel 3 menunjukkan hasil kenaikan akurasi yang sebanding dengan penambahan data.

Table 3. Hasil kenaikan akurasi

Jumlah Data Training	Proses Cleaning	Akurasi
1157 (percobaan pertama)	Ubah jadi <i>lower case</i> , menghapus kata "RT", menghapus	74%

Jumlah Data Training	Proses Cleaning	Akurasi
	<i>username, link url</i> , tanda baca, emoticon dan karakter-karakter yang tidak penting.	
1271	Ubah jadi <i>lower case</i> , menghapus kata "RT", menghapus <i>username, link url</i> , tanda baca, emoticon dan karakter-karakter yang tidak penting.	75%
1372	Ubah jadi <i>lower case</i> , menghapus kata "RT", menghapus <i>username, link url</i> , tanda baca, emoticon dan karakter-karakter yang tidak penting + dilakukan Stemmer menggunakan sastrawi stemmer	75%

Seperti yang terlihat di tabel 3, prosentase akurasi mengalami kenaikan seiring dengan penambahan training dataset, dikarenakan dalam penelitian kali ini metode yang digunakan adalah naïve bayes/ metode *mechine learning*, yang berarti dapat disimpulkan bahwa semakin banyak dataset yang digunakan akan semakin handal model tersebut dalam melakukan klsifikasi testing set.

## PENUTUP

Penelitian ini mencoba melakukan analisis sentimen pada twitter untuk mengetahui bagaimana sentimen

masyarakat terhadap indihome dan first media. Penelitian ini berfokus pada tweet berbahasa Indonesia. Dokumen yang digunakan dalam penelitian ini diperoleh dari akun resmi indihome @indihome dan first media (@firstmediacares) sebagai pembanding.

Hasil klasifikasi sentimen positif dan negatif dari model yang terbentuk untuk masing-masing vendor menunjukkan indihome mendapatkan respon negatif (135 respon) lebih besar daripada respon positif (58 respon), begitu pula dengan first media yang memperoleh 54 respon positif dan 141 respon negatif dengan nilai akurasi 74%, recall 66% dan presisi 83 %.

## DAFTAR PUSTAKA

- [1] Kompas, "https://tekno.kompas.com/read/2019/10/30/16062477/pengguna-aktif-harian-twitter-indonesia-diklaim-terbanyak," 30 September 2019. [Online]. Available: <https://tekno.kompas.com/read/2019/10/30/16062477/pengguna-aktif-harian-twitter-indonesia-diklaim-terbanyak>. [Accessed 06 2020].
- [2] O. Jagdale, V. Harmalkar, S. Chavan and N. Sharma, "Twitter Mining Using R," pp. 252-256, 2017.
- [3] B. Liu, "Sentiment analysis and subjectivity," Handbook of natural language processing, vol. 2, pp. 627--666, 2010.
- [4] E. Prasetyo, "Konsep dan Aplikasi Menggunakan Matlab," 2012.
- [5] M. Muljono, D. P. Artanti, A. Syukur, A. Prihandono and D. R. M. Setiadi, Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes, 2018.
- [6] S. Hanggara, T. M. Akhriza and M. Husni, Aplikasi Web Untuk Analisis Sentimen Pada Opini Produk Dengan Metode Naive Bayes Classifier, 2017.
- [7] A. Siregar, "KLASIFIKASI ALGORITMA TF DAN NEUTRAL NETWORK DALAM SENTIMEN ANALISIS," AIMS: Jurnal Accounting Information System, vol. 1, pp. 1-8, 2018.

- [8] W. Zhang and F. Gao, "An Improvement to Naive Bayes for Text Classification," *Procedia Engineering*, vol. 15, pp. 2160-2164, 2011.
- [9] Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh and E. Dempsey, "Simplified Text Processing".
- [10] R. Habibi, D. B. Setyohadi and E. Wati, "Analisis sentimen pada Twitter mahasiswa menggunakan metode backpropagation," *Jurnal Informatika*, vol. 12, 2016.
- [11] ramaprakoso. [Online]. Available: <https://github.com/ramaprakoso/analisis-sentimen/blob/master/kamus/kbba.txt>

..