

IMPLEMENTASI OPINION MINING UNTUK PROVIDER INTERNET MENGGUNAKAN METODE NAIVE BAYES.

Devin Abipraya¹⁾ Viny Christianti M²⁾ Novario Jaya P³⁾

^{1) 2) 3)} Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia

email: devinabipraya@gmail.com ¹⁾, email: viny@untar.ac.id ²⁾, email: novariojp@fti.untar.ac.id ³⁾

ABSTRAK

The development of information technology is growing from year to year. To support the smooth flow of information, there are many internet service providers circulating in Indonesia to support their needs. Some of the largest internet service providers in Indonesia such as Indihome, First media, and Biznet Home definitely have their own advantages and disadvantages. At this time, internet provider providers only accept customer complaints or suggestions through the customer service (CS) call center. Meanwhile, many young Indonesians currently use one of the popular Social Media services, namely Twitter as a user-friendly microblogging service so that users can easily use it, especially in delivering messages in the form of tweets. Therefore, a sentiment analysis program was designed for several internet providers in Indonesia. Opinions or Opinions will be analyzed to determine public sentiment. These sentiments will be classified into 3 sentiments, namely negative, positive, and neutral sentiments. The sentiment classification process can be done manually, but if there is too much data, it requires a system equipped with a classification method, so that the determination of classification can be done quickly. The design of this program applies the Naive Bayes Classifier method. Because this method is supervised learning, it requires training datasets with labels. Labeling will be done automatically using the K-means method. K-Means will sort tweets into groups which are divided into 3 labels. The results of the K-means clustering accuracy are 73.4%. The results of this application are divided into 2 parts, namely a pie chart which is divided into slices that describe the results of the percentage of tweet classifications and a table of classification results containing the number, content of the tweet, and the results of the classification. The best level of accuracy in testing

uses 220 training data, and 54 training data. The results of the accuracy of 83.3%.

1. Pendahuluan

Perkembangan Teknologi informasi semakin berkembang dari tahun ke tahun. Untuk mendukung kelancaran informasi, banyak provider layanan internet yang beredar di Indonesia untuk menunjang kebutuhan mereka. Beberapa penyedia layanan provider internet terbesar di Indonesia seperti Indihome, First media, dan Biznet Home pasti mempunyai keunggulan dan kekurangan tersendiri.

Sedangkan generasi muda Indonesia saat ini, banyak yang menggunakan salah satu layanan Social Media yang sedang populer yaitu Twitter sebagai layanan microblogging yang userfriendly sehingga pengguna mudah menggunakannya, Terutama dalam penyampaian pesan berupa tweet. Oleh karena itu, dirancanglah program analisis sentimen mengenai beberapa provider internet di Indonesia.

Opini atau Pendapat akan dianalisis untuk mengetahui sentimen masyarakat. Sentimen tersebut akan diklasifikasikan menjadi 3 sentimen yaitu sentimen negatif, positif, dan netral. Perancangan program ini menerapkan metode Naive Bayes Classifier.

Hasil dari aplikasi ini dibagi menjadi 2 bagian, yaitu diagram lingkaran (pie chart) yang dibagi menjadi irisan-irisan yang menggambarkan hasil presentase klasifikasi tweet dan tabel hasil klasifikasi yang berisi no, isi tweet, dan hasil klasifikasi. Tingkat akurasi terbaik dalam pengujian menggunakan data latih sebanyak 220 data, dan data latih sebanyak 54 data. Hasil keakuratan sebesar 83,3%.

2. Dasar Teori

Beberapa teori dan konsep yang digunakan dalam implementasi sistem analisis sentimen ini adalah Analisis Sentimen, *Preprocessing Text*, *Naive Bayes Classifier*, Klasterisasi (*K-Means*).

2.1 Analisis Sentimen

Analisis Sentimen adalah sebuah studi dalam menganalisis opini atau pendapat orang, penilaian, evaluasi, sikap, dan emosi terhadap suatu entitas seperti layanan, jasa, produk, organisasi, isu-isu, peristiwa dan topik. Tujuan utama dari analisis sentimen adalah untuk mengklasifikasi pendapat pengguna Twitter ke dalam suatu kelas. Contoh Tweet bisa dilihat pada Gambar 1.



Gambar 1. Contoh Tweet Provider Indihome.

2.2 Preprocessing Text

Tahap text preprocessing merupakan tahap awal dari text mining. Text preprocessing merupakan proses menggali, mengolah, mengatur informasi dengan cara menganalisis hubungannya, aturan-aturan yang ada di data tekstual semi terstruktur atau tidak terstruktur [1] Tahapan dalam text processing terdiri dari case folding, tokenizing, filtering, stopword removal dan stemming. Pada tahapan ini terdapat tahap tambahan yaitu convert emoticon. Berikut tahapan akan dijelaskan pada subab berikut:

1. Case Folding

Case Folding adalah tahap preprocessing task yang mengubah semua huruf dalam dokumen menjadi huruf kecil bertujuan untuk semua teks, hanya huruf 'a' sampai 'z'. Karakter selain huruf tersebut dihilangkan dan dianggap sebagai pembatas.

2. Tokenizing

Tokenizing adalah tahap yang menjadikan sebuah kalimat menjadi kata dengan cara memecah kalimat. Setiap potongan kata disebut juga dengan token. Proses ini melakukan penguraian deskripsi yang semula berupa kalimat berisi kata-kata dan tanda pemisah lain menjadi kata saja, baik merupakan kata penting maupun tidak penting.

3. Stopword Removal

Stopword adalah kata-kata yang sering muncul dalam jumlah besar dan tidak memiliki makna, sehingga dengan dibuangnya kata-kata tersebut ukuran kosakata menjadi berkurang dan harapannya sisa dari kata-kata memiliki bobot yang tinggi.

4. Stemming

Stemming adalah tahap mengembalikan kata menjadi kata dasar yang telah dihasilkan dari tahap stopword removal. Pada tahap ini kata akan dilakukan pemotongan kata imbuhan. Stemming pada bahasa Indonesia akan lebih kompleks dibandingkan dengan bahasa Inggris dikarenakan tidak ada rumus pasti.

2.2 Naive Bayes Classifier

Algoritma naive bayes adalah metode yang menggunakan prinsip probability untuk membuat model prediksi klasifikasi. Dengan memanfaatkan data tentang kejadian di masa lalu, model dapat membuat perkiraan apa yang terjadi di masa depan [2]. Metode ini pada prinsipnya menghitung seberapa tinggi kemungkinan satu model dalam suatu observasi masuk ke suatu kelompok atau kelas tertentu. Model klasifikasi yang memakai metode ini akan memanfaatkan training dataset untuk menghitung kemungkinan setiap kelas berdasarkan nilai-nilai feature yang di dalamnya. Metode ini merupakan conditional probability, yaitu kemungkinan munculnya suatu kejadian A bila suatu kejadian B muncul. Karena kejadian A bergantung kepada kejadian B, itu sebabnya disebut conditional (bersyarat).

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

Keterangan:

X: Data dengan class (label) yang belum diketahui.

H: Hipotesis data merupakan suatu class spesifik.

P(H|X) : Probabilitas hipotesis H bedasar kondisi X (posterior probability).

P(H) : Probabilitas hipotesis H (prior probability)

P(X|H) : Probabilitas X bedasar kondisi pada hipotesis H.

P(X) : Probabilitas X.

Pada Naive Bayes Classification setiap tweet direperesentasikan dalam pasangan atribut (a1, a2, a3, ..., an) dimana a1 adalah kata pertama a2 adalah kata kedua dan seterusnya, sedangkan V adalah himpunan kelas. Pada saat klasifikasi, metode ini akan menghasilkan kategori/kelas yang paling tinggi probabilitasnya (VMAP) dengan memasukkan atribut (a1, a2, a3, ..., an). Adapun rumus VMAP dapat dilihat pada persamaan 2 berikut:

$$V_{\text{map}} = \operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2, a_3, \dots, a_n | V_j) P(V_j)}{P(a_1, a_2, a_3, \dots, a_n)} \quad (2)$$

Karena $P(a_1, a_2, a_3, \dots, a_n)$ nilainya konstan untuk semua V_j sehingga persamaan diatas dapat ditulis dengan persamaan 3 berikut:

$$V_{\text{map}} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) \quad (3)$$

Setelah diperoleh perhitungan untuk masing-masing kategori, maka kategori yang dipilih adalah yang memiliki nilai V_{map} terbesar. Nilai $P(V_j)$ ditentukan pada saat pelatihan, yang nilainya berdasarkan persamaan:

$$P(v_j) = \frac{\text{docs}_j}{|\text{contoh}|} \quad (4)$$

Keterangan:

$P(V_j)$ = probabilitas setiap dokumen terhadap sekumpulan dokumen.

$|\text{docs}_j|$ = banyak dokumen yang memiliki kategori j dalam pelatihan.

$|\text{contoh}|$ = banyaknya dokumen dalam contoh yang digunakan saat pelatihan.

$$P(w_k | v_j) = \frac{n_k + 1}{n + |\text{kata}|} \quad (5)$$

Keterangan:

$P(w_k | V_j)$ = probabilitas kemunculan kata w_k pada suatu dokumen dengan kategori V_j .

n_k = frekuensi munculnya kata w_k dalam dokumen yang berkategori V_j .

n = banyaknya seluruh kata dalam dokumen berkategori V_j .

$|\text{kosakata}|$ = banyaknya kata dalam contoh pelatihan.

2.3 Klasterisasi (K-Means)

Klasterisasi merupakan metode yang digunakan untuk mengelompokkan data sehingga elemen yang serupa atau saling terkait di Cluster yang sama. Prinsip dasar dari K-Means adalah melakukan proses iteratif (berulang) untuk menggeser Centroid, yaitu suatu titik imajiner didalam setiap cluster agar letaknya tepat ada ditengah cluster. Proses penggeseran centroid diulangi berkali-kali hingga diperoleh posisi centroid yang stabil, kemudian kita anggap pembentukan cluster selesai. K adalah suatu konsanta yang menentukan jumlah centroid, yang berarti juga sama dengan jumlah cluster yang akan dibentuk. Terdapat beberapa metode untuk menghitung centroid terdekat salah satunya adalah Euclidian Distance. Tahapan pengelompokan data menggunakan metode K-Means adalah[3]:

1. Menentukan jumlah cluster
2. Mengelompokkan data sehingga terbentuk K buah cluster dengan titik centroid dari setiap cluster merupakan titik centroid yang telah dipilih sebelumnya.
3. Hitung pusat centroid dari data yang ada di masing-masing kelompok. Tempat centroid setiap

kelompok diambil dari rata-rata (mean) semua nilai data pada setiap fiturnya.

4. Alokasikan masing-masing data pada centroid terdekat.

Untuk mengukur jarak data ke pusat centroid dapat dilakukan dengan rumus Euclidean distance sebagai berikut:

$$\text{Euclidean Distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (6)$$

Keterangan:

(x_1, y_1) adalah koordinat titik awal dan (x_2, y_2) adalah koordinat titik tujuan.

Ulangi langkah 2-4 hingga nilai dari titik centroid tidak lagi mengalami perubahan

2.4 Purity K-Means

Purity digunakan untuk mengevaluasi hasil pengelompokkan oleh K-Means. Purity menghitung rasio antara jumlah dokumen yang pengelompokkannya benar dengan total dokumen yang dianalisis, nilai dari purity berkisar nol sampai satu. Semakin mendekati satu, maka semakin banyak dokumen yang telah sesuai atau benar pengelompokkannya dan sebaliknya. Jika nilai mendekati nol, maka menunjukkan semakin sedikit dokumen yang benar pengelompokkannya[4]. Rumus untuk mengukur akurasi model adalah sebagai berikut:

$$\text{purity}(\Omega, c) = \frac{1}{n} \sum_k^{\max} |w_k \cap C_j| \quad (7)$$

Keterangan:

Ω = $\{w_1, w_2, \dots, w_k\}$ adalah himpunan cluster

W_k = Himpunan dokumen dalam w_k

C = $\{c_1, c_2, \dots, c_j\}$ adalah himpunan kelas

C_j = Himpunan dokumen dalam C_j

N = Jumlah Data

\sum_j^{\max} = Banyak elemen yang sama

3 Rancangan dan Pembuatan

3.1 Rancangan Sistem

Sistem yang dirancang merupakan aplikasi berbasis website yang ditujukan untuk melakukan analisis sentimen terhadap komentar-komentar yang diambil dari sosial media (twitter). Pembuatan aplikasi ini bertujuan untuk memberikan informasi analisis sentimen pelanggan terhadap layanan provider internet.

Aplikasi ini menerima input berupa file csv hasil scraping komentar di sosial media dengan menggunakan 3rd party API. Sistem yang dibangun terdiri dari pengumpulan data, preprocessing, Klasterisasi (K-Means) dan klasifikasi data menggunakan Naive Bayes.

Penelitian ini diawali dengan pengumpulan tweet diambil secara live sebanyak 300 tweet. Tweet yang diambil merupakan 300 tweet terbaru mengenai penyedia layanan internet yaitu Indihome, First media, dan Biznet.

Setelah data terkumpul akan dilakukan proses text mining yaitu tahap preprocessing, dan pebelan kata akan dilakukan secara otomatis menggunakan klasterisasi K-means pada data training.

Sistem akan memberikan output suatu sentimen apakah opini bersifat negatif, netral, atau positif. Setelah seluruh data memiliki kelas masing-masing kemudian dilakukan pemisahan data menjadi dua bagian yaitu data training dan data testing. Data latih digunakan untuk mengklasifikasikan tweet pada kelas sentimennya memiliki jumlah data yang sama. Data uji digunakan untuk mengukur sejauh mana classifier berhasil melakukan klasifikasi dengan benar.

3.1.1 Perencanaan Sistem

Pada tahap perencanaan sistem, ditentukan tujuan, batasan rancangan, dan metode yang akan digunakan pada aplikasi analisis sentimen. Tujuan perancangan sistem ini untuk memberikan informasi mengenai sentimen pengguna terhadap layanan provider internet. Batasan yang digunakan dalam perancangan sistem ini yakni data masukan berupa file csv hasil scraping dengan menggunakan 3rd party API. Sehingga, dalam proses pengujian tidak dilakukan pengumpulan data kembali. Hal ini dilakukan untuk mencegah kemungkinan kegagalan saat melakukan scraping dengan 3rd party API. Data yang digunakan sebagai input sistem merupakan data yang berkaitan dengan komentar pengguna internet provider di Indonesia.

Metode yang digunakan sistem untuk melakukan analisis sentimen adalah Naive Bayes Classifier dengan metode evaluasi Confusion Matrix dan Akurasi. Proses klasifikasi akan sepenuhnya menggunakan bahasa pemrograman Python. Hal ini dikarenakan, Python memiliki library yang cukup lengkap yang dapat membantu dalam text processing.

3.1.2 Analisis Sistem

Pada tahap analisis sistem, dilakukan analisis perangkat keras dan perangkat lunak yang

dibutuhkan untuk pengembangan aplikasi sentimen analisis. Spesifikasi perangkat keras yang digunakan yaitu:

- Processor
- Memori
- Hard Drive
- Graphics
- Layar

Spesifikasi perangkat lunak yang digunakan yaitu:

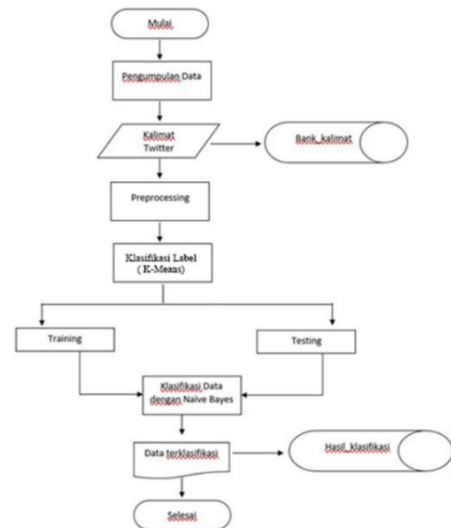
- Sistem operasi Windows 10
- Notepad++ 7.9.3
- Python 3.7
- XAMPP

3.1.3 Perancangan Sistem

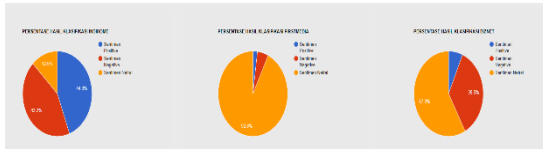
Pada tahap perancangan sistem, dilakukan perancangan sistem sentiment analisis sebelum nantinya dilanjutkan ke tahap pengembangan sistem. Pada tahap ini, akan dibuat diagram alur sistem, rancangan diagram hirarki, rancangan basis data, dan rancangan tampilan antarmuka.

3.1.3.1. Rancangan Diagram Alur

Diagram alur sistem digunakan untuk menggambarkan alur proses yang terjadi pada sistem. Hal ini dilakukan, untuk memudahkan programmer untuk memahami alur proses sistem yang akan dibuat. Diagram alur sistem analisis sentimen dapat dilihat pada Gambar 2.



Gambar 2. Diagram Hirarki



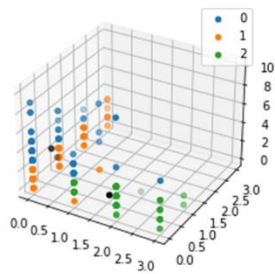
Gambar 9. Modul Hasil Presentasi Klasifikasi

4 Pengujian dan Pembahasan

Pengujian Rancangan yang digunakan untuk mengembangkan aplikasi ini adalah *Blackbox Testing*. *Blackbox Testing* adalah pengujian yang dilakukan hanya mengamati hasil eksekusi melalui data uji dan memeriksa beberapa fungsional dari perangkat lunak. *Blackbox Testing* hanya mengevaluasi penampilan luar / tampilan interface saja tanpa mengetahui yang terjadi pada proses detailnya. Selanjutnya, Pengujian yang dilakukan terhadap data adalah akurasi dari klasifikasi sentimennya. Data tweet yang akan diuji sebanyak 337 tweet. Tweet tersebut diambil dari 3 Provider Internet yaitu Biznet, First Media, dan Indihome. Akurasi dari klasterisasi *K-means* akan dievaluasi menggunakan metode *Purity K-Means* dan Klasifikasi *Naive Bayes* menggunakan *Confusion Matrix*.

4.1 Pengujian Data Klasterisasi *K-Means*

Pengujian terhadap data akan dibagi menjadi 3 kelas cluster yaitu positif, negatif, dan netral. Pada setiap cluster terdapat titik pusat (centroid) yang mempresentasikan cluster tersebut, jarak antara data dan pusat cluster akan dihitung menggunakan Euclidean Distance untuk menghitung jarak semua data ke titik pusat cluster. Proses penentuan pusat cluster dan penempatan data dalam cluster akan diulangi sampai nilai centroid tidak berubah lagi. Hasil *K-means* dapat dilihat pada Gambar 10.



Gambar 10. Hasil Klasterisasi *K-Means*

Dari Hasil *K-Means* tersebut menghasilkan 3 cluster sentimen, yaitu cluster 0 terdapat 93 Sentimen, Cluster 1 berisi 127 sentimen dan Cluster 2 terdiri dari 77 Sentimen. *K-Means* ini

akan digunakan untuk pelabelan secara otomatis pada data, sehingga penulis tidak perlu melakukan pelabelan tiap kelas secara manual untuk proses klasifikasi pada metode *Naive Bayes Classifier*.

4.1.1 Pengujian Evaluasi *Purity K-Means*

Pengujian evaluasi *purity k-means* pada program ini dilakukan untuk mengukur keakuratan hasil klaster yang telah dihasilkan pada tahap clustering. *Purity test* yang dilakukan dengan data sebanyak 297 sentimen dengan $M=3$ sebesar 0,73 menggunakan persamaan rumus (7). Hasil *purity test* yang sudah dilakukan dapat dilihat pada Tabel 1.

Tabel 1. Hasil *Purity K-Means*

	Cluster 0	Cluster 1	Cluster 2	Purity
Cluster 0	58	21	14	0,623
Cluster 1	31	86	10	0,677
Cluster 2	3	0	74	0,961
Hasil Purity				0,734

Hasil pengujian *purity test* yang sudah dilakukan menggunakan data set sebanyak 297 Data dengan mengkombinasikan parameter M sebagai titik pusat cluster menghasilkan nilai *purity* terbaik sebesar 0,73 dengan kombinasi M sebesar 3. Sebanyak 218 dataset telah benar pengelompokkannya dari 297 Data. Hasil pengujian mendekati angka 1 yang mengindikasikan bahwa semakin banyak dokumen yang telah sesuai atau benar pengelompokkannya.

Dari hasil pengujian *purity* yang sudah dilakukan dapat disimpulkan bahwa pengelompokkan terbaik dengan menggunakan data set sebanyak 297 data dengan $M=3$ sebesar 0,73. Hasil klasterisasi sentimen terbagi menjadi 3 cluster adalah sebagai berikut:

- a. Pada Cluster 0 memiliki titik pusat cluster kata netral dan jumlah sentimen pada cluster ini sebanyak 93 Data.
- b. Pada Cluster 1 memiliki titik pusat cluster kata negatif dan jumlah sentimen pada cluster ini sebanyak 127 Data.
- c. Pada Cluster 2 memiliki titik pusat cluster kata positif dan jumlah sentimen pada cluster ini sebanyak 77 Data.

4.1.2 Evaluasi *Confusion Matrix*

Untuk Pengujian *Confusion Matrix* menggunakan semua data dan akan diuji sebanyak 3 kali, dengan skenario pembagian data training

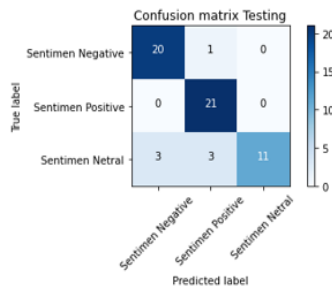
dan data testing. Untuk tabel penjujian Confusion Matrix dapat dilihat pada Tabel 2.

Tabel 2. Tabel Pengujian *Confusion Matrix*

L	Data Traning	Data Testing	Proporsi Pembagian
1	238	59	80%/20%
2	208	89	70%/30%
3	179	118	60%/40%

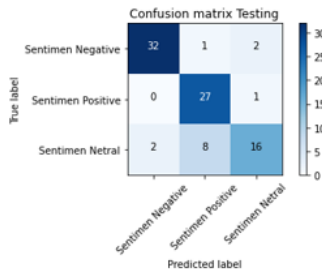
Hasil Confusion Matrix pada pengujian pertama dapat dilihat pada Tabel 3. Sumbu X adalah Hasil prediksi dan Y adalah label yang seharusnya

Tabel 3. Pengujian *Confusion Matrix* ke-1



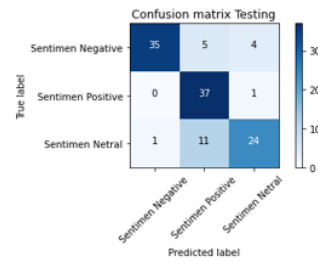
Pengujian ke-1 didapatkan hasil akurasi sebesar 88,13%, dengan data training sebanyak 238 data, dan data testing sebanyak 59 data. Selanjutnya hasil untuk skenario pengujian confusion matrix ke-2, dapat dilihat pada Tabel 4.

Tabel 4. Pengujian *Confusion Matrix* ke-2



Pengujian ke-2 didapatkan hasil akurasi sebesar 84,26%, dengan data training sebanyak 208 data, dan data testing sebanyak 89 data. Selanjutnya hasil untuk skenario pengujian confusion matrix ke-3, dapat dilihat pada Tabel 5

Tabel 5. Pengujian *Confusion Matrix* ke-3



Pengujian ke-3 didapatkan hasil akurasi sebesar 81,35%, dengan data training sebanyak 179 data, dan data testing sebanyak 118 data. Dari hasil pengujian dari 3 skenario diatas, didapatkan hasil terbaik diperoleh oleh pengujian pertama dengan pembagian proporsi data training 238 data dan data testing 59 data, didapatkan hasil akurasi sebesar 88,13% berdasarkan pengujian menggunakan confusion matrix. Pengujian ke-2 mendapatkan hasil akurasi sebesar 84,26% dan untuk pengujian ke-3 mendapatkan hasil terendah dengan hasil akurasi 81,35%.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang didapat setelah dilakukannya tahap pengujian terhadap sistem aplikasi analisis opini mining untuk provider internet dengan metode Naive Bayes adalah sebagai berikut:

1. Aplikasi ini dapat melakukan klasifikasi data training secara otomatis menggunakan klusterisasi k-means.
2. Metode Naive Bayes Classifier dapat menganalisis sentimen pelanggan terhadap beberapa provider internet pada Twitter.
3. Aplikasi yang telah dibuat dapat mengklasifikasikan sentimen menjadi 3 label yaitu positif, negatif, dan netral.
4. Hasil akurasi terbaik pada tahap pengujian menggunakan metode Purity K-Means sebesar 0,73 dan metode Naive Bayes pada pengujian ke-1 sebesar 88,13%.
5. Hasil pengujian pada modul-modul pada aplikasi ini berjalan dengan baik tanpa adanya error/bug.

5.2 Saran

Beberapa saran untuk program sistem aplikasi analisis opini mining untuk provider internet dengan metode Naive Bayes adalah sebagai berikut:

1. Memperbanyak kata dasar untuk mendukung fitur K-Means.

2. Kata dasar ditambah dengan kata dasar Bahasa Inggris, dikarenakan beberapa tweet menggunakan bahasa Inggris yang tidak dapat diklasifikasikan oleh fitur K-Means.

3. Melakukan analisa sentimen berdasarkan makna semantik dalam kalimat.

REFERENSI

[1] Navi Atri Lestari, Tubagus Mohammad Akhriza, and Eka Yuniar, 2020, "Metode Naïve Bayes Classifier Dengan Textblob Untuk Analisis Sentimen Terhadap Pelayanan Indihome Dan First Media", Seminar Nasional Teknologi Informasi Dan Komunikasi STI&K (SeNTIK).

[2] Fajar Ratnawati, 2018, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter", Riau.

[3] Muhammad Sholeh Hudin, M Ali Fauzi, Sigit Adinugoro, 2018, " Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi", Malang.

[4] Ika Kurnia Saputri, Tedy Setiadi, Lisna Zahtorun, 2018, "Penerapan Text Mining Pengelompokkan Judul Kerja Praktek Menggunakan Metode K-Means Clustering dengan Cosine Similarity", Yogyakarta.

Devin Abipraya, mahasiswa S1, program studi Teknik Informatika, Fakultas Teknologi Informasi Universitas Tarumanagara.

Viny Christanti M., M.Kom, memperoleh gelar S.Kom dari Universitas Tarumanagara, Jakarta tahun 2004 dan gelar M.Kom dari Universitas Indonesia, Depok tahun 2008. Saat ini sebagai Staf Pengajar program studi teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara.

Novario Jaya Perdana, S.Kom., M.T memperoleh gelar S.Kom dari ITS Jakarta tahun 2011 dan gelar M.T dari Universitas Indonesia tahun 2016