ANALISIS SENTIMEN TERHADAP *BRAND SKINCARE* LOKAL MENGGUNAKAN *NAÏVE BAYES CLASSIFIER*

Kaswili Sriwenda Putri[1)], Iwan Rizal Setiawan[2)], Agung Pambudi[3)]
Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi
Email: swendap1115@gmail.com
Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi
Email: myfrank5150@gmail.com
Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi
Email: agungpambd@ummi.ac.id

Abstrak

Sejalan dengan perkembangan teknologi informasi yang makin meningkat serta akses internet yang semakin mudah, banyak masyarakat yang menyuarakan opini mereka di media sosial salah satunya yaitu Twitter. Salah satu isu atau topik yang sering dibahas di twitter adalah mengenai perawatan kulit atau *skincare* terutama dalam mengulas produk-produk *skincare* dari suatu *brand*. Ulasan serta opini terhadap *brand skincare* terutama *brand skincare* lokal seperti Avoskin, Azarine dan Somethinc di twitter dijadikan sumber data untuk mengetahui persepsi masyarakat terhadap *brand skincare* lokal tersebut. Data *tweet* yang digunakan dibagi kedalam 3 dataset berdasarkan ulasan terhadap *brand* yang dituju yaitu Avoskin, Azarine dan Somethinc. Untuk mendapatka hasil yang jelas, maka dilakukan proses klasifikasi. Algoritma yang dapat digunakan untuk mengklasifikasikan suatu data salah satunya adalah *naïve bayes classifier* dengan mengklasifikasikan data kedalam 2 jenis, yaitu positif dan negatif. Proses klasisfikasi yang menggunakan *naïve bayes classifier* ini menghasilkan nilai akurasi sebesar 79% untuk dataset Avoskin, 78% untuk dataset Azarine dan 75% untuk dataset Somethinc. Sedangkan pengujian dengan *k-fold cross validation* menghasilkan nilai sebesar 79% untuk dataset Avoskin serta Somethinc dan 78% untuk dataset Azarine.

Kata kunci— *Twitter, Brand,* Analisis Sentimen, *Naive Bayes Classifier*

PENDAHULUAN

Sejalan dengan perkembangan teknologi yang setiap waktunya terus meningkat, akses internet juga semakin mudah serta biaya aksesnya murah, menjadikan penggunaan media sosial pun secara global semakin meningkat seiring berjalannya waktu. Di Indonesia sendiri, sebanyak 63 jiwa penduduknya menggunakan internet dengan penggunaan untuk mengakses social media sebesar 95% (Kominfo, 2022). Hal ini dikarenakan pengguna pada media sosial memungkinkan mereka untuk mengekspesikan diri, berkomunikasi dengan sesama pengguna media sosial lainnya bahkan bertukar informasi secara *real time* (Nasrullah, 2015). Salah satu media sosial yang banyak digunakan oleh masyarakat Indonesia yaitu Twitter. Dengan jumlah pengguna sebanyak lebih dari 19,5 juta serta menempati urutan ke 5 sebagai pengguna terbanyak di serluruh dunia (Kominfo, 2022).

Twitter adalah media sosial yang memungkinkan penggunanya melakukan komunikasi dalam pesan singkat dengan jumlah karakter maksimal 280 karakter. Pengguna twitter biasanya menumpahkan apa yang mereka rasakan kedalam sebuah cuitan. Di tambah dengan fitur-fiturnya yang lebih komprehensif, menjadikannya banyak digunakan masyarakat baik itu perorangan maupun institusi-institusi lainnya yang menjadikan twitter sebagai riset pasar mereka. Salah satu fitur yang ada di twitter yaitu *trending*, dimana fitur ini menampilkan topik atau isu yang banyak atau sedang populer. Hal ini memungkinkan pengguna untuk melihat hal apa saja yang sedang menjadi berita terpanas di seluruh dunia atau di wilayah tertentu. Kemudahan interaksi pada twitter ini

menjadikannya salah satu media sosial yang akrab dengan masyarakat Indonesia.

Salah satu topik yang hampir setiap hari dibahas di twitter yaitu mengenai perawatan kulit atau *skincare*. Bahasan yang tidak habisnya mulai dari tahapan perawatan kulit yang baik dan benar hingga ulasan mengenai *brand* dari produk skincare. Ketertarikan dan kesadaran masyarakt terhadap perawatan kulit ini menarik perhatian beberapa individual atau perusahaan lokal untuk membuat suatu *brand* yang menawarkan produk *skincare* yang mengandung banyak manfaat disertai dengan harga yang terjangkau. Hal ini juga membuat beberapa *brand* lokal banyak diperbincangkan di media social terutama twitter. mulai dari kualitas produk tersebut, promosi yang dilakukan, hingga harga yang apakah sebanding dengan hasil yang diberikan.

Ulasan serta opini terhadap beberapa *brand skincare* lokal ini dijadikan sebagai objek penelitian dengan tujuan untuk melihat kecenderungan apakah bersentimen positif atau negatif. Agar hasil yang diinginkan jelas, maka dilakukan proses klasifikasi yang merujuk pada proses analisis sentimen. Analisis sentimen sendiri merupakan proses implementasi teknologi pemrosesan Bahasa alami, linguistic komputasi dan analisis teks untuk mengenali pendapat subyektif dalam suatu dokumen.

Untuk mendapatkan gambaran dalam memecahkan masalah yang sedang diteliti ini, dilakukan analisis terhadap beberapa penelitian serupa yang dimana akan dijadikan sebagai tolak ukur untuk penelitian yang akan dilakukan. Layaknya pada penelitian yang dilakukan oleh Miranti Alysha Zula Larasati dan kawan-kawan pada jurnal yang berjudul "Penerapan Metode *K-Means Clustering* dalam menganalisis Sentimen Masyarakat Terhadap K-Popers Pada Twitter" yang dimana pada penelitian ini menghasilkan nilai silhouette sebesar 0.687974 (Larasati et al., 2022). Sedangkan penelitian yang dilakukan oleh Yonathan Sari Mahardhika dan Eri Zuliarso dengan judul "Analisis Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma *Naïve Bayes Classifier*" menghasilkan nilai akurasi sebesar 79% (Mahardika & Zuliarso, 2018).

Atas pertimbangan dari dua penelitian tersebut, peneliti menyimpulkan algoritma yang dapat digunakan untuk mengatasi permasalahan di penelitian ini adalah *naïve bayes classifier* yang bekerja berdasarkan probabilitas kemunculan kata-kata dari suatu teks.

METODE PENELITIAN

2.1 Pengumpulan Data

Dikarenakan data yang digunakan bersumber dari twitter, maka dilakukan proses *crawling* untuk proses pengambilan datanya. data yang digunakan adalah *tweet* yang mengandung komentar atau *review* terhadap *brand skincare* lokal Avoskin, Azarine dan Somethinc. Dimana nantinya data tersebut akan disimpan kedalam *dataset* yang berbeda-beda berdasarkan *brand* yang dimaksud pada *tweet* tersebut.

2.2 Pelabelan Data

Selanjutnya pelabelan data yang dilakukan terhadap data latih ini memberikan label positif atau negatif pada teks dokumen.

2.3 Pre-Processing

Tahapan ini bertujuan untuk mengganti data yang tidak terstruktur menjadi terstuktur (Muel, 2020). Proses ini adalah tahapan yang paling penting dalam proses analisis sentimen karena semakin terstruktur data yang digunakan maka akan prediksi yang akan dihasilkan akan semakin akurat. Tahapan dari proses *preprocessing* adalah sebagai berikut.

2.3.1 Cleaning

Tahapan ini bertujuan untuk menghilangkan data duplikat, *missing value*, data tidak valid atau *noise* terhadap data. Selain itu dihilangkan atribut-atribut yang tidak akan berpengaruh terhadap hasil klasifikasi seperti *mention*, *hastag* dan *link* (Firmansyah & Puspitasari, 2021).

Tabel 1. Contoh *Cleaning*

Sebelum *cleaning*	**Sesudah *cleaning***
@ohmybeautybank Aku pertama coba exfo toner avoskin, dan emang	Aku pertama coba exfo toner

nampol banget. https://t.co/krvXNsaOiE	avoskin, dan emang

2.3.2 Case Folding

Proses berikutnya adalah *case folding*, proses ini bertujuan untuk mengkonversi huruf kapital menjadi huruf kecil.

Tabel 2. Contoh *Case Folding*

Sebelum *case folding*	**Sesudah *case folding***
Aku pertama coba exfo toner avoskin, dan emang nampol banget.	aku pertama coba exfo toner avoskin, dan emang nampol banget.

2.3.3 Tokenizing

Kemudian, tahapan berikutnya yaitu tokenizing yang bertujuan untuk membagi teks menjadi potongan kalimat.

Tabel 3. Contoh *Tokenizing*

Sebelum *tokenizing*	**Sesudah *tokenizing***
aku pertama coba exfo toner avoskin emang nampol banget.	'aku', 'pertama', 'coba', 'exfo', 'toner', 'avoskin', 'emang', 'nampol', 'banget'

2.3.4 Stopword

Pada proses ini akan dihilangkan kata yang tidak bermakna atau tidak berpengaruh terhadapa hasil akurasi nantinya seperti kata penghubung.

Tabel 4. Contoh *Stopword*

Sebelum *stopword*	**Sesudah *stopword***
'aku', 'pertama', 'coba', 'exfo', 'toner', 'avoskin', 'emang', 'nampol', 'banget'	'coba', 'exfo', 'toner', 'avoskin', 'emang', 'nampol', 'banget'

2.3.5 Slangword

Tahapan berikutnya yaitu slangword, yang dimana proses ini bertujuan untuk menggati kata-kata salng kedalam bentuk yang sebenarnya.

Tabel 5. Contoh *Slangword*

Sebelum *slangword*	**Sesudah *slangword***
bgt	banget
abis	Habis
bgs	Bagus
pake	pakai
lucuk	lucu

2.3.6 Stemming

Tahapan terakhir dalam proses preprocessing yaitu stemming, dimana proses ini bertujuan untuk menghilangkan sufiks dan prefis dan mengubahnya kedala bentuk dasarnya.

Tabel 6. Contoh *Stemming*

Sebelum *stemming*	**Sesudah *stemming***
azarine bagus nahan minyak banget tapi kurang lembab di kulitku	azarine bagus nahan minyak banget tapi kurang lembab di kulit

2.4 Pembobotan Kata

Proses ini dilakukan dengan memanfaatkan metode TF-IDF yang merupakan metode pembanding untuk proses pembobotan kata dengan menghitung frekuensi kemunculan suatu kata dalam sebuah dokumen serta frekuensi kebalikan dari dokumen yang menganding kata tersebut (Andriani & Wibowo, 2021). persamaan untuk menghitung TF-IDF adalah sebagai berikut.

$$W_{dt} = TF_{dt} * IDF_{ft} \qquad (1)$$

Keterangan:

W_{dt} : bobot dokumen ke-d terhadap kata ke-t

TF_{dt} : banyaknya kata yang dicari pada sebuah dokumen

IDF_{ft} : Inverse Document Frequency $log\frac{N}{df}$

2.5 Pemodelan Algoritma

Pemodelan algoritma pada penelitian ini menggunakan *naïve bayes classifier* yang dimana pada proses klasifikasinya berdasarkan *teorema bayes*. *Teorema bayes* sendiri mengasumsikan setiap variabel adalah idependen. Data yang digunakan dalam *naïve bayes classifier* dibagi menjadi dua jenis data yaitu data latih dan data uji. data latih merupakan data yang digunakan untuk menentukan probabilitas yang aka terjadi, sedangkan data uji merupakan data yang dipakai dalam melakukan prediksi dari probabilitas yang terbentuk (SETYAWATI, 2020).

Persamaan yang menyatakan suatu probabilitas atau peluang bersyarat adalah sebagai berikut (Suyanto, 2019).

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (4)$$

Keterangan:

X	:	Bukti
H	:	Hipotesis
P(H|X)	:	Posterior probabilitas
P(H)	:	Prior probabilitas
P(X|H)	:	Probabilitas X berdasarkan pada kondisi hipotesis H
P(X)	:	Probabilitas X

2.6 Evaluasi

Tahapan evaluasi dilakukan dengan tujuan agar mengetahui tingkat akurasi performa dari algoritma *naïve bayes classifier* dengan menggunakan *confusion matrix* dan *K-fold Cross Validation.*

2.6.1 Confusion Matrix

Confusion matrix digunakan sebagai mengevaluasi kinerja dari algoritma *naïve bayes classifier* melalui pengukuran dan pengujian. Pada matriks yang berukuran N x N ini, nilai akurasi digunakan untuk variable uji yang diperoleh dari tabel matriks berikut (Firmansyah & Puspitasari, 2021).

Confusion Matrix		Kelas Aktual	
		Positif	Negatif
Kelas Prediksi	Positif	TP	FP
	Negatif	FN	TN

Gambar 1. Tabel *Confusion Matrix*

Dengan keterangan sebagai berikut

1. TP : *True Positive*, data yang diperdikasi positif dan benar-benar positif.
2. TN : *True Negative*, data yang diprediksi negatif dan benar-benar negatif.
3. FP : *False Positive* data yang diprediksi positif akan tetapi ternyata negatif.
4. FN : *False Negative*, data yang diprediksi negatif akan tetapi ternyata positif.

2.6.2 K-fold Cross Validation

k-fold Cross Validation dipakai untuk mengukur kinerja naïve bayes classifier dengan membagi sampel data secara acak sebanyak nilai K. hal ini dilakukan untuk mengestimasi kesalahan dalam proses prediksinya (Patro, 2021).

HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data yang berupa *tweet* menggunakan bahasa Indonesia dengan kata kunci Avoskin, Azarine serta Somethinc. Pengambilan datanya dilakukan melalui proses *crawling* menggunakan Bahasa pemrograman *python*. Pengambilan data dilakukan dari bulan Januari sampai dengan Maret 2023, dengan jumlah data yang didapatka adalah sebanyak 4500 data yang dibagi kedalam 3 dataset berdasarkan kata kunci pencariannya, dengan masing-masing dataset berjumlah 1500 data.

3.2 Pelabelan Data

Setelah data terkumpul, data pada masing-masing dataset akan dibagi kedalam 2 bagian. Yang dimana 80% dari data keseluruhan akan digunakan sebagi data latih dan sisanya yang sebagai 20% akan digunakan sebagai data uji. pada data latih disetiap dataset akan diberi label. Terdapat 2 kategori label yang digunakan yaitu label positif dan negatif yang dimana proses pelabelan ini dilakukan secara manual. Berikut rincian jumlah data yang berlabel positif dan negatif disetiap dataset.

Tabel 7. Jumlah Data Latih

No	Dataset	Jumlah Data	
		Positif	Negatif
1	Avoskin	919	281
2	Azarine	924	276
3	Somethinc	872	328
Jumlah Keseluruhan		2715	885

3.3 Preprocessing

Proses *preprocessing* yang diawali dengan tahap *cleaning*, lalu tahap *case folding, tokenizing, stopword, slangword* hingga *stemming* ini dilakukan dengan menggunakan Bahasa pemrograman *python* dengan beberapa *library* yang dipakai seperti gambar berikut.

Gambar 2. *Library Python* untuk *Preprocessing*

Hasil dari proses *preprocessing* ini akan disimpan dalam *file* CSV. Serta berikut data yang telah selesai dilakukan proses *preprosessing* pada salah satu dataset.

Gambar 3. Hasil *Preprocessing*

3.4 Pembobotan Kata

Pengimplementasian perhitungan TF-IDF pada kode program *python* dengan menggunakan *library sklearn.feature_extraction.text* adalah sebagai berikut.

Gambar 4. Library TF-IDF

3.5 Pemodelan Algoritma

Setelah proses pembobotan dengan TF-IDF selesai, kemudian dilakukan proses klasifikasi menggunakan *naïve bayes classifier*. Dimana data akan dibagi menjadi 2 bagian yaitu sebagai data latih sebanyak 80% dan data uji sebanyak 20%. Data yang telah selesai dilatih, modelnya akan disimpan dalam bentuk *pickle* yang nantinya data itu akan dimuat kembali untuk melakukan proses klasifikasi pada data uji. implementasi proses klasifikasi menggunakan *naïve bayes classifier* dalam bentuk kode program dapat dilihat pada gambar berikut.

Gambar 5. Kode Program *Naïve Bayes Classifier*

3.6 Evaluasi

3.6.1 Confusion Matrix

Analisis sentimen yang dihasilkan menggunakan *naïve bayes classifier* di evaluasi dengan menghitung presisi, *recall* dan *F1-score*. Hal ini dilakukan untuk melihat akurasi performa dari algoritma *naïve bayes classifier*. Tahapan ini dilakukan dengan menggunakan *confusion matrix* dengan hasilnnya pada masing-masing dataset dapat dilihat pada tabel berikut.

Tabel 8. Hasil *Confusion Matriks*

Dataset	*Precision*		*Recall*		*F1-Score*	
	P	N	P	N	P	N
Avoskin	0.79	1.00	1.00	0.04	0.89	0.08
Azarine	0.79	1.00	1.00	0.04	0.88	0.07
Somethinc	0.75	1.00	1.00	0.09	0.86	0.17

3.6.2 K-fold Cross Validation

Selanjutnya dilakukan pengujian ulang menggunakan *K-fold Cross Validation* dengan tujuan agar hasil uji serta evaluasi kinerja algoritma memperoleh hasil yang maksimal.

Pada penelitian ini menggunakan nilai k sebanyak 10 pada masing-masing dataset dengan rata-rata hasil pengujiannya dapat dilihat pada tabel berikut.

Tabel 9. Hasil *K-fold Cross Validation*

Brand	**Akurasi**	**presisi**	***recall***	***F1-score***
Avoskin	79%	85%	57%	57%
Azarine	78%	88%	56%	55%
Somethinc	79%	83%	58%	59%

KESIMPULAN

Dari penelitian yang telah dilakukan, dapat disimpulkan bahwa ketiga brand tersebut dengan data yang diambil dari bulan Januari hingga Maret 2023 ini memiliki hasil klasifikasi bersentimen positif, dengan presentase sebanyak 76.6% untuk brand avoskin, 77% untuk brand avoskin dan 72.6% untuk brand Somethinc. Dengan tingkat akurasi penerapan algoritma *naïve bayes classifier* menggunakan *confusion matrix* untuk masing-masing *brand* adalah sebesar 79% untuk *brand* Avoskin, 78% untuk *brand* Azarine dan 75% untuk brand Somethinc. Sedangkan dengan menggunakan *k-fold cross validation*, rata-rata hasil akurasi dengan k sebanyak 10 adalah sebesar 79% untuk *brand* avoskin dan somethinc serta 78% untuk *brand* azarine.

REFERENSI

Andriani, N., & Wibowo, A. (2021). Implementasi Text Mining Klasifikasi Topik Tugas Akhir Mahasiswa Teknik Informatika Menggunakan Pembobotan TF-IDF dan Metode Cosine Similarity Berbasis Web. *Senamika*, *September*, 130–137. https://conference.upnvj.ac.id/index.php/senamika/article/view/1807%0Ahttps://conference.upnvj.ac.id/index.php/senamika/article/download/1807/1350

Firmansyah, Z., & Puspitasari, N. F. (2021). Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Berdasarkan Opini Pada Twitter Menggunakan Algoritma Naive Bayes. *Jurnal Teknik Informatika*, *14*(2), 171–178. https://doi.org/10.15408/jti.v14i2.24024

Kominfo. (2022). *Kominfo : Pengguna Internet di Indonesia 63 Juta Orang*. https://www.kominfo.go.id/index.php/content/detail/3415/Kominfo+:+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker

Larasati, M. A. Z., Winarsih, N. A. S., Rohman, M. S., & Saraswati, G. W. (2022). Penerapan Metode K-Means Clustering Dalam Menganalisis Sentimen Masyarakat Terhadap K-Popers Pada Twitter. *Progresif: Jurnal Ilmiah Komputer*, *18*(2), 201. https://doi.org/10.35889/progresif.v18i2.877

Mahardika, Y. S., & Zuliarso, E. (2018). Analisis Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma Naives Bayes. *Prosiding SINTAK 2018*, *2015*, 409–413.

Muel, S. S. (2020). *ANALISIS SENTIMEN PADA TWITTER GOJEK DENGAN METODE NAÏVE BAYES CLASSIFIER MENGGUNAKAN VISUALISASI LATENT DIRICHLET ALLOCATION*. UNIVERSITAS MUHAMMADIYAH SEMARANG.

Nasrullah, R. (2015). *Media Sosial; Persfektif Komunikasi, Budaya dan Sosioteknologi*. Simbiosa Rekatama Media.

Patro, R. (2021). *Cross-Validation: K Fold vs Monte Carlo*. Medium. https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

SETYAWATI, I. (2020). *IMPLEMENTASI DATA MINING DENGAN ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI TKP KRIMINALITAS DI KABUPATEN PONOROGO*. UNIVERSITAS MUHAMMADIYAH PONOROGO.

Suyanto, D. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data* (Revisi). Informatika Bandung.