PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* DAN FITUR EKSTRAKSI N-GRAM DALAM ANALISIS SENTIMEN BERBASIS ASPEK

Robi Nurhidayat[1], Kania Evita Dewi[2]

[1,2] Program Studi Teknik Informatika, Universitas Komputer Indonesia
Jl. Dipatiukur No. 112-116 Bandung
E-mail : kania.evita.dewi@email.unikom.ac.id[2]

Abstrak

Pertumbuhan dan perkembangan teknologi yang begitu cepat dan pesat menjadikan membeli produk secara *online* semakin meningkat dan disukai yaitu membeli produk kecantikan. Banyak pertimbangan untuk mengetahui kualitas dari produk, salah satu caranya yaitu melihat ulasan produk kecantikan. Tujuan dari penelitian untuk mengevaluasi performansi dari metode *K-Nearest Neighbor* dan fitur ekstraksi *N-Gram* dalam melakukan analisis sentimen berbasis aspek pada produk kecantikan. Metodologi yang digunakan adalah pengumpulan data, *preprocessing*, ekstraksi fitur *N-Gram*, pembobotan kata dengan TF-IDF, klasifikasi dengan *K-Nearest Neighbor* ,Multi Label dengan binari ova, dan terakhir evaluasi performansi. Pembagian data dibagi menjadi tiga skenario yaitu 80:20, 70:30:, dan 60:40. Pengujian dilakukan dengan dataset original dan data yang diseimbangkan menggunakan metode *Random Over Sampling*. Hasil pengujian menunjukkan bahwa data yang seimbang menghasilkan nilai akurasi yang lebih baik daripada data yang tidak seimbang. Pada KNN dengan nilai k = 1 pada dataset seimbang, menghasilkan akurasi tertinggi. Akurasi pada aspek aroma, harga, kemasan dan efektivitas secara berturut-turut adalah 91,9%; 95,4%; 98,6%; 88,8%. Berdasarkan hasil pengujian yang telah dilakukan pada setiap aspek, didapatkan akurasi tertinggi dengan nilai akurasi 98,6% dari aspek kemasan pada skenario data 80:20.
Kata kunci : Analisis Sentimen, KNN, *N-Gram, ROS,* Multilabel.

Abstract

The growth and development of technology is so fast and rapid that buying products online is increasing and preferred, namely buying beauty products. There are many considerations to determine the quality of the product, one way is to look at beauty product reviews. The purpose of the research is to evaluate the performance of the K-Nearest Neighbor method and N-Gram extraction features in performing aspect-based sentiment analysis on beauty products. The methodology used is data collection, preprocessing, N-Gram feature extraction, word weighting with TF-IDF, classification with K-Nearest Neighbor, Multi Label with ova binaries, and finally performance evaluation. The data division is divided into three scenarios, namely 80:20, 70:30, and 60:40. Testing is done with the original dataset and balanced data using the Random Over Sampling method. The test results show that balanced data produces better accuracy values than unbalanced data. In KNN with a value of k = 1 on a balanced dataset, it produces the highest accuracy. Accuracy in the aspects of aroma, price, packaging and effectiveness are 91.9%; 95.4%; 98.6%; 88.8%, respectively. Based on the test results that have been carried out on each aspect, the highest accuracy is obtained with an accuracy value of 98.6% from the packaging aspect in the 80:20 data scenario.
Keywords : *Sentiment Analysis, KNN, N-Gram, ROS, Multilable.*

1. PENDAHULUAN

Pertumbuhan dan perkembangan teknologi yang begitu cepat dan pesat menjadikan membeli produk secara *online* semakin meningkat dan disukai, salah satunya yaitu membeli produk kecantikan. Ketika hendak ingin membeli produk kecantikan, dibutuhkan banyak pertimbagan untuk mengetahui kualitas dari produk kecantikan tersebut. Salah satu caranya yaitu dengan melihat ulasan produk kecantikan. Ulasan produk kecantikan dapat memberikan informasi tentang mutu dan kualitas dari produk kecantikan tersebut[1].

Analisis sentimen merupakan sebuah proses untuk mengenali pendapat atau opini seseorang terhadap suatu topik atau produk tertentu dan dapat diklasifikasikan ke dalam kategori positif, negatif, atau

netral. Proses ini dilakukan dengan tujuan untuk menganalisis atau memahami sikap seseorang terhadap topik atau produk tersebut[2]. Analisis sentimen dibagi menjadi 3 tingkat yaitu, tingkat dokumen, tingkat kalimat dan tingkat aspek[1]. Pada Penelitian[3], menggunakan metode *K-Nearest Neighbor* dan *N-Gram* hasil eksperimen menunjukkan 7 persen perbaikan akurasi dari metode SVM pada analisis sentimen. Eksperimennya dilakukan pada data bahasa Inggris dan saran untuk penelitian selanjutnya pengujian dilakukan pada bahasa lain[3]. Sebuah penelitian yang dilakukan dengan menggunakan pendekatan *N-gram* dan metode *K-Nearest Neighbor*. Dalam penelitian tersebut, digunakan bigram dan menghasilkan tingkat akurasi sebesar 77,01% dengan nilai K = 7, menggunakan teknik 10-fold cross validation pada 100 data yang digunakan.[4]. Dalam kedua penelitian di tersebut, analisis sentimen yang dilakukan masih sebatas analisis tingkat dokumen, padahal dalam beberapa komentar terkadang membahas beberapa aspek dan beda sentimen. Analisis sentimen pada level aspek menunjukkan performa yang lebih baik[5]. Pada penelitian[2], Melakukan analisis sentimen tingkat aspek dengan akurasi 88% pada aspek Makanan, 76% pada aspek Layanan, dan 84% pada aspek Atmosfir. Namun analisis sentimen tingkat aspek menghasilkan permasalahan baru yaitu pengklasifikasian menjadi multi label. Pengklasifikasian multi label dapat diatasi dengan cara menggunakan *binary relevance*[6].

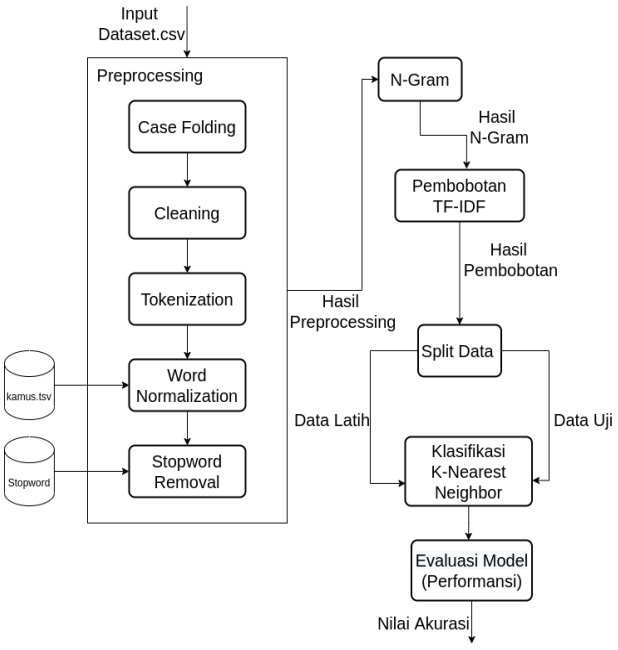
Beberapa teknik telah dikembangkan dan diterapkan untuk menganalisis sentimen, salah satunya adalah metode *K-Nearest Neighbor* (KNN). KNN adalah metode klasifikasi yang memanfaatkan jarak atau kemiripan antara suatu data dengan data lainnya. Dalam melakukan klasifikasi, KNN mengidentifikasi sejumlah k tetangga terdekat dari suatu data dan menentukan jenis kelas data tersebut berdasarkan mayoritas kelas tetangga terdekat tersebut. KNN termasuk algoritma yang relatif mudah dibandingkan dengan algoritma lain karena tidak membangun model pada saat pembelajaran mesin, melainkan hanya mengandalkan memori[7].

Berdasarkan beberapa penelitian sebelumnya dan latar belakang yang telah dijelaskan, penelitian ini bertujuan untuk menganalisis sentimen pada ulasan produk kecantikan dengan mempertimbangkan aspek yang relevan menggunakan algoritma *K-Nearest Neighbor* dan fitur ekstraksi *n-gram.* Tujuan dari penelitian ini adalah untuk mengevaluasi kinerja dari metode *K-Nearest Neighbor* dan *N-Gram* dalam melakukan analisis sentimen berbasis aspek pada produk kecantikan.

2. METODOLOGI

2.1 Alur Penelitian

Alur penelitian adalah proses awal penelitian hingga penelitian berakhir. proses penelitian akan melewati beberapa tahapan yaitu pengumpulan data, preprocessing, ekstraksi fitur dengan menggunakan *N-Gram* dan pembobotan kata dengan TF-IDF, proses klasifikasi *K-Nearest Neighbor*, dan terakhir evaluasi perfomansi. Alur penelitian dapat dilihat pada gambar 1.



Gambar 1. Alur penelitian

2.2 Dataset

Data yang digunakan pada penelitian ini yaitu ulasan produk kecantikan berasal dari situs kaggle yaitu https://www.kaggle.com/datasets/hafidahmusthaanah/skincare-review?select=00.+Review.csv. Data berjumlah 2554 ulasan yang sudah dilabeli secara manual. Terdapat empat aspek yang akan digunakan dalam penelitian ini yaitu harga, produk, kemasan, efektivitas. Setiap aspek pada ulasan akan diberi nilai sentimennya, nilai "0" untuk tidak memiliki sentimen, nilai "1" untuk aspek yang mempunyai sentimen positif, dan nilai "-1" untuk aspek yang mempunyai sentimen negatif.

2.3 Preprocessing

Tahap *preprocessing* dilakukan agar dataset lebih mudah digunakan untuk proses klasifikasi. Selain itu, tahapan *preprocessing* membuat dataset lebih seragam dan menghilangkan *noise* pada dataset. Tahapan dalam *preprocessing* yang digunakan adalah sebagai berikut:

A. *Case Folding*
   Case folding adalah tahap merubah jika ada huruf kapital atau besar di dalam dataset jadi huruf kecil[8].
B. *Cleaning*
   Cleaning adalah tahap membersihkan data seperti menghilangkan angka, tanda baca dan simbol dan *emoticon* pada dataset[9].
C. *Tokenization*
   Tokenization adalah tahap mengubah kalimat menjadi token token pada dataset[9].
D. *Word Normalization*
   Word Normalization adalah tahap membenarkan kata singkatan atau salah eja menjadi kata yang benar. Proses normalization ini berfungsi untuk merubah dimensi kata yang artinya sama tetapi memiliki ejaan yang salah[10].
E. *Stopword Removal*
F. Pada tahap ini, dilakukan pembuangan kata-kata yang dianggap tidak penting seperti: 'di', 'ke','dari', 'yang', 'dan', 'atau', 'ini', dan lainnya[11].

2.4 N-Gram

N-Gram adalah sejumlah pecahan kata yang dihasilkan dari sebuah kalimat. Metode *N-Gram* bisa juga digunakan untuk membangkitkan kata atau karakter[12]. *N-Gram* dapat dibentuk sedemikian rupa berdasarkan dari kata-kata sebelumnya dan berikutnya[13]. N-Gram dapat dibagi menjadi beberapa jenis berdasarkan jumlah pecahan kata atau substring yang dihasilkan. Jenis-jenis N-Gram ini terdiri dari Unigram, Bigram, Trigram, dan seterusnya sesuai dengan jumlah n dalam N-Gram[14]. *Unigram*, *Bigram*, dan *Trigram* adalah jenis *N-Gram* yang sering digunakan dalam pemrosesan bahasa alami. *Unigram* terdiri dari satu kata, sedangkan *Bigram* terdiri dari dua kata dan *Trigram* terdiri dari tiga kata. Ketiga jenis N-Gram ini umum digunakan dalam proses analisis bahasa alami untuk memahami hubungan antara kata-kata dalam teks. Jumlah pemotongan N-Gram ini ditentukan oleh jumlah gram yang diinginkan. Misalnya, jika kita ingin memecah kalimat "produknya tidak cocok dengan kulit saya" maka:

1. *Unigram*
   {" produknya"," tidak"," cocok", " dengan"," kulit"," saya"}
2. *Bigram*
   {"produknya_tidak","tidak_cocok","cocok_dengan","kulit_saya"}
3. *Trigram*
   {"produknya_tidak_cocok","tidak_cocok_dengan","cocok_dengan_kulit","dengan_kulit_saya "}

2.5 Pembobotan TF-IDF

Salah satu metode pembobotan kata yang paling populer untuk kasus klasifikasi teks adalah TF-IDF. TF-IDF ini dipakai sebagai pemberi bobot nilai menurut *level* kepentingan *term* terhadap kategori atau dokumen didalam setiap dokumen[9]. Dalam analisis dokumen, dapat dilakukan penghitungan frekuensi kemunculan suatu kata dalam dokumen tertentu, yang disebut *Term Frequency* (TF), serta kemungkinan suatu kata muncul dalam berbagai dokumen, yang disebut *Inverse Document Frequency* (IDF). Dengan cara ini, dapat diukur tingkat kepentingan suatu kata dalam dokumen dan menentukan kata-kata yang paling relevan dalam memahami teks tersebut[15]. Hasil dari TF-IDF yaitu berupa *matrix* yang dihasilkan oleh perkalian *Inverse Document Frequency* dengan *Term Frequency*. *Term Frequency* atau TF merupakan bobot nilai dari frekuensi term atau kata yang sering muncul pada dokumen. Sedangkan, *Inverse Document*

Frequency atau IDF adalah jumlah dokumen terkait yang mengandung kata tertentu[9]. Perhitungan TF-IDF dapat dilakukan menggunakan persamaan sebagai berikut[1]:

$$TF * IDF\ (d,t) = TF(d,t) * log\ (\frac{N}{df(t)}) \quad (1)$$

Keterangan:

$TF * IDF\ (d,t)$: Bobot *Term Frequency - Inverse Document Frequency*
$TF(d,t)$: jumlah munculnya term t pada dokumen d.
N : total dokumen.
$df(t)$: Jumlah dari dokumen yang terdapat term t.

2.6 K-Nearest Neighbor

Pada awal tahun 1960, *K-Nearest Neighbor* (KNN) diperkenalkan sebagai suatu metode. Pada masa itu, metode ini tidak populer dan hanya bekerja secara intensif ketika diberikan data training yang besar. Namun, seiring dengan perkembangan teknik komputasi, penggunaan metode ini menjadi semakin populer[4]. KNN merupakan algoritma klasifikasi *supervised learning*. KNN adalah salah satu algoritma sederhana dalam machine learning yang berbasis jarak. Perhitungan jarak dengan metode KNN dilakukan dengan cara menghitung jarak dari satu data yang berasal dari data test dengan seluruh data dari data train menggunakan *Euclidean Distance*[9]. Kelas target, yang memiliki jarak rata-rata minimum dari contoh uji, dipilih sebagai kelas dari data uji[16].Terdapat beberapa tahapan untuk menentukan kelas atau label pada data, yaitu:

1. Menentukan nilai *k*
2. Menghitung jarak terdekat dengan persamaan *Euclidean Distance*
3. Menjumlahkan hasil perhitungan jarak
4. Mengurutkan hasil penjumlah secara ascending
5. Memilih kategori tetangga terdekat sebanyak *k*
6. Memilih kelas berdasarkan frekuensi tertinggi terhadap *k* tertangga terdekat

Untuk menghitung jarak, dapat menggunakan persamaan *Euclidean Distance* :

$$deuclid = \sqrt{\sum_{i=1} |P_i - Q_i|^2} \quad (2)$$

Keterangan:

$deuclid$: jumlah fitur atau dimensi
P_i : fitur ke *i* pada data uji
Q_i : fitur ke *i* pada data latih

Didalam dataset terdiri dari 4 aspek, dimana hal ini menyebabkan data menjadi multilabel. Oleh karena itu, untuk melakukan proses klasifikasi dengan knn di lakukan sebanyak empat kali berdasarkan teknik yang digunakan yaitu *binary relevance*.

2.7 Confusion Matrix

Terdapat metode yang lebih efektif untuk mengevaluasi kinerja klasifikasi, yaitu menggunakan *confusion matrix*. *Confusion matrix* adalah sebuah tabel yang digunakan untuk menampilkan jumlah data uji yang diklasifikasikan dengan benar dan salah, sehingga memudahkan dalam mengevaluasi akurasi suatu sistem klasifikasi. Dengan menggunakan *confusion matrix*, kita dapat melihat secara detail kinerja suatu sistem klasifikasi dan mengidentifikasi di mana terjadi kesalahan klasifikasi.[17]. *Confusion matrix* merupakan sebuah teknik yang mudah dan efektif dalam mengukur kinerja sistem klasifikasi. Tujuan utama dari *confusion matrix* adalah untuk menilai performa atau akurasi dari suatu sistem klasifikasi dalam mengklasifikasikan data uji [18].

Tabel 1. Confusion Matrix

Confusion Matrix		Prediksi		
		Positif	Non Aspek	Negatif
Aktual	Positif	TPosPos	PFNon	PFNeg
	Non Aspek	NonFP	TNonNon	NonFNeg
	Negatif	NegFP	NegFNon	TNegNeg

Confusion Matrix digunakan untuk mengevaluasi tingkat akurasi dari suatu proses klasifikasi yang telah dilakukan. Tingkat akurasi ini mengindikasikan proporsi jumlah prediksi yang benar. Untuk menghitung akurasi, presisi, dan recall, digunakan rumus sebagai berikut:

$$Total = \text{TPosPos} + \text{PFNon} + \text{PFNeg} + \text{NonFP} + \text{TNonNon} + \text{NonFNeg} + \text{NegFP} + \text{NegFNon} + \text{TNegNeg} \quad (3)$$

$$accuracy = \frac{TPosPos+TNonNon+TNegNeg}{Total} \quad (4)$$

$$precision = \frac{TPosPos}{TPosPos+NonFP+NegFP} \quad (5)$$

$$recall = \frac{TPosPos}{TPosPos+PFNon+PFNeg} \quad (6)$$

3. HASIL DAN PEMBAHASAN

Dataset pada pengujian ini menggunakan data sebanyak 2554 ulasan produk kecantikan. Ulasan produk kecantikan tersebut yang mengandung aspek aroma dengan 502 memiliki sentimen positif, 130 sentimen negatif dan 1922 non sentimen. Untuk aspek harga dengan 340 memiliki sentimen positif, 118 sentimen negatif dan 2096 non sentimen. Untuk aspek kemasan dengan 134 memiliki sentimen positif, 27 sentimen negatif dan 2393 non sentimen. Untuk aspek efektivitas dengan 1771 memiliki sentimen positif, 686 sentimen negatif dan 97 non sentimen. Untuk lebih jelasnya jumlah setiap polaritas pada tiap aspek dapat dilihat pada tabel 2.

Tabel 2. Jumlah polaritas pada tiap aspek

Aspek	Polaritas		
	Positif (1)	Negatif (-1)	Non Sentimen (0)
Aroma	502	130	1922
Harga	340	118	2096
Kemasan	134	27	2393
Efektivitas	1771	686	97

Pengujian yang dilakukan pada sentiment tiap aspek dengan menggunakan split data sebanyak 80:20, 70:30, 60:40 antara data latih dan data uji. Pengujian juga dilakukan menggunakan jumlah *ngram* sebanyak *unigram*, *bigram*, *trigram* dan gabungan dari *unigram*, *bigram*, dan *trigram*. Pada proses klasifikasi menggunakan *K-Nearest Neighbor*. Nilai parameter k yang digunakan yaitu k = 23. Hasil pengujian akurasi akan dihitung menggunakan metode confusion matrix pada tiap aspek.

3.1 Pengujian Data Tidak Seimbang

Pada tahap ini dilakukan pengujian dengan data yang tidak seimbang. Pengujian dilakukan pada tiap aspek, hasil pengujian tiap aspek sebagai berikut.

A. Aspek Aroma

Pengujian model yang dibentuk oleh *K-Nearest Neighbor* terhadap aspek aroma dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapat dilihat hasilnya pada tabel 3.

Tabel 3. Hasil pengujian aspek aroma data tidak seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.765	0.761	0.753	0.773
70% dan 30%	0.776	0.767	0.752	0.773
60% dan 40%	0.771	0.767	0.751	0.77

Pada table 3 dapat dilihat hasil pengujian dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 70:30, dengan akurasi sebesar 0.776 atau sama dengan 77,6% dengan jenis *gram* yaitu *unigram*. Akurasi terkecil didapatkan dari pembagian data 60:40 dengan akurasi 0.751 atau sama dengan 75,1% yang didapatkan dari *trigram.*

B. Aspek Harga

Pada tabel 4 diperlihatkan hasil pengujian model *K-Nearest Neighbor* pada aspek harga dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram.*

Tabel 4. Hasil pengujian aspek harga data tidak seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.828	0.843	0.82	0.838
70% dan 30%	0.83	0.837	0.821	0.839
60% dan 40%	0.827	0.839	0.821	0.836

Dapat dilihat pada tabel 4 dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 80:20, dengan akurasi sebesar 0.843 atau sama dengan 84,3% dengan jenis n-*gram* yaitu *bigram*. Akurasi terkecil didapatkan dari pembagian data 80:20 dengan akurasi yang diperoleh adalah 0.82 atau sama dengan 82% dengan jenis n-*gram* yaitu *trigram.* Dari hasil perhitungan ini, dapat dilihat penggunaan *n-gram* dapat meningkatkan akurasi.

C. Aspek Kemasan

Hasil pengujian model yang dibentuk oleh *K-Nearest Neighbor* terhadap aspek Kemasan dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapat dilihat pada table 5.

Tabel 5. Hasil pengujian aspek kemasan data tidak seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.937	0.937	0.937	0.937
70% dan 30%	0.937	0.937	0.937	0.937
60% dan 40%	0.936	0.936	0.936	0.936

Dapat dilihat pada table 5 dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 80:20 dan 70:30, dengan akurasi sebesar 0.937 atau sama dengan 93,7% dengan jenis n-*gram* yaitu *unigram, bigram, trigram dan gabungan unigram,bigram, trigram*. Akurasi terkecil didapatkan dari pembagian data 60:40 dengan akurasi 0.936 atau sama dengan 93,6% yang didapatkan dari *unigram, bigram, trigram dan gabungan unigram, bigram, trigram.*

D. Aspek Efektivitas

Hasil pengujian model *K-Nearest Neighbor* terhadap aspek efektivitas dengan menggunakan ekstraksi fitur yang digunakan adalah *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapat dilihat pada tabel 6.

Tabel 6. Hasil pengujian aspek efektivitas data tidak seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.738	0.699	0.695	0.722
70% dan 30%	0.734	0.709	0.694	0.729
60% dan 40%	0.733	0.705	0.696	0.73

Dapat dilihat pada table 6 hasil akurasi dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 80:20, dengan akurasi sebesar 0.738 atau sama dengan 73,8% dengan jenis *gram* yaitu *unigram.* Akurasi terkecil didapatkan dari pembagian data 70:30 dengan akurasi 0.694 atau sama dengan 69,4% yang didapatkan dari *trigram.*

Berdasarkan hasil pengujian yang telah dilakukan menggunakan confusion matrix, dapat dilihat dari aspek aroma *unigram* hasil prediksi tidak mendeteksi satupun kelas "positif ", pada aspek harga *trigram* hasil prediksi tidak mendeteksi satupun kelas "positif dan negatif ", pada aspek kemasan *unigram* hasil prediksi tidak mendeteksi satupun kelas "positif dan negatif ". Sehingga untuk mengatasi hal itu akan dilakukan proses penyeimbangan kelas. Metode *random over sampling* dapat digunakan untuk mengatasi data tidak seimbang[19]. Kemudian untuk mengetahui nilai k terbaik untuk digunakan pada metode *random oversampling*, maka akan dilakukan pencarian nilai k dengan memanfaatkan evaluasi model dengan menggunakan grid search. Metode populer yang biasanya dipakai dalam mencari parameter paling optimal untuk suatu model salah satu nya yaitu grid search. Metode *grid search* melakuan optimasi parameter berdasarkan nilai parameter yang digunakan oleh *user*. Cara kerja dari metode ini, yaitu menggabungkan semua parameter yang telah ditentukan oleh peneliti, kemudian metode *grid search* akan menampung hasil parameter yang sudah dicampurkan ke dalam *grid*. Setelah itu, akan dipilih parameter terbaik dan menampilkan hasil yang terbaik juga. [20].Setelah pencarian nilai k terbaik dengan *grid search*, didapatkan nilai k terbaik yaitu k = 1.

3.2 Pengujian Data Seimbang

Pada tahap ini dilakukan pengujian dengan data yang seimbang. Pengujian dilakukan pada tiap aspek, hasil pengujian tiap aspek sebagai berikut.

A. Aspek Aroma

Hasil pengujian model yang dibentuk oleh K-*Nearest Neighbor* pada aspek aroma dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapay dilihat pada tabel 7.

Tabel 7. Hasil pengujian aspek aroma data seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.908	0.912	0.829	0.919
70% dan 30%	0.899	0.902	0.931	0.901
60% dan 40%	0.882	0.884	0.785	0.883

Berdasarkan tabel 7 pengujian dilakukan dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 70:30, dengan akurasi sebesar 0.931 atau sama dengan 93,1% dengan jenis *gram* yaitu *trigram.* Akurasi terkecil didapatkan dari pembagian data 60:40 dengan akurasi 0.785 atau sama dengan 78,5% yang didapatkan dari *trigram.* Terdapat peningkatan akurasi penggunaan *n-gram*

B. Aspek Harga

Hasil pengujian model yang dibentuk oleh K-*Nearest Neighbor* pada aspek Harga dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapay dilihat pada tabel 8.

Tabel 8. Hasil pengujian aspek harga data seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.954	0.943	0.947	0.948
70% dan 30%	0.945	0.932	0.937	0.942
60% dan 40%	0.936	0.924	0.928	0.932

Berdasarkan tabel 8 pengujian dilakukan dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 80:20, dengan akurasi sebesar 0.954 atau sama dengan 95,4% dengan jenis *gram* yaitu *unigram.* Akurasi terkecil didapatkan dari pembagian data 60:40 dengan akurasi 0.924 atau sama dengan 92,4% yang didapatkan dari *bigram.*

C. Aspek Kemasan

Hasil pengujian model yang dibentuk oleh K-*Nearest Neighbor* pada aspek Kemasan dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapay dilihat pada tabel 9.

Tabel 9. Hasil pengujian aspek kemasan data seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.984	0.984	0.986	0.983
70% dan 30%	0.981	0.981	0.984	0.978
60% dan 40%	0.979	0.976	0.982	0.976

Berdasarkan table 9 pengujian dilakukan dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 80:20, dengan akurasi sebesar 0.986 atau sama dengan 98,6% dengan jenis *n-gram* yaitu *unigram.* Akurasi terkecil didapatkan dari pembagian data 60:40 dengan akurasi 0.976 atau sama dengan 97,6% yang didapatkan dari *bigram.* Terdapat kenaikan akurasi sebesar 0.2% dari *unigram* dan *bigram* ke *trigram.*

D. Aspek Efektivitas

Hasil pengujian model yang dibentuk oleh K-*Nearest Neighbor* pada aspek Efektivitas dengan menggunakan ekstraksi fitur *unigram*, *bigram*, *trigram* dan gabungan *unigram*, *bigram*, *trigram* dapay dilihat pada tabel 10.

Tabel 10. Hasil pengujian aspek kemasan data seimbang

Jumlah Pembagian Data	Hasil Akurasi			
	Unigram	Bigram	Trigram	Unigram, Bigram, Trigram
80% dan 20%	0.880	0.879	0.886	0.888
70% dan 30%	0.876	0.863	0.855	0.874
60% dan 40%	0.856	0.853	0.847	0.856

Berdasarkan tabel 10 pengujian dilakukan dengan menggunakan skema pembagian dataset menjadi 80:20, 70:30, dan 60:40 antara data latih dan data uji menghasilkan akurasi yang beragam. Hasil akurasi tertinggi didapatkan dari pembagian dataset dengan ukuran 80:20, dengan akurasi sebesar 0.888 atau sama dengan 88.8% dengan jenis *n-gram* yaitu gabungan *unigram, bigram, trigram*. Akurasi terkecil didapatkan dari pembagian data 60:40 dengan akurasi 0.847 atau sama dengan 84.7% yang didapatkan dari *trigram.*

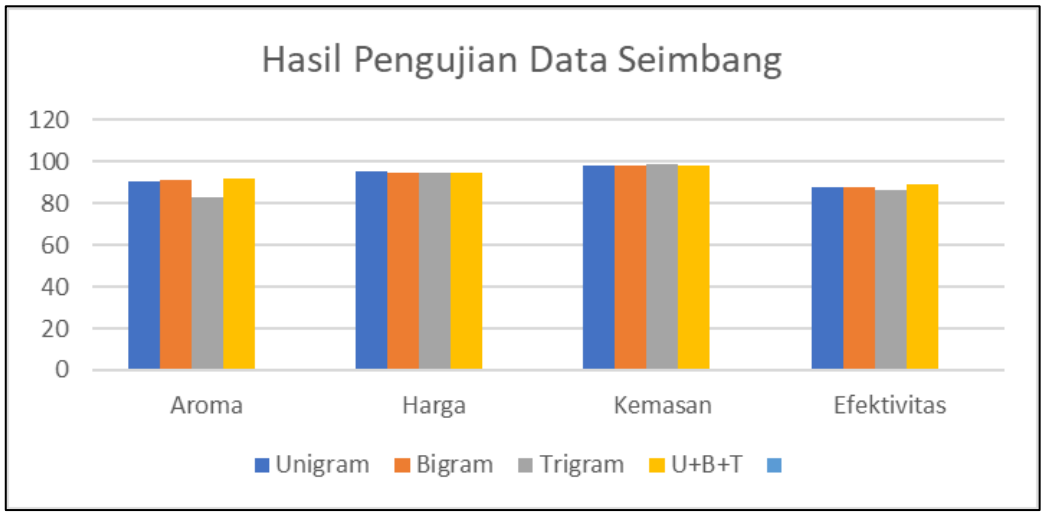
3.3 Hasil Pengujian

Hasil pengujian diambil berdasarkan dari akurasi tertinggi yang didapatkan. Berdasarkan hasil pengujian, didapatkan hasil akurasi tertinggi setelah dilakukan proses penyeimbangan kelas tiap aspek nya. Hasil akurasi tertinggi didapatkan dari ukuran split data 80:20 antara data latih dan data uji. Berikut ini hasil akurasi yang didapatkan dapat dilihat pada tabel 11.

Tabel 10. Kesimpulan hasil pengujian

Akurasi	Aroma	Harga	Kemasan	Efektivitas
Unigram	90,8%	95,4%	98,4%	88%
Bigram	91,2%	94,3%	98,4%	87,9%
Trigram	82,9%	94,7%	98,6%	86,6%
Unigram, Bigram, Trigram	91,9%	94,8%	98,3%	88,8%

Dari tabel 11, didapatkan hasil akurasi tertinggi yaitu 98.6% dari aspek kemasan dengan menggunakan *trigram*. Sedangkan hasil akurasi terrendah diperoleh oleh asepk efektivitas sebesar 86.6%.
Pada gambar 2 terdapat grafik dari hasil pengujian data seimbang untuk setiap aspek.



Gambar 2. Kesimpulan hasil pengujian

Berdasarkan gambar 2 dapat dilihat bahwa dengan menggunakan penyeimbangan data kebanyakan akurasi meningkat dan untuk semua aspek dengan menggunakan skema pembagian data yang berbeda akurasi berasa diatas 80%.

4. PENUTUP

Berdasarkan hasil dari semua tahapan yang telah dilakukan dalam analisis sentimen berbasis aspek dengan menggunakan metode *K-Nearest Neighbor* dengan *N-Gram* pada kasus ulasan produk kecantikan mendapatkan hasil akurasi tertinggi yaitu 98.6% pengujian sentimen aspek kemasan dengan data yang seimbang menggunakan *trigram*. Terjadi peningkatan akurasi sebesar 0.4% dan 1.1% pada pengujian aspek aroma dengan menggunakan *Bigram* dan gabungan dari *unigram*, *bigram*, dan *trigram* menghasilkan akurasi 91.2% dan 91.9% sedangkan menggunakan *unigram* menghasilkan akurasi 90.8%. Terjadi peningkatan akurasi sebesar 0.2% pada pengujian aspek kemasan dengan menggunakan *trigram* menghasilkan akurasi 98,6% sedangkan menggunakan *unigram* 98,4%.

Penelitian ini dapat dilakukan pengembangan lebih lanjut agar kedepannya mendapatkan hasil yang lebih baik dengan menggunakan teknik penyeimbangan data yang lebih baik, karena didalam penelitian ini dapat dilihat bahwa dengan menyeimbangkan data akurasi dapat meningkat.

DAFTAR PUSTAKA

[1] N. F. Putri, S. Al Faraby, and M. Dwifebri, "Analisis Sentimen pada Produk Kecantikan dari Ulasan Female Daily Menggunakan Information Gain dan SVM Classifier," *e-Proceeding of Enginering*, vol. 8, no. 5, pp. 10068–10079, 2021, doi: https://doi.org/10.34818/eoe.v8i5.15731.

[2] W. Parasati, F. A. Bachtiar, and N. Y. Setiawan, "Analisis Sentimen Berbasis Aspek pada Ulasan Pelanggan Restoran Bakso President Malang dengan Metode Naïve Bayes Classifier," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 4, pp. 1090–1099, 2020, [Online]. Available: http://j-ptiik.ub.ac.id.

[3] S. Kaur, G. Sikka, and L. K. Awasthi, "Sentiment Analysis Approach Based on N-gram and KNN Classifier," *ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun.*, pp. 13–16, 2018, doi: 10.1109/ICSCCC.2018.8703350.

[4] R. Sari, "Analisis Sentimen Pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma K-Nearest Neighbor (K-Nn)," *EVOLUSI J. Sains dan Manaj.*, vol. 8, no. 1, pp. 10–17, 2020, doi: 10.31294/evolusi.v8i1.7371.

[5] F. A. Hirzani, W. Maharani, and M. A. Bijaksana, "Analisis Sentimen Review Produk Menggunakan

Pendekatan Berbasis Kamus," *e-Proceeding of Enginering*, vol. 2, no. 2, pp. 5891–5898, 2015, doi: https://doi.org/10.34818/eoe.v2i2.2991.

[6] A. J. Rivera and M. J. Jesus, *Francisco Herrera, Francisco Charte, Antonio J. Rivera, María J. del Jesus (auth.) - Multilabel Classification _ Problem Analysis, Metrics and Techniques (2016, Springer International Publishing) - libgen.lc.pdf*. .

[7] I. Indriati and A. Ridok, "Sentiment Analysis for Review Mobile Applications Using Neighbor Method Weighted K-Nearest Neighbor (Nwknn)," *J. Enviromental Eng. Sustain. Technol.*, vol. 3, no. 1, pp. 23–32, 2016, doi: 10.21776/ub.jeest.2016.003.01.4.

[8] F. Pramono, D. Rosiyadi, and W. Gata, "Integrasi N-gram, Information Gain, Particle Swarm Optimation di Naïve Bayes untuk Optimasi Sentimen Google Classroom," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 383–388, 2019, doi: 10.29207/resti.v3i3.1119.

[9] E. Y. Prastika S., S. Al Faraby, and M. D. Purbolaksono, "Analisis Sentimen pada Ulasan Produk Kecantikan Menggunakan K-Nearest Neighbor dan Information Gain," *eProceedings Eng.*, vol. 8, no. 5, pp. 10091–10105, 2021, doi: https://doi.org/10.34818/eoe.v8i5.15729.

[10] A. N. Indraini, I. Ernawati, and A. Zaidah, "Analisis Sentimen Terhadap Pembelajaran Daring Di Indonesia Menggunakan Support Vector Machine (SVM)," 2021.

[11] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 422, 2021, doi: 10.30865/mib.v5i2.2845.

[12] N. Arifin, U. Enri, and N. Sulistiyowati, "Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 6, no. 2, p. 129, 2021, doi: 10.30998/string.v6i2.10133.

[13] A. Kulkarni and A. Shivananda, *Natural Language Processing Recipes*. 2019.

[14] F. Fitriyani and T. Arifin, "Penerapan Word N-Gram Untuk Sentiment Analysis Review Menggunakan Metode Support Vector Machine (Studi Kasus: Aplikasi Sambara)," *Sistemasi*, vol. 9, no. 3, p. 610, 2020, doi: 10.32520/stmsi.v9i3.954.

[15] K. K. Purnamasari and N. I. Widiastuti, "Perbandingan Algoritma K-Means Dan K-Nearest Neighbors Pada Sistem Peringkasan Otomatis," *Komputa J. Ilm. Komput. dan Inform.*, vol. 6, no. 2, pp. 57–66, 2017, doi: 10.34010/komputa.v6i2.2478.

[16] S. Mukhopadhyay and P. Samanta, *Advanced Data Analytics Using Python*. 2023.

[17] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 5, no. 2, pp. 697–711, 2021.

[18] R. Syafaat Amardita and M. Dwifebri Purbolaksono, "Analisis Sentimen terhadap Ulasan Paris Van Java Resort Lifestyle Place di Kota Bandung Menggunakan Algoritma KNN," *J. Ris. Komputer)*, vol. 9, no. 1, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i1.3793.

[19] L. Qadrini, H. Hikmah, and M. Megasari, "Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 386–391, 2022, doi: 10.47065/josyc.v3i4.2154.

[20] M. Fajri and A. Primajaya, "Komparasi Teknik Hyperparameter Optimization pada SVM untuk Permasalahan Klasifikasi dengan Menggunakan Grid Search dan Random Search," vol. 7, no. 1, pp. 14–19, 2023.