Analisis Sentimen Aplikasi *TikTok* menggunakan Metode BM25 dan *Improved K-NN* Fitur *Chi-Square*

Annida Purnamawati[1*], Monikka Nur Winarto[2], Mely Mailasari[3]
[1]Sistem Informasi Kampus Kota Yogyakarta, Universitas Bina Sarana Informatika
[2]Sistem Informasi Kampus Kota Pontianak, Universitas Bina Sarana Informatika
[3]Sistem Informasi, Universitas Bina Sarana Informatika
*email: annida.npr@bsi.ac.id

DOI: https://doi.org/10.31603/komtika.v7i1.8938

Received: 01-04-2023, Revised: 25-04-2023, Accepted: 29-04-2023

ABSTRACT

Currently technological advances are very rapid, as well as the use of the internet is increasing. This change is supported by the development of communication media which increases the number of internet users and encourages the rapid dissemination of information through social media applications. One of the most popular social media applications in Indonesia today is the TikTok application. The TikTok application provides a place to make videos with a duration of 60 seconds and has many features such as adding music, changing voices, providing filters, adding effects and stickers. Users of the application range from minors to the elderly, so not a few users give positive or negative reviews. Therefore this study helps users to analyze the review data by conducting experiments using the sentiment classification technique using the BM25 method as word weighting, and Improved K-NN as a determinant in choosing sentiment by adding the Chi-Square feature to reduce the number of words in the classification. . The test uses 5 feature ratio tests and then gets the best results from a feature ratio of 50% and k = 20 so that the best results are obtained, namely the value of precision 70.03%, recall 67.22%, accuracy 83.33% and f-measure 66. 26%. It can be concluded that the addition of feature selection can help improve recall, f-measure, precision, and accuracy results.

Keywords: *TikTok, sentiment, BM25, improved K-NN, Chi-Square*

ABSTRAK

Saat ini kemajuan teknologi sangat pesat, begitu juga halnya penggunaan internet semakin meningkat. Perubahan tersebut didukung dengan berkembangnya media komunikasi yang membuat jumlah penggunaan internet meningkat dan mendorong persebaran informasi sangat cepat melalui aplikasi sosial media. Aplikasi *TikTok* merupakan salah satu sosial media di Indonesia yang sangat popular saat ini. Aplikasi *TikTok* memberikan wadah untuk membuat video dengan durasi 60 detik dan mempunyai banyak fitur seperti menambahkan musik, mengubah suara, memberikan filter, menambahkan efek dan stiker. Pengguna aplikasi tersebut dari anak dibawah umur sampai dengan yang sudah tua maka tidak sedikit pengguna memberikan ulasan positif maupun negatif. Maka dari itu pada penelitian ini membantu pengguna untuk menganalisis data ulasan tersebut dengan melakukan eksperimen menggunakan teknik klasifikasi sentimen menggunakan metode BM25 sebagai pembobotan kata, dan *Improved K-NN* sebagai penentu dalam memilih sentimen dengan menambahkan fitur *Chi-Square* guna untuk mengurangi jumlah kata dalam klasifikasi. Pengujian menggunakan 5 kali pengujian rasio fitur kemudian di dapatkan hasil terbaik dari rasio fitur 50% dan k = 20 sehingga memperoleh hasil terbaik yaitu nilai *precision* 70,03%, *recall* 67,22%, *accuracy* 83,33% dan *f-measure* 66,26%. Dapat disimpulkan untuk penambahan fitur seleksi dapat membantu meningkatkan hasil *recall*, *f-measure*, *precision*, dan *accuracy*.

Keywords: *TikTok*, Sentimen, *BM25*, *improved K-NN*, *Chi-Square*

PENDAHULUAN

Kemajuan teknologi pada dunia internet sangat pesat, mulai dari perdagangan, pendidikan sampai dengan komunikasi sosial. Pengguna internet pada periode 2022 sampai dengan 2023 di Indonesia sendiri sudah mencapai 215,63 juta pengguna berdasarkan hasil survei dari Asosiasi Penyelenggara Jasa Internet Indonesia(APJII) . Peningkatan jumlah penggunaan internet ada mengalami kenaikan 2,67% jika dibandingkan dengan tahun lalu yaitu sebesar 210,03 juta pengguna [1]. Perubahan data tersebut didukung dengan berkembangnya media komunikasi yang membuat jumlah penggunaan internet meningkat dan mendorong persebaran informasi sangat cepat melalui aplikasi sosial media. Salah satu aplikasi sosial media yang sangat popular di Indonesia saat ini adalah aplikasi *TikTok* [2]. Aplikasi *TikTok* merupakan platform media sosial dan juga video musik yang dibuat oleh Zhang Yiming pada bulan September 2016. Aplikasi *TikTok* mendorong penggunanya untuk dapat kreatif membuat video-video yang mereka inginkan. Aplikasi *TikTok* memberikan wadah untuk dapat membuat video dengan durasi 60 detik dan dapat menambahkan banyak fitur-fitur, misalkan menambahkan music, mengubah suara, memberikan filter, menambahkan stiker maupun efek. Aplikasi *TikTok* juga terdapat kolom komentar, like, dapat menyimpan video dan dapat membagikan video ke orang lain. Pengguna aktif bulanan *TikTok* di berbagai dunia sudah mengalami peningkatan yang cukup pesat sejak awal tahun pandemik 2020 hingga menurut Business of Apps, data kuartal II 2022 aplikasi *TikTok* mempunyai 1,46 miliar pengguna aktif bulanan dikutip dari databoks [3]. Aplikasi *TikTok* tersebut dapat di download pada *Google Play Store* sehingga tak sedikit pula yang memberikan penilaian pada aplikasi tersebut. Dalam berkomentar pengguna bebas memberikan opini positif terhadap aplikasi maupun negatif. Maka dari itu, penelitian menggunakan analisis sentimen untuk menggali informasi lebih lanjut untuk dapat menganalisis opini yang di berikan oleh pengguna aplikasi *TikTok* pada ulasan *Google Play Store.*

Analisis sentimen juga merupakan cara dimana kita dapat memberikan sebuah opini yang bersifat publik di sosial media yang memuat pelayanan publik dan isu terkini [4]. Adanya ulasan yang diberikan pada aplikasi *TikTok* pada *Google Play Store* maka analisis sentimen memiliki pengaruh dan manfaat besar. Analisis sentimen merupakan bidang studi untuk melakukan analisis penilaian, opini, evaluasi terhadap sikap seseorang dan sentimen terhadap organisasi individu, produk dan lain-lain sehingga diklasifikasikan menjadi dua yaitu positif dan negatif [5].

Ada beberapa algoritma yang sering digunakan untuk klasifikasi pada data teks, salah satunya metode *K-Nearest Neighbor (*K-NN*)*. K-NN merupakan sebuah algoritma pada machine learning yang dapat digunakan dalam proses klasifikasi dengan berdasarkan seberapa dekat lokasi atau jarak suatu data dengan data lainnya [6]. Dengan hal tersebut proses algoritma K-NN untuk dapat melakukan klasifikasi dataset dengan mencari jumlah k yang terdekat untuk bisa diambil kelas mayoritasnya dari data yang akan diolah. Salah satu kelebihan yang dimiliki oleh algoritma K-NN adalah kinerja yang bagus dari metode tersebut dan juga hasil yang akurat dalam klasifikasi teks. K-NN dapat menghasilkan *accuracy* 90% dari jumlah k sebanyak 5 sehingga K-NN dapat disimpulkan mempunyai kinerja yang bagus [7]. Pemeringkatan dan juga pembobotan dokumen merupakan komponen dalam klasifikasi teks yang sangat penting . Salah satu metode yang membantu proses pemeringkatan dokumen yaitu *Best Matching 25*(BM25). BM25 merupakan metode untuk pemeringkatan yang dapat mencocokan pengurutan antara

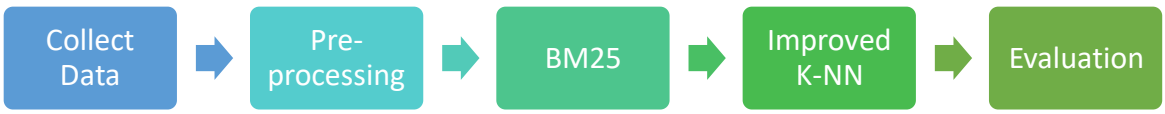
kata kunci dan dokumen koleksi data yang ada. Menurut penelitian Tinega,dkk 2018 menjelaskan bahwa pemeringkatan penggunaan BM25 mempunyai nilai yang jauh lebih baik jika dibandingkan dengan *vector space model* dan *boolean model* [8].

Studi sebelumnya yang telah dilakukan Pardede dkk ditahun 2015 menggunakan metode BM25 menghasilkan nilai rata-rata F-measure 61,649 dan pada metode PLSA rata-rata F-measure 56,8877 hal ini menunjukkan peforma yang dihasilkan pada metode BM25 lebih baik dan efisien jika dibandingkan dengan metode PLSA dalam hal pemeringkatan dokumen. Kemudian penelitian Nathania dkk pada 2018 mengklasifikasikan spam twitter menggunakan algoritma *Improved K-NN* menunjukkan algoritma *Improved K-NN* menghasilkan *accuracy* yang lebih efektif dibandingkan algoritma K-NN dalam pengklasifikasian tersebut [9].

Berdasarkan hal tersebut penelitian ini menerapkan kolaborasi metode BM25 dan metode Improved K-NN dengan menambahkan seleksi fitur *Chi-Square*. Klasifikasi dengan fitur *Chi-Square* diyakinkan dapat menghilangkan fitur yang mengganggu dalam proses klasifikasi sehingga dapat diperoleh hasil nilai *accuracy* yang maksimal dalam klasifikasi teks [10]. Penelitian ini diharapkan dapat memberikan nilai *accuracy* yang baik dan tepat sehingga dapat dijadikan tolak ukur untuk penelitian selanjutnya.

METODE

Metode yang dilakukan dalam proses perancangan system dimulai dari pengumpulan data yang diambil dari ulasan *Google Playstore*, preprocessing, seleksi fitur dengan *Chi-Square*, menerapkan metode BM25 yang fungsinya untuk penghitungan pembobotan kata dan melakukan *Improved K-NN* untuk proses klasifikasi. Tahapan sistem yang dilakukan secara umum dijelaskan pada Gambar 1.



Gambar 1. Tahap Penelitian

Berikut merupakan penjelasan dari metode yang telah di rancang:

1. *Collcet Data*

Collect data (pengumpulan data) dengan pendekatan kuantitatif yaitu melakukan survei dari lituratur akademis dari bidang keilmuan yang sama dengan tujuan untuk mendapatkan konsep yang relevan sehingga dapat menyalurkan inovasi penelitian untuk kebijakan publik [11]. Dalam penelitian ini dataset yang digunakan merupakan ulasan aplikasi *TikTok* pada *Google Play Store* di bulan Desember 2022 sampai dengan Februri 2023 sebanyak 400 ulasan. Pengumpulan data menggunakan teknik scraping, yaitu pengambilan dokumen dari halaman website dengan bantuan *webHarvy*.

2. *Pre-processing*

Tahap *pre-processing* dilakukan dengan mengubah data tidak terstruktur menjadi terstruktur. Tujuan dari *pre-processing* text yaitu menghilangkan noisy pada data sehingga nantinya didapatkan hasil yang lebih optimal. Selain itu untuk melakukan ekstrak *vector feature* menjadi berkualitas tinggi untuk setiap dokumen yang digunakan, sehingga values yang diambil adalah yang nilainya tinggi dan penting [12].

a. Data Cleaning

Data Cleaning merupakan proses penyaringan data. Proses ini dilakukan untuk menghapus hastag, tanda baca, angka ,nama pengguna maupun karakter khusus pada tabel 1.

Tabel 1. Hasil *Data Cleaning*

Input	Output
ini sangat seru untuk menonton jika melamun,Saya beri nilai lima untuk tiktok,tiktok salah satu aplikasi yang saya suka dan sering saya nonton dari aplikasi lain,Segitu saja ,terimakasih	ini sangat seru untuk menonton jika melamun Saya beri nilai lima untuk tiktok tiktok salah satu aplikasi yang saya suka dan sering saya nonton dari aplikasi lain Segitu saja terimakasih

b. Case Folding

Case Folding merupakan tahapan yang berfungsi untuk mengubah huruf kapital menjadi huruf kecil (a-z). Hasil *case folding* disajikan pada tabel 2.

Tabel 2. Hasil *Data Cleaning*

Input	Output
ini sangat seru untuk menonton jika melamun Saya beri nilai lima untuk tiktok tiktok salah satu aplikasi yang saya suka dan sering saya nonton dari aplikasi lain Segitu saja terimakasih	ini sangat seru untuk menonton jika melamun saya beri nilai lima untuk tiktok tiktok salah satu aplikasi yang saya suka dan sering saya nonton dari aplikasi lain segitu saja terimakasih

c. Tokenization

Tokenization adalah salah satu bagian penting dari awal proses pengolahan data teks yaitu proses untuk memisah atau memotong sebuah kalimat, dokumen dan paragraph menjadi sebuah token. Hasil *tokenization* disajikan pada tabel 3.

Tabel 3. Hasil *Data Cleaning*

Input	Output
ini sangat seru untuk menonton jika melamun saya beri nilai lima untuk tiktok tiktok salah satu aplikasi yang saya suka dan sering saya nonton dari aplikasi lain segitu saja terimakasih	‘ini’, ‘sangat’, ‘seru’, ‘untuk’, ‘menonton’, ‘jika’, ‘melamun’, ‘saya beri’, ‘nilai’, ‘lima’, ‘untuk’, ‘tiktok’, ‘tiktok’ ,’salah’, ‘satu’, ‘aplikasi’, ‘yang’, ‘saya’, ‘suka’, ‘dan’, ‘sering’, ‘saya’, ‘nonton’, ‘dari’, ‘aplikasi’, ‘lain’, ‘segitu’, ‘saja’, ‘terimakasih’

d. Stopwords Removal

Pada tahap ini proses proses untuk menentukan apakah kata tertentu harus termasuk ke dalam *stopword* . Proses *stopword removal* akan melakukan penghilangan terhadap kata yang tidak memiliki arti. Adapun *term* yang didapat dari tahap *tokenization* akan dilihat kedalam suatu daftar *stopword*, jika kata tersebut termasuk kedalam daftar *stopword* maka kata tersebut tidak akan diproses ketahap selanjutnya. Berikut contoh *stop word removal* yaitu “yang”, “di”, “ke”, “itu”, “ini” dan lain sebagainya. Hasil *Stopwords Removal* disajikan pada tabel 4.

Tabel 4. Hasil *Stopwords Removal*

Stopwords
‘sangat’, ‘seru’, ‘menonton’, ‘jika’, ‘melamun’, ‘saya’, ‘beri’, ‘nilai’, ‘lima’, ‘untuk’, ‘tiktok’, ‘tiktok’, ’salah’, ‘satu’, ‘aplikasi’, ‘saya’, ‘suka’, ‘dan’, ‘sering’, ‘saya’, ‘nonton’, ‘aplikasi’, ‘segitu’, ‘saja’, ‘terimakasih’

e. Steamming

Steamming adalah proses untuk mengembalikan kata dasar dengan cara menghilangkan sisipan, awalan, maupun akhiran. Proses ini dilakukan karena hanya sebuah kata dasar yang tersimpan di dalam *database*. Hasil dari *Steamming* disajikan pada tabel 5.

Tabel 5. Hasil *Steamming*

Input	Output
'sangat', 'seru', 'menonton', 'jika', 'melamun', 'saya', beri', 'nilai', 'lima', 'untuk', 'tiktok', 'tiktok' ,'salah', 'satu', 'aplikasi', 'saya', 'suka', 'dan', 'sering', 'saya', 'nonton', 'aplikasi', 'segitu', 'saja', 'terimakasih'	'sangat', 'seru', 'nonton, 'jika', 'lamun, 'saya', 'beri', 'nilai', 'lima', 'untuk', 'tiktok', 'tiktok' ,'salah', 'satu', 'aplikasi', 'saya', 'suka', 'dan', 'sering', 'saya', 'nonton', 'aplikasi', 'segitu', 'saja', 'terimakasih'

3. BM25

Metode BM25 adalah suatu sistem perangkingan yang dapat melakukan mengurutkan hasil kecocokan terhadap sebuah dokumen. Metode ini juga formula terbaik dalam kelas *best match*, karena ada formula yang efektif dan mempunyai tingkat ketepatan saat mengurutkan dokumen berdasarkan *query* yang dicari [13]. Metode BM25 berfungsi untuk menghitung tiga (tiga) faktor yang mempengaruhi peringkat dokumen dibandingkan dengan dokumen term. Faktor yang pertama adalah *Term Frequency* (TF) karena nilai fungsi peringkat dokumen semakin besar seiring dengan nilai TD. TF yang dimaksud adalah banyaknya kemunculan dari *term* yang terdapat dalam sebuah dokumen. Faktor kedua yaitu *Invers Document Frequency Of Term* (IDF) adalah nilai *invers* dari seluruh dokumen yang terdapat kata yang akan dicari. Dengan kata lain, nilai IDF yang lebih tinggi menunjukkan bahwa lebih banyak kata yang akan dibuat pada dokumen. Faktor ketiga adalah panjang dari dokumen, yang berarti jumlah kata dalam dokumen yang digunakan [14].

Peringkat dari N dokumen dijabarkan dengan persamaan 1 (satu):

$$BM25(d_j, q_{1:N}) = \sum_{i=1}^{N} IDF(q1)\ \frac{TF(qi,dj).(k+1)}{TF\ (q1,dj)+k.(1+b+b.\frac{|dj|}{L})} \quad (1)$$

Dengan L dijabarkan pada persamaan 2 (dua):

$$L = \frac{\sum i|di|}{N} \quad (2)$$

IDF dapat dilihat pada persaman 3:

$$IDF(qi) = log\frac{N-DF(qi)+0,5}{DF(qi)+0,5} \quad (3)$$

Keterangan : *Qi* yaitu *term* yang dicari, |*dj*| yaitu panjang dari dokumen *dj*, *DF(qi)* yaitu jumlah dokumen berupa *qi*, *TF(qi,dj)* yaitu banyaknya kemunculan kata *qi* didalam dokumen *dj*, *IDF(qi)* yaitu nilai *invers* dari seluruh dokumen yang terdapat kata *qi*, *L* yaitu rerata panjang dari dokumen dari *N* dan *k* konstanta dalam melakukan evaluasi yang rentannya 1,2-2, *b* ketetapan dari konstanta yang mempunyai rentang 0,75-0,8.

4. *Improved K-NN*

Metode *Improved K-NN* adalah bagian dari pengembangan klasifikasi dengan metode K-NN yang memiliki perbedaan yaitu dalam proses klasifikasi objek yang baru, *Improved K-NN* menggunakan tahap yang sama dengan K-NN saat melakukan proses perhitungan [15]. Namun pada metode *Improved K-NN* proses perhitungan bobot setiap data latih diuji berdasarkan data uji dan data latih lainnnya, kemudian masing-masing kelas data uji ditentukan oleh bobot latih, untuk dapat menghitung kesamaan antar dokumen, menggunakan metode *cosine similarity*. Rumus untuk menghitung yaitu sebagai berikut :

$$\cos\ (\theta_{QD} = \frac{-b\pm\sqrt{b^2-4ac}}{\sqrt{\sum_{i=1}^{n}(Q_i)^2}\sqrt{\sum_{i=1}^{n}(D_i)^2}} \quad (4)$$

Keterangan : Cos(θ_{QD}) merupakan kesamaan dokumen dari Q terhadap dokumen D, Q merupakan data uji, D merupakan data latih, N merupakan banyaknya data yang akan digunakan.

Dari rumus tersebut dilakukan untuk mencari *k-values* yang tepat dari algoritma *Improved K-NN*, dengan hasil dari perhitungan similiaritas dari setiap kategori diurutkan dahulu.

HASIL DAN PEMBAHASAN

Peneliti melakukan pengujian dengan 2 cara yaitu pertama pengujian rasio dengan menggunakan jumlah fitur dan kedua nilai k kemudian pengaruh dari seleksi fitur. Penelitian ini melakukan pengujian rasio dengan jumlah fitur menggunakan rasio fitur yaitu 25%, 50% dan 75%, kemudian nilai awal k yang di tentukan yaitu: k = 2, k = 3, k = 5, k = 10, k = 20 dan k = 30, k = 40, k = 50, k = 60 dan k = 70. Kemudian pengujian seleksi fitur *Chi-Square* dilakukan dengan rasio fitur dan juga nilai k yang sehingga pada pengujian pertama mendapatkan rata-rata yang paling baik. Kemudian hasil yang dilakukan perbandingan dengan tanpa seleksi fitur *Chi-Square* pada pengujian kedua.

1. Pengujian Rasio Jumlah Fitur

Berikut hasil dari pengujian dengan jumlah fitur yang telah direncanakan yaitu 25%, 50% dan 75%, disajikan pada Gambar 2.

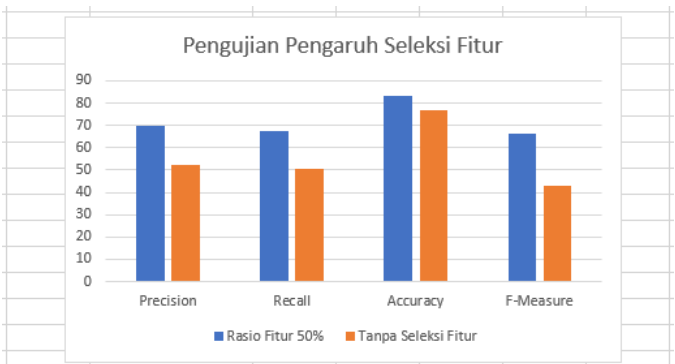
Gambar 2. Grafik dari hasil pengujian dengan rasio fitur 25%,50 % dan 75%

Dilihat dari grafik hasil pengujian dengan masing-masing rasio yang telah di dapatkan dapat dilihat fold dengan rata-rata terbaik. Untuk nilai yang terbaik didapatkan pada rasio dengan fitur 50%. Dari rasio tersebut menggunakan nilai k = 20. Hasil dari pengujian setelah diambil yang terbaik kemudian disimpulkan untuk *precision* 70,03%, *recall* 67,22%, nilai

accuracy 83,33% dan *f-measure* 66,26%. Hasil tersebut merupakan hasil yang telah dilakukan dengan menggunakan seleksi fitur *Chi-Square*.

2. Pengujian Pengaruh Seleksi Fitur

Pengujian yang kedua dilakukan perbandingan tanpa menggunakan seleksi fitur *Chi-Square*. Hasil pengujian seleksi fitur *Chi-Square* disajikan seperti pada Gambar 3.



	Precision	Recall	Accuracy	F-Measure
Rasio Fitur 50%	70,03968	67,22222	83,33333	66,26662
Tanpa Seleksi Fitur	52,41546	50,55556	77,08333	43,02326

Gambar 3. Grafik perbandingan hasil dengan menggunakan fitur *Chi-Square* dan tidak menggunakan fitur *Chi-Square*

Dari Gambar 3 dapat disimpulkan untuk perhitungan dengan menggunakan rasio fitur yang sama yaitu rasio fitur 50% dengan tanpa menambahkan seleksi fitur *Chi-Square* mendapatkan hasil *precision* 52,41%, *recall* 50,55%, accuarcy 77,08% dan *f-measure* 43,02%. Dapat disimpulkan dari gambar 5 bahwa pada pengujian pertama memiliki peforma lebih baik dibanding pengujian kedua. Artinya seleksi fitur *Chi-Square* dapat dikategorikan dokumen uji yang tepat dengan rasio kesalahan lebih kecil dari pada hasil yang diperoleh tanpa menggunakan seleksi fitur *Chi-Square*.

3. Evaluation

Dalam penelitian ini menggunakan pengujian *Confusion Matrix* untuk mengukur hasil evaluasi, pengukuran evaluasi dengan kriteria diantaranya*: f-measure, precission, recall* dan *accuracy* yang ditampilkan dalam bentuk presentase disajikan pada tabel 6.

Tabel 6. *Confusion Matrix Result 1*

	precision	***recall***	***accuracy***	***f-measure***
Rasio Fitur 25%	49,55%	47,22%	63,33%	46,26%
Rasio Fitur 50%	70,03%	67,22%	83,33%	66,26%
Rasio Fitur 75%	64,04%	61,22%	80,33%	63,26%

Pada tabel 6 tersebut menunjukkan pengujian pertama dengan menggunakan seleksi fitur *Chi-Square* dan diperoleh pengujian yang baik dengan rasio fitur 50% dengan hasil rata-rata *precision* 70,03%, rata-rata *recall* 67,22%, rata-rata nilai *accuracy* 83,33% dan rata-rata *f-measure* 66,26%. Sedangkan hasil pengujian kedua disajikan seperti pada tabel 7.

Tabel 7. *Confusion Matrix Result 2*

Rasio Fitur 50%	***precision***	***recall***	***accuracy***	***f-measure***
Dengan seleksi fitur *Chi-Square*	67,22%	69,87%	83,33%	66,26%
Tanpa seleksi fitur *Chi-Square*	52,41%	50,55%	77,08%	43,02%

Pada tabel 7 tersebut menunjukkan pengujian ke dua hasil menggunakan seleksi fitur Chi-Square yaitu *precision* 70,03%, *recall* 67,22%, nilai *accuracy* 83,33% dan *f-measure* 66,26%. Dan hasil tanpa menggunakan seleksi fitur Chi-Square yaitu yaitu *precision* 52,41%, *recall* 50,55%, nilai *accuracy* 77,08% dan *f-measure* 43,02%.

KESIMPULAN

Dalam penelitian melakukan percobaan dengan membandingkan literatur yang ada dan melakukan perbandingan kolaborasi antara penggunaan fitur *Chi-Square* dan tidak menggunakan fitur tersebut. Pengujian klasifikasi pada rasio yang jumlah fiturnya berbeda-beda yaitu rasio fitur 25%, 50% dan 75%. Sehingga diperoleh hasil yang baik dengan menggunakan rasio fitur 50% dengan nilai k yang di tetapkan k=20 menghasilkan *precision* 70,03%, *recall* 67,22%, nilai *accuracy* 83,33% dan *f-measure* 66,26%. Dapat disimpulkan dari hasil tersebut untuk fitur yang digunakan semakin rendah maka hasilnya menjadi kurang makimal begitu pula sebaliknya jika fitur yang digunakan terlalu tinggi maka hasilnya juga kurang maksimal. Setelah mendapatkan hasil yang baik dari klasifikasi *improved K-NN* dengan seleksi fitur *Chi-Square* kemudian pada penelitian ini juga melakukan perbandingan jika tidak menggunakan tambahan seleksi fitur *Chi-Square.* Dari perbandingan hasil antara menggunakan tambahan seleksi fitur *Chi-Square* dengan tanpa menggunnakan seleksi fitur *Chi-Square* terlihat bahwa fitur *Chi-Square* dapat menjadikan hasil dari penelitian lebih baik. Selain itu dapat juga dikatakan sebagai kategori dokumen secara tepat dengan rasio kesalahan yang lebih kecil. Saran untuk eksperimen di masa mendatang dapat membandingkan juga dengan seleksi fitur antara *Chi-Square* dengan fitur *Gain* atau seleksi fitur yang lain sesuai dengan literatur dari pakar yang sesuai pakar yang ada, agar nilai yang dihasilkan tepat dan sesuai.

DAFTAR PUSTAKA

[1] S. Sadya, "APJII: Pengguna Internet Indonesia 215,63 Juta pada 2022-2023," *dataindonesia.id*, 2023. https://dataindonesia.id/digital/detail/apjii-pengguna-internet-indonesia-21563-juta-pada-20222023

[2] Y. dkk Sri Mulyani, "Pemanfaatan Media Sosial *TikTok* Untuk Pemasaran Bisnis Digital Sebagai Media Promosi," *Penelit. manfaat media Sos. untuk Pemasar.*, vol. 11, no. 1, p. 3, 2022, [Online]. Available: http://stp-mataram.e-journal.id/JHI

[3] C. M. Annur, "Jumlah Pengguna *TikTok* Terus Bertambah, Ini Data Terbarunya," *databoks.katadata.co.id*,2022.https://databoks.katadata.co.id/datapublish/2022/09/06/jumlah-pengguna-*TikTok*-terus-bertambah-ini-data-terbarunya

[4] C. F. Hasri and D. Alita, "Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 3, no. 2, pp. 145–160, 2022, [Online].

Available: http://jim.teknokrat.ac.id/index.php/informatika

[5] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[6] M. N. Muttaqin and I. Kharisudin, "Analisis Sentimen Pada Ulasan Aplikasi Gojek Menggunakan Metode Support Vector Machine dan K Nearest Neighbor," *UNNES J. Math.*, vol. 10, no. 2, pp. 22–27, 2021, [Online]. Available: http://journal.unnes.ac.id/sju/index.php/ujm

[7] G. N. Bagaskoro, M. A. Fauzi, and P. P. Adikara, "Penerapan Klasifikasi Tweets Pada Berita Twitter Menggunakan Metode K-Nearest Neighbor Dan Query Expansion Berbasis Distributional Semantic," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 10, pp. 3849–3855, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[8] G. A. Tinega, P. W. Mwangi, and D. R. Rimiru, "Text Mining in Digital Libraries using OKAPI BM25 Model," *Int. J. Comput. Appl. Technol. Res.*, vol. 7, no. 10, pp. 398–406, 2018, doi: 10.7753/ijcatr0710.1003.

[9] D. Zakia Nathania and F. Abdurrachma Bachtiar, "Klasifikasi Spam Pada Twitter Menggunakan Metode Improved K-Nearest Neighbor," vol. 2, no. 10, pp. 3948–3956, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[10] M. Danil, S. Efendi, and R. Widia Sembiring, "The Analysis of Attribution Reduction of K-Nearest Neighbor (K-NN) Algorithm by Using *Chi-Square*," *J. Phys. Conf. Ser.*, vol. 1424, no. 1, 2019, doi: 10.1088/1742-6596/1424/1/012004.

[11] A. Purnamawati, M. N. Winnarto, and M. Mailasari, "Analisis Cart (Classification and Regression Trees) Untuk Prediksi Pengguna Sepeda Berdasarkan Cuaca," *J. Teknoinfo*, vol. 16, no. 1, p. 14, 2022, doi: 10.33365/jti.v16i1.1478.

[12] R. T. Handayanto, Herlawati, P. D. Atika, F. N. Khasanah, A. Y. P. Yusuf, and D. Y. Septia, "Analisis Sentimen Pada Situs Google Review dengan Naïve Bayes dan Support Vector Machine," *J. Komtika (Komputasi dan Inform.*, vol. 5, no. 2, pp. 153–163, 2021, doi: 10.31603/komtika.v5i2.6280.

[13] J. Pardede, M. Gustiana Husada, and R. Riansyah, "Implementasi Dan Perbandingan Metode Okapi BM25 Dan PLSA Pada Aplikasi Information Retrieval," *Itenas Repos.*, pp. 1–10, 2018.

[14] E. Tjandra and M. Widiasri, "Sistem Repositori Tugas Akhir Mahasiswa dengan Fungsi Peringkat Okapi BM25," vol. 2, no. 2, 2016, [Online]. Available: http://e-journal.unair.ac.id/index.php/JISEBI

[15] A. N. Mahardika, A. W. Widodo, and M. A. Rahman, "Diagnosis Penyakit Mata menggunakan Metode Improved K-Nearest Neighbor," vol. 3, no. 11, pp. 10531–10537, 2019.