Sentiment Analysis pada *Movie Review* Menggunakan *Feature Selection Mutual Information* dan *K-Nearest Neighbour Classifier*

Moch. Febriansyah Trisnadi[1], Said Al Faraby[2], Mahendra Dwifebri[3]

[1,2,3] Universitas Telkom, Bandung
[1]mfebriansyaht@students.telkomuniversity.ac.id, [2]saidalfaraby@telkomuniversity.ac.id, [3]mahendradp@telkomuniversity.ac.id,

Abstrak

Sentiment analysis* adalah sebuah cabang baru pada penambangan teks, yang meliputi kegiatan memproses dan mengekstraksi data dalam bentuk teks. *Sentiment analysis* memiliki tujuan agar dapat mengetahui apakah ulasan tersebut positif atau negatif. *Sentiment analysis* pun dapat digunakan dalam sebuah *movie review*. Melalui *movie review*, penonton dapat mengetahui kulitas film tersebut baik atau tidak. Untuk mendapatkan informasi tentang film, dibutuhkan banyak usaha bagi para penonton untuk membaca banyak *movie review*. Berdasarkan pada kondisi tersebut, sehingga membuat *sentiment analysis* pada *movie review* menjadi sebuah topik yang menarik diselesaikan dengan *machine learning*. Pada penelitian ini menerapkan *sentiment analysis* pada *movie review* berbahasa inggris menggunakan metode *K-Nearest Neighbor* dan *feature selection mutual information*. Dataset yang digunakan yaitu *Polarity* v2.0 dari *Cornell movie review dataset* . Pada penelitian ini menunjukkan bahwa penggunaan metode *K-Nearest Neighbor dan feature selection Mutual Information* mendapatkan nilai akurasi sebesar 73,32% dengan nilai K yaitu 32 dengan batas threshold 0,04 dan tanpa menggunakan *feature selection Mutual Information* mendapatkan akurasi sebesar 77,06%. Hal ini membuktikan bahwa *feature selection Mutual Information* tidak dapat meningkatkan performansi pada *K-Nearest Neighbor

Kata kunci : ***K-Nearest Neighbor, Movie Review, Mutual Information, Sentiment analysis.***

Abstract

Sentiment analysis is a new branch of text mining, which includes processing and extracting data in text form. Sentiment analysis has the aim of knowing whether the review is positive or negative. Sentiment analysis can also be used in a movie review. Through movie reviews, viewers can find out whether the quality of the film is good or not. To get information about movies, it takes a lot of effort for the audience to read a lot of movie reviews. Based on these conditions, making sentiment analysis on movie reviews an interesting topic, it is solved by machine learning. In this study, sentiment analysis is applied to the English-language movie review using the K-Nearest Neighbor method and feature selection mutual information. The dataset used is Polarity v2.0 from the Cornell movie review dataset. This study shows that the use of the K-Nearest Neighbor method and the Mutual Information feature selection get an accuracy value of 73.32% with a K value of 32 with a threshold limit of 0.04 and without using the Mutual Information feature selection an accuracy of 77.06%. This proves that the Mutual Information feature selection cannot improve performance on K-Nearest Neighbor
Keywords: ***K-Nearest Neighbor, Movie Review, Mutual Information, Sentiment analysis.***

1. Pendahuluan

Latar Belakang

Perkembangan zaman yang semakin pesat dan modern membuat *platform* dunia maya semakin beragam, dapat ditemukan banyak informasi seperti pandangan, perasaan, dan penilaian pada topik tertentu di dalam dunia maya. Informasi tersebut disimpan dalam bentuk teks, sehingga penambangan teks memiliki potensi nilai komersial yang tinggi [1].

Sentiment analysis adalah sebuah cabang baru pada penambangan teks, yang meliputi kegiatan memproses dan mengekstraksi data dalam bentuk teks [2]. *Sentiment analysis* memiliki tujuan agar dapat mengetahui apakah ulasan tersebut positif atau negatif [3]. *Sentiment analysis* pun dapat digunakan dalam sebuah *movie review*. Melalui *movie review*, penonton dapat mengetahui apakah *movie review* tersebut positif atau negative .

Untuk mendapatkan informasi tentang film, dibutuhkan banyak usaha bagi para penonton untuk membaca banyak *movie review*. Berdasarkan pada kondisi tersebut, sehingga membuat *sentiment analysis* pada *movie review* menjadi sebuah topik yang menarik diselesaikan dengan *machine learning*. Dengan menggunakan *machine learning* bisa membantu untuk mengklasifikasikan serta dapat mempersingkat waktu yang diproses, sehingga membuat suatu hal menjadi efektif dan efisien [4].

Pada penelitian sebelumnya [5] yang dilakukan oleh Octavani Faomasi Daeli dan Adiwijaya pada tahun 2019 yang berjudul *"Sentiment Analysis On Movie Reviews Using Information Gain And K-Nearest Neighbor*" memiliki tujuan untuk menemukan K optimal yang memiliki acuan pada ambang *Information Gain*, dan untuk menemukan ambang *Information Gain* terbaik. Pada penelitian ini digunakan polaritas v2.0 dari *dataset review*

film Cornell untuk menguji *K-Nearest Neighbor* melalui pemilihan fitur *Information gain* untuk mencapai kinerja yang baik. Pada penelitian ini disimpulkan bahwa K terbaik untuk *K-Nearest Neighbor* adalah sebesar 3 untuk dataset Polarity v2.0 . Pada penelitian ini hasil akurasi dari *K-Nearest Neighbor* lebih beasar dibandingkan dengan metode pembelajaran mesin lain seperti NB, SVM dan RF.

Berdasarkan penelitian sebelumnya, penulis tertarik menggunakan *K-Nearest Neighbor* (KNN) sebagai metode *machine learning* untuk digunakan pada penelitian ini. Pemilihan fitur yang digunakan yaitu *Mutual Information*. Karena pada penelitian sebelumnya [6] menunjukkan bahwa pemilihan fitur *Mutual Information* meningkatkan akurasi dari 96,2% menjadi 97,9%. Oleh karena itu, kombinasi antara *K-Nearest Neighbor* dan *Mutual Information* dapat meningkatkan akurasi pada penelitian ini.

Topik dan Batasannya

Dalam penelitian ini penulis akan membangun model untuk *sentiment analysis*pada *movie review*. Penelitian ini berfokus pada klasifikasi, *feature selection*, dan *preprocessing*. Pada klasifikasi penulis menggunakan metode *K-Nearest Neighbor*. Pada proses *feature selection*, penulis mencari perbedaan antara penggunaan *Mutual Information* dengan tidak menggunakan *Mutual Information*. Selanjutnya pada proses *preprocessing*, penulis membandingkan pengaruh antara penggunaan *stopword removal* dengan tidak menggunakan *stopword removal*. Adapun beberapa batasan masalah pada penelitian ini, yaitu dataset *movie review* berbahsa inggris dengan jumlah 2000 ulasan yaitu 1000 *positive review* dan 1000 *negative review.*

Tujuan

Tujuan dari penelitian ini adalah untuk membandingkan pengaruh dari sistem pengklasifikasian dengan menggunakan metode *K-Nearest Neighbor* (KNN) dan *feature selection Mutual Information* dibandingkan dengan tanpa penggunaan *feature selection Mutual Information*. Penelitian ini pun meneliti pengaruh penggunaan *stopword removal* pada tahap preprocessing terhadap hasil dari *sentiment analysis.*

Organisasi Tulisan

Pada bagian selanjutnya yaitu bagian 2 membahas tentang penelitian terkait, dan kajian pustaka, pada bagian 3 membahas perancangan sistem yang dibangun pada penelitian ini, pada bagian 4 membahas evaluasi dari hasil pemgujian yang telah dilakukan, dan pada bagian 5 membahas kesimpulan dari penelitian ini dan memberikan saran untuk penelitian selanjutnya.

2. Kajian Pustaka

2.1 Penelitian Terkait

Dalam penelitian ini menggunakan beberapa referensi, yaitu penelitian yang dilakukan oleh Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, dan Sweta Tiwari pada tahun 2016 yang berjudul *"Sentiment Analysis of Review Datasets using Naïve Bayes and K-NN Classifier"* [7] yang bertujuan untuk mengevaluasi kinerja untuk klasifikasi sentiment dalam hal *accuracy, precision and recall.* Para peneliti membandingkan dua algoritma pembelajaran mesin yaitu *naïve bayes* dan *K-Nearest Neighbor* untuk klasifikasi sentimen dari ulasan film dan ulasan hotel. Hasil percobaan menunjukkan bahwa *naïve bayes* memberikan hasil yang lebih baik dari *K-Nearest Neighbor* untuk ulasan film dengan nilai akurasi di atas 80%. Namun untuk ulasan hotel, keakuratannya banyak lebih rendah dan kedua pengklasifikasi menghasilkan hasil yang serupa. Para peneliti mengatakan pengklasifikasi *Naïve Bayes* dapat digunakan berhasil menganalisis ulasan film.

Pada penelitian berikutnya [5] yang dilakukan oleh Octavani Faomasi Daeli dan Adiwijaya pada tahun 2019 yang berjudul *"Sentiment Analysis On Movie Reviews Using Information Gain And K-Nearest Neighbor*" memiliki tujuan untuk menemukan K optimal yang memiliki acuan pada ambang *Information Gain*, dan untuk menemukan ambang *Information Gain* terbaik. Pada penelitian ini digunakan polaritas v2.0 dari *dataset review* film Cornell untuk menguji *K-Nearest Neighbor* melalui pemilihan fitur *Information gain* untuk mencapai kinerja yang baik. Pada penelitian ini disimpulkan bahwa K terbaik untuk *K-Nearest Neighbor* adalah sebesar 3 untuk dataset Polarity v2.0 . Pada penelitian ini *K-Nearest Neighbor* telah dibandingkan dengan metode pembelajaran mesin lain seperti NB, SVM dan RF.

Penelitian yang dilakukan oleh Sari Widya Sihwi, Insan Prasetya Jati dan Rini Anggrainingsih pada tahun 2018 [8] yang berjudul "*Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifi*er" memiliki tujuan untuk mendapatkan informasi apakah tweet tersebut merupakan opini positif, opini negatif, atau opini netral. Para peneliti menggunakan algoritma *naïve bayes classifier* karena kekuatan akurasinya serta dikombinasikan dengan metode pemilihan fitur *Information.* Para peneliti mengumpulkan dataset tweet dari 12 judul film popular. Hasil dari penelitian memiliki akurasi yang didapatkan dengan menggunakan metode *naïve bayes classifier* dan *information gain* sebesar 82,19% dengan 0,006 sebagai ambang *gain optimal*.

Pada penelitian berikutnya [6] yang dilakukan oleh Maria Arista Ulfa, Budi Irmawati, and Ario Yudo Husodo pada tahun 2018 yang berjudul "*Twitter Sentiment Analysis using Naive Bayes Classifier with Mutual Information Feature Selection*". Pada penelitian ini penulis menggunakan metode *Naïve Bayes Classifier* dengan fitur seleksi

mutual information. Alasan para penulis menggunakan *Mutual Information* sebagai fitur seleksi karena *Mutual Information* membantu memilih fitur yang memiliki kontribusi tinggi. Pada penelitian tersebut Hasil penelitian menunjukkan bahwa pemilihan fitur *Mutual Information* meningkatkan akurasi dari 96,2% menjadi 97,9%. Ini juga berkontribusi untuk meningkatkan presisi dan perolehan. Selain itu, waktu klasifikasi juga berkurang 51,52%.

2.2 Sentiment Analysis

Sentiment Analysis merupakan kombinasi antara *data mining* dan *text mining*, atau atau cara untuk mengolah berbagai opini yang diberikan oleh pengguna atau para pakar melalui berbagai media, mengenai sebuah produk, jasa ataupun sebuah instansi. Untuk mendapatkan sebuah *sentiment* yang terkandung dalam sebuah opini digunakan *sentiment analysis* sebagai sebuah metode untuk memahami, mengolah data opini, dan mengolah data tekstual secara otomatis [9]. Ada 2 jenis *sentiment* pada *movie review* yaitu *sentiment* positif dan *sentiment* negatif. *Sentiment* positif pada *movie review* menunjukan bahwa *movie review* tersebut menunjukan ulasan yang positif, sedangkan *sentiment* negatif menunjukkan bahwa *movie review* tersebut menunjukan ulasan yang negatif.

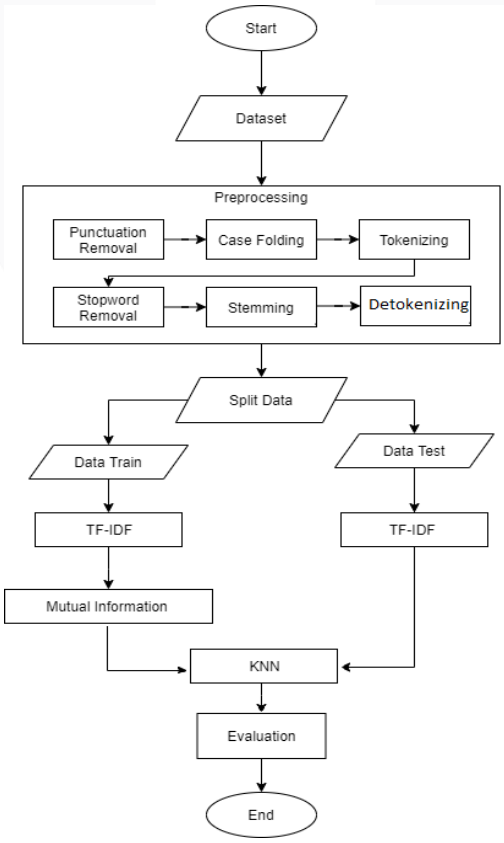
2.3 Movie Review

Website dapat menjadi tempat bagi pengguna internet untuk mengemukakan pendapatnya. Maka dari itu banyak opini yang muncul dari pengguna internet mengenai suatu hal yang spesifik. Salah satunya yaitu informasi mengenai film. Pengguna internet dapat mengetahui informasi mengenai suatu film melalui pencarian website yang berisikan pendapat-pendapat penulis atau disebut juga *movie review*. Namun diantara pendapat dari setiap penulis tidak akan selalu sama antara satu dengan yang lain, karena banyaknya pendapat yang muncul pada *website* maka akan semakin sulit untuk menemukan informasi penting yang sesuai kebutuhan pengguna. Tetapi jika data *movie review* diolah secara baik maka akan didapatkan informasi mengenai kualitas film. Proses klasifikasi informasi *movie review* dapat mempermudah pengguna untuk menarik kesimpulan berdasarkan pada opini orang lain [10].

3. Sistem yang Dibangun

Pada penelitian ini, melakukan proses preprocessing meliputi, *punctuation removal, case folding, tokenizing, stopword removal, stemming* dan *detokenizing, split dataset, feature extraction* menggunakan TF-IDF, *feature selection* menggunakan *mutual information, classification* menggunakan *K-Nearest Neighbor* dan *evaluation* menggunakan *confusion matrix*. Berikut merupakan gambaran umum sistem yang dibangun:

Gambar 1. Gambaran Umum SIstem



3.1 Dataset

Dataset yang digunakan pada penelitin yaitu *Polarity* v2.0 dari *Cornell movie review dataset* dan pada dataset tersebut telah dilabeli secara manual dengan batas 20 ulasan per penulis (jumlah 312 penulis) per kategori [19]. *Dataset* tersebut berisi 2000 ulasan film berbahasa inggris dan mempunyai 2 label yaitu label 1 untuk *movie review* positif yang berjumlah 1000 dataset dan label 0 untuk *movie review* negatif yang berjumlah 1000 dataset.

3.2 Preprocessing

Preprocessing merupakan langkah langkah dalam mengolah *dataset* yang berikutnya akan dimasukkan kedalam sistem klasifikasi Berikut adalah tahapan yang penulis lakukan pada tahap *preprocessing*.

a) Punctuation Removal
Pada tahap ini, dilakukan pembersihan data dengan menghapus tanda baca, angka atau karakter khusus selain kata.

b) Case Folding
Tahapan ini melakukan konversi sebuah teks ke dalam bentuk yang sama yaitu diubahnya semua huruf menjadi huruf kecil. Pada tahap ini dilakukan agar membuat konsistensi pada teks untuk mendapatkan hasil yang lebih optimal.

c) Tokenizing
Tahapan ini terdapat proses memisahkan sebuah kalimat menjadi potongan kata yang terpisah.

d) Stopword Removal
Tahapan ini yaitu proses penghapusan teks yang tidak memiliki data dan hanya mengambil kata-kata penting seperti pada kata umum yang biasanya muncul dalam jumlah besar

e) Stemming
Tahapan ini, dilakukan pemotongan kasar yang memiliki tujuan mereduksi sebuah kata menjadi bentuk kata dasarnya atau menghapus kata imbuhan.

f) Detokenizing
Tahapan ini terdapat proses mengembalikan potongan kata yang terpisah menjadi sebuah kalimat.

3.3 Split Dataset

Pada tahap *split data*, *dataset* tersebut akan dibagi menjadi 2 yaitu *data train* dan *data test*, dimana *data train* menggunakan data sebanyak 1600 data *movie review* dan data test akan menggunakan data sebanyak 400 data *movie review*.

3.4 *Feature Extraction* TF-IDF

Dalam tahap ini, dataset akan melalui metode feature extraction TF-IDF. Metode TF-IDF (*Term Frequency – Inverse Document Frequency*) dipakai agar dapat memberi bobot pada term sebagai strategi untuk pengklasifikasian sebuah dokumen [11]. Prosesnya antara lain terdiri dari menghitung nilai TF *(term frequency*) dan IDF (*Inverse Document Frequency*). Kegunaan dari TF (*term frequency*) adalah menghitung frekuensi dari kemunculan kata yang selanjutnya dimasukan ke dalam nilai log tf, sedangkan IDF (*Inverse Document Frequency*) adalah perhitungan dari suatu term yang selanjutnya didistribusikan pada keseluruhan dokumen. TF-IDF pun berfungsi menghitung bobot pada setiap kata yang berasal dari *preprocessing*. Untuk mengesktrak kata menjadi format numerik atau angka, digunakan pembobotan kata tersebut yang berfungsi mempresentasikan data secara menyeluruh [12]. Berikut merupakan proses perhitungan pembobotan kata TF-IDF [13]

$$tf_{t,d} = f_{t,d} \tag{1}$$

Berdasarkan persamaan (1), tf adalah jumlah kemunculan term pada suatu dokumen.

$$idf_t = log \frac{N}{df_t} \tag{2}$$

Persamaan (2) adalah perhitungan untuk idf, dimana N adalah jumlah dokumen, sedangkan df merupakan jumlah dokumen yang mengandung term t

$$w = tf_{t,d} \times idf_t \tag{3}$$

Persamaan (3), w adalah perhitungan bobot nilai dari setiap term, dimana tf adalah jumlah kemunculan term dalam suatu dokumen dan idf adalah nilai *inverse document frequency*.

3.5 Feature Selection Mutual Information

. Pada penelitian ini, penulis menggunakan *feature selection Mutual information. Mutual information* merupakan metode seleksi fitur yang menghitung seberapa banyak informasi yang terkandung dalam term, sehingga berkontribusi untuk membuat klasifikasi yang tepat [14]. Pada tahap ini, dataset akan dilakukan pemilihan fitur-fitur yang paling relevan terhadap masing-masing kelas untuk digunakan pada proses klasifikasi. Berikut merupakan proses perhitungan nilai Mutual Information [15]

$$\mathrm{I(U,C)} = \sum_{et\in\{1.0\}} \sum_{ec\in\{1.0\}} P(U=et, C=ec) log_2 \frac{P(U=et, C=ec)}{P(U=et)P(C=ec)} \quad (4)$$

Berdasarkan persamaan (4), perhitungan nilai *Mutual Information* dapat diuraikan lebih detail menjadi persamaan (5) seperti berikut :

$$I(U,C) = \frac{N_{11}}{N} log_2 \frac{N.N_{11}}{N_1.N_1} + \frac{N_{01}}{N} log_2 \frac{N.N_{01}}{N_0.N_1} + \frac{N_{10}}{N} log_2 \frac{N.N_{10}}{N_1.N_0} + \frac{N_{00}}{N} log_2 \frac{N.N_{00}}{N_0.N_0} \quad (5)$$

Keterangan :

- N = Jumlah dokumen yang memiliki et (term t) dan ec (kelas c) atau ($N_{00} + N_{01} + N_{10} + N_{11}$)
- N_1. = jumlah dokumen yang memiliki et (term t) atau ($N_1. = N_{10} + N_{11}$)
- N_1 = Jumlah dokumen yang memiliki ec (kelas c) atau ($N_1 = N_{01} + N_{11}$)
- N_0. = Jumlah dokumen yang tidak memiliki et (term t) atau ($N_0. = N_{01} + N_{00}$)
- N_0 = Jumlah dokumen yang tidak memiliki ec (kelas c) atau ($N_0 = N_{10} + N_{00}$)

3.6 Clasification K-Nearest Neighbor

Tahap selnajutnya adalah klasifikasi. Pada tahap ini penulis menggunakan Algoritma *K-Nearest Neighbor* (KNN) untuk melakukan klasifikasi. *K-Nearest Neighbor* adalah sebuah metode untuk melakukan pengelompokan terhadap objek berdasarkan pada data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Terdapat beberapa kelebihan dari algoritma *K-Nearest Neighbor* yaitu memiliki ketahanan terhadap data latih yang memiliki banyak noise dan efektif apabila data latih tersebut besar [16]. Jarak jauh atau dekatnya dapat dihitung berdasarkan *Euclidean Distance* yang direpresentasikan pada persamaan berikut:

$$distance = \sqrt{\sum_{i=1}^{n} (x^i_{training} - x_{testing})^2} \quad (6)$$

Keterangan :

- $x^i_{training}$: data training ke-i,
- $x_{testing}$: data testing,
- i : record (baris) ke-i
- n : jumlah data training.

3.7 Evaluation Confusion Matrix

Confusion matrix berisi tentang informasi mengenai klasifikasi yang diprediksi secara benar oleh sebuah sistem klasifikasi [17]. *Confusion matrix* tersebut hasilnya menjadi 4 bagian yaitu *false positive, true positif, false negative,* dan *true negative.* Berikut tabel contoh dari *confusion matrix*[18].

Tabel 1. Confusion Matrix

	Actual Positive	*Actual Negative*
Predicted Positive	*True Positive (TP)*	*False Positive (FP)*
Predicted Negative	*False Negative (FN)*	*True Negative (TN)*

Keterangan dari :

- *True positive* disebut sebagai data yang diprediksi positif dan klasifikasi aktual positif.
- *False positive* disebut sebagai data yang diprediksi positif dan tetapi klasifikasi aktualnya negatif
- *False Negative* disebut sebagai data yang diprediksi negatif tetapi aktualnya positif.
- *True negative* disebut sebagai data yang diprediksi negatif dan klasifikasi aktual negative..

Berdasarkan 4 hasil diatas, bisa dihitung evaluasi berdasarkan klasifikasi yang telah dikerjakan. Terdapat rumus-rumus untuk menghitung evaluasi yaitu :

- *Accuracy* merupakan seberapa akurat model memprediksi data prediksi dari data sebenarnya. Berikut rumus dari *accuracy* :

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (7)$$

- *Precision* merupakan seberapa akurat model memprediksi data positif dari data sebenarnya yang positif. Berikut rumus dari *precision* :

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

- *Recall* merupakan dengan seberapa banyak data positif yang dapat ditangkap dari model klasifikasi. Berikut rumus dari *recall* :

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

- *F1 Score* merupakan perbandingan rata-rata *precision* dan *recall*. Berikut merupakan rumus dari *F1 Score*

$$\text{F1 Score} = \frac{2(Recall * Precision)}{(Recall + \text{Precision})} \quad (10)$$

4. Evaluasi

Dalam tahap evaluasi, penulis melakukan 3 skenario pengujian. Pengujian skenario yang pertama dilakukan untuk mengetahui pengaruh *stopword removal* pada tahap *preprocessing*. Pengujian skenario kedua dilakukan untuk mengetahui pengaruh dari *feature selection mutual information* pada metode *K-Nearest Neighbor*. Pengujian skenario yang ketiga dilakukan agar dapat mengetahui pengaruh anatara nilai K terhadap hasil klasifikasi dengan metode *K-Nearest Neighbor*.

4.1 Pengujian Pengaruh *Stopword Removal*

Pada pengujian skenario yang pertama ini dilakukan untuk mengetahui pengaruh dari proses pada penggunaan *preprocessing stopword removal*. .Pada skenario pertama ini penulis melakukan 2 pengujian *dataset.* yang pertama melakukakan pengujian *dataset* menggunakan *stopword removal* dan yang kedua melakukan pengujian dataset tanpa tahap *stopword removal*. *Dataset* yang digunakan sudah dibagi menjadi 2 yaitu *data train* 80% dan *data test* 20% , menggunakan *feature selection mutual information* dengan batas threshold 0,04, dan menggunakan metode *K-Nearest Neighbor* dengan nilai K = 32. Berikut merupakan hasil dari pengujian skenario pertama :

Tabel 2 Hasil Pengujian Skenario 1

Stopword	Akurasi	*Precission*	*Recall*	*F1-Score*
Y	**73,32%**	73,30%	73,21%	73,25%
-	52,12%	76,00%	50,26%	60,50%

Dapat dilihat pada table diatas nilai akurasi dengan menggunakan mutual information dan menggunakan *stopword removal* mendapatkan nilai akurasi 73% sedangkan dengan tanpa menggunakan *stopword removal* mendapatkan nilai akurasi 52,11% .

Tabel 3 Top 10 fitur tanpa menggunakan stopword removal

Terms	**TF-IDF**	**Nilai MI**
the	481.317888	0.119297
and	224.347085	0.642722
of	213.989785	0.340085
to	199.718287	0.310078
is	160.148584	0.189631
in	136.208862	0.000000
it	118.113142	0.083193
that	101.376915	0.000000
film	77.558486	0.205188
as	74.763374	0.305213

Dapat dilihat pada tabel 4 bahwa fitur yang termasuk list stopword pada percobaan tanpa menggunakan *stopword removal* memiliki nilai bobot yang cukup besar, kemungkinan hal tersebut menyebabkan sistem tidak stabil sehingga menyebabkan salah dalam klasifikasi dan menyebabkan hasil akurasi menjadi kecil.

4.2 Pengujian Pengaruh *Feature Selection*

Pada Pengujian skenario kedua ini, dilakukan untuk mengetahui pengaruh *feature selection mutual information* pada metode *K-Nearest Neighbor* yang dapat memberikan hasil akurasi yang terbaik. Dataset pada pengujian sekenario pertama ini telah melalui tahap *preprocessinng*, tahap split data data dengan cara membagi *dataset* menjadi 2 yaitu data train 80% dan data test 20%, tahap *feature extraction* menggunakan TF-IDF, pengklasifikasian menggunakan *K-Nearest Neighbor dan* pada pengujian skenario kedua ini penguji menggunakan nilai K yaitu K=32. Berikut merupakan hasil dari pengujian skenario kedua :

Tabel 4 Hasil pengujian perbandingan *Feature Selection*

Threshold	**Fitur yang digunakan**	**Akurasi**	***Precission***	***Recall***	***F1-Score***
Tanpa Menggunakan *Feature Selection*	27897	**77,06%**	77,06%	77,10%	77,07%
0,01	13513	55,86%	77,01%	54,15%	63,58%
0,02	12990	58,10%	66,70%	56,66%	61,27%
0,03	12487	68,83%	72,99%	69,58%	71,24%
0,04	12029	73,32%	73,30%	73,21%	73,25%
0,05	11585	56,86%	63,81%	55,41%	59,31%

Dapat dilihat pada tabel 2 yaitu hasil pengujian pengaruh terhadap *feature selection mutual information,* pada percobaan pertama yaitu tidak menggunakan *feature selection mutual information* mendapatkan hasil akurasi tertinggi sebesar 77,06%. Pada percobaan kedua menggunakan *feature selection* dengan nilai *threshold* 0,01 mendapatkan hasil akurasi sebesar 55,86%. Pada percobaan ketiga menngunakan nilai *threshold* 0,02 mendapatkan hasil akurasi sebesar 58,10%. Pada percobaan keempat menggunakan nilai *threshol*d 0,03 mendapatkan hasil akurasi sebesar 68,83%. Pada percobaan kelima menggunakan nilai *threshold* 0,04 mendapatkan hasil akurasi sebesar 73,32%. Dan pada percobaan terahir menggunakan nilai *threshold* 0,05 mendapatkan hasil akurasi sebesar 56,86%.

Tabel 5 Perbandingan akurasi *data train* dan *data test*

	Test	*Train*
Tanpa *Mutual Information*	77,06%	77,00%
Mutual Information Threshold 0,01	55,86%	56,06%
Mutual Information Threshold 0,02	58,10%	63,31%
Mutual Information Threshold 0,03	68,83%	74,31%
Mutual Information Threshold 0,04	73,32%	76,94%
Mutual Information Threshold 0,05	56,86%	63,81%

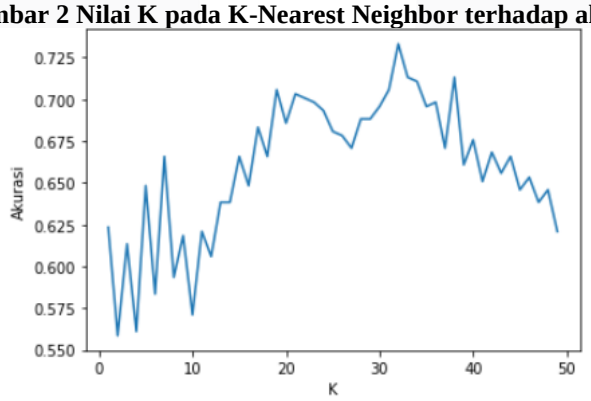
Dapat dilihat pada tabel diatas bahwa semua hasil akurasi dengan menggunakan *mutual information* mengalami *overfitting*, sehingga hasil akurasi dengan menggunakan *mutual information* mendapatkan hasil yang lebih rendah dibandingkan dengan tanpa menggunakan *mutual information*. *Overfitting* pada model tersebut kemungkinan disebabkan karena pada model yang menggunakan *mutual information* terlalu fokus terhadap data training, sehingga model tersebut tidak bisa memprediksi dengan tepat jika diberikan dataset lain dan mendapatkan hasil akurasi yang rendah.

Sehingga dapat disimpulkan dari pengujian skenario bahwa penggunaan *feature selection mutual information* tidak dapat bekerja dengan baik pada metode *K-Nearest Neighbor* karena pada model dengan menggunakan *mutual information* mengalami *overfitting* yang menyebabkan hasil akurasinya rendah dibandingkan dengan tanpa menggunakan *feature selection mutual information*.

4.3 Pengujian Pengaruh Nilai K pada *K-Nearest Neighbor*

Pada Pengujian skenario pertama ini, dilakukan untuk mengetahui nilai K yang optimal pada metode *K-Nearest Neighbor* yang dapat memberikan hasil akurasi paling optimal. *Dataset* pada pengujian sekenario pertama ini telah melalui tahap *preprocessinng*, tahap *split data* data dengan cara membagi dataset menjadi 2 yaitu *data train* 80% dan *data test* 20%,, *feature extraction* menggunakan TF-IDF dan menggunakan *feature selection mututal information* dengan batas threshold 0,04. Pada skenario ini, penulis membandingkan nilai K pada metode *K-Nearest Neighbor* dalam rentang 1-50.

Gambar 2 Nilai K pada K-Nearest Neighbor terhadap akurasi



Tabel 6 Nilai K dengan akurasi tertinggi

Nilai K	Akurasi	Precission	Recall	F1-Score
32	73,32%	73,30%	73,21%	73,25%

Dapat dilihat pada gambar 2, penulis melakukan percobaan untuk mencari nilai K terbaik pada metode *K-Nearest Neighbor*. Percobaan tersebut mendapatkan nilai K yang paling optimal yaitu K=32 dan mendapatkan akurasi sebesar 73,32%.. Berdasarkan pengujian skenario ketiga dengan membandingkan nilai K pada metode *K Nearest Neighbor* dalam rentang 1-50 dapat disimpulkan bahwa untuk mendapatkan tingkat akurasi yang optimal dengan menggunakan metode *K-Nearest Neighbor* diperlukan untuk menemukan nilai K yang tepat.

5. Kesimpulan dan Saran

Berdasarkan hasil dari ketiga skenario pengujian yang telah dilakukan untuk *sentiment analysis* pada *movie review* berbahasa inggris dengan melakukan proses *preprocessing* (*punctuation removal, case folding,tokenizing,stopword removal, stemming, detokenizing*), *feature extraction TF-IDF, feature selection mutual information* menggunakan batas nilai threshold 0,04 dan klasifikasi menggunakan metode *K-Nearest Neighbor* dengan nilai K=32 menghasilkan akurasi sebesar 73,32%. Sedangkan tanpa menggunakan *feature selection mutual information* mendapatkan nilai akurasi sebesar 77,06%. Dapat disimpulkan bahwa penggunaan *feature selection mutual information* tidak dapat meningkatkan akurasi klasifikasi dari *K-Nearest Neighbor* karena pada model dengan menggunakan *mutual information* mengalami *overfitting* yang menyebabkan hasil akurasinya rendah dibandingkan dengan tanpa menggunakan *feature selection mutual information*. Selain itu, untuk mendapatkan tingkat akurasi yang optimal dengan menggunakan metode *K-Nearest Neighbor* diperlukan untuk menemukan nilai K yang tepat dan pada *preprocessing* penggunaan *stopword remova*l dapat meningkatkan akurasi pada penelitian ini karena pada *stopword removal* dapat menghapus kata-kata yang tidak penting karena kata-kata yang tidak penting mempunyai bobot nilai TF-IDF yang besar, kemungkinan hal tersebut menjadi penyebab salah dalam klasifikasi dan system menjadi tidak stabil saat melakukan prediksi .

Saran untuk penelitian selanjutnya, agar mendapatkan hasil yang lebih baik diharapkan menambahkan jumlah dataset, hal tesebut dilakukan agar bertambahnya variasi data. Dan juga melakukan kombinasi metode *K-Nearest Neighbor* dengan fitur seleksi selain *mutual information* sehingga mendapatkan hasil yang lebih optimal.

Referensi

[1] M. M and S. Mehla, "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers," *Int. J. Comput. Appl.*, vol. 182, no. 50, pp. 25–28, 2019, doi: 10.5120/ijca2019918756.

[2] Y. Nurdiansyah, S. Bukhori, and R. Hidayat, "Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method," *J. Phys. Conf. Ser.*, vol. 1008, no. 1, 2018, doi: 10.1088/1742-6596/1008/1/012011.

[3] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis: Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2. 2008.

[4] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002, doi: 10.1145/505282.505283.

[5] N. Octaviani Faomasi Daeli, "Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor," *J. Data Sci. Its Appl.*, vol. 3, no. 1, pp. 1–007, 2020, [Online]. Available: http://commdis.telkomuniversity.ac.id/jdsa/index.php/jdsa/article/view/22.

[6] P. T. Informatika, F. Teknik, and U. Mataram, "Twitter Sentiment Analysis using Na ¨ ıve Bayes Classifier with Mutual Information Feature Selection," vol. 2, no. 2, pp. 106–111, 2018.

[7] L. Dey, "Sentiment Analysis of Review Datasets Using Naïve Bayes ' and K -NN Classifier," no. December, 2016, doi: 10.5815/ijieeb.2016.04.07.

[8] S. Widya Sihwi, I. Prasetya Jati, and R. Anggrainingsih, "Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 190–195, 2018, doi: 10.1109/ISEMANTIC.2018.8549757.

[9] E. M. Sipayung, H. Maharani, I. Zefanya, and D. S. Informasi, "No Title," vol. 8, no. 1, pp. 958–965, 2016.

[10] W. C. Widyaningtyas, A. Adiwijaya, and S. Al Faraby, "Klasifikasi Sentiment Analysis Pada Review Film Berbahasa Inggris Dengan Menggunakan Metode Doc2vec Dan Support Vector Machine (svm)," *eProceedings Eng.*, vol. 5, no. 1, pp. 1570–1578, 2018.

[11] D. Zhu and J. Xiao, "R-tfidf , a Variety of tf-idf Term Weighting Strategy in Document Categorization," pp. 83–90, 2011, doi: 10.1109/SKG.2011.44.

[12] J. T. Medler, "The types of Flatidae (Homoptera) in the Stockholm Museum described by Stål, Melichar, Jacobi and Walker," *Insect Syst. Evol.*, vol. 17, no. 3, pp. 323–337, 1986, doi: 10.1163/187631286X00251.

[13] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 406–414, 2021.

[14] T. Akhir, "Analisis Pengaruh Seleksi Fitur Information Gain dan Mutual Information pada Klasifikasi Sentimen Ulasan Film Menggunakan Support Vector Machine Program Studi Sarjana S1 Informatika Fakultas Informatika Universitas Telkom Bandung," 2019.

[15] A. Hanafi, A. Adiwijaya, and W. Astuti, "Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, pp. 357–364, 2020, doi: 10.32736/sisfokom.v9i3.980.

[16] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," vol. 9, no. 1, pp. 57–68, 2012.

[17] 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.

[18] M. Awad and R. Khanna, *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, no. May 2016. 2015.

[19] Pang B and Lee L *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts* 2004 *July* ACL 271

-.