Analisis Sentimen Pengaruh Kombinasi Ekstraksi Fitur TF-IDF dan *Lexicon* Pada Ulasan Film Menggunakan Metode KNN

Setyo Adji Pratomo[1], Said Al Faraby[2], Mahendra Dwifebri Purbolaksono[3]

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

Abstrak

Industri film selalu berkembang dengan cepat seiring bertambahnya waktu. Terdapat banyak pertimbangan untuk menentukan film yang berkualitas, ulasan film merupakan salah satu faktor yang memiliki peran penting untuk menentukan film yang berkualitas. Situs IMDb merupakan salah satu platform yang digunakan untuk menampung sentimen seseorang terhadap film. Metode machine learning dapat digunakan untuk memudahkan kita dalam hal meringkas atau melakukan analisis terhadap banyak opini yang ada dalam platform tersebut. Dalam Tugas Akhir ini dibangun sistem analisis sentiment dengan menggunakan metode machine learning KNN dengan menggabungkan fitur ekstraksi TF-IDF dan Lexicon SentiWordNet. Data yang digunakan dalam Tugas Akhir ini adalah data ulasan film dari website iMDB sebanyak 2000. Hasil akhir dari penelitian yang dilakukan ini yaitu analisis terhadap penggabungan metode fitur ekstraksi TF-IDF dengan Lexicon SentiWordnet dan menguji penggunaan fitur seleksi *Information Gain* (IG). Dari hasil percobaan yang telah dilakukan, penggabungan fitur ekstraksi TF-IDF dengan Lexicon SentiWordnet memiliki hasil akurasi yang tidak lebih tinggi dibandingkan dengan hanya menggunakan fitur ekstraksi TF-IDF yaitu 73.31%, dan penggunaan fitur seleksi IG dengan *threshold* yang tepat mampu mengoptimasi hasil performansi.

Kata kunci : analisis sentimen, KNN, *Lexicon SentiwordNet,* TF-IDF, ulasan film.

Abstract

The film industry has always developed rapidly over time. There are many considerations to determine a quality film, film reviews are one of the factors that have an important role in determining a quality film. The IMDb site is one of the platforms used to accommodate one's sentiments towards films. Machine learning methods can be used to make it easier for us to summarize or analyze the many opinions that exist on the platform. In this final project, a sentiment analysis system was built using the KNN machine learning method by combining the extraction features of TF-IDF and Lexicon SentiWordNet. The data used in this final project is 2000 film review data from the iMDB website. The final result of this research is an analysis of combining the TF-IDF feature extraction method with Lexicon SentiWordnet and testing application selection feature Information Gain (IG). From the results of experiments, combining the TF-IDF extraction feature with Lexicon SentiWordnet has an accuracy result that is not higher than using only the TF-IDF extraction feature, which is 73.31%, and the use of the IG selection feature with the right threshold is able to optimize the performance results.

Keywords: sentiment analysis, KNN, *Lexicon SentiwordNet,* TF-IDF, movie review.

1. Pendahuluan [10 pts/Bold]

Latar Belakang

Industri film selalu berkembang dengan cepat seiring bertambahnya waktu, selain menjadi media hiburan, film dapat dijadikan sebagai media untuk memberikan edukasi. Dengan pilihan film yang beragam seringkali membuat seseorang bingung untuk memilih film yang akan mereka saksikan. Terdapat banyak pertimbangan untuk menentukan film yang berkualitas, ulasan film merupakan salah satu faktor yang memiliki peran penting untuk menentukan film yang berkualitas. Untuk menentukan film yang sesuai, seseorang membuat keputusan berdasarkan pengalaman masa lalu, sentimen atau opini yang dilalui orang lain sebelumnya[1]. Dengan sentimen analisis membuat tugas peringkasan opini lebih mudah dengan mengekstrak sentimen yang diungkapkan oleh pengulas[2].

Trend sentimen analisis berkembang dengan signifikan seiring pesatnya penggunaan jejaring sosial, aplikasi, dan forum[3]. Sentimen analisis telah digunakan pada setiap bisnis. Karena, opini merupakan kegiatan manusia yang sentral untuk menentukan sebuah keputusan[4]. Sekarang, hampir seluruh *platform* menyediakan kolom komentar untuk meninggalkan pesan terkait sebuah produk, pelayanan ataupun film[3]. Salah satu *platform* tersebut adalah situs iMDb, website ini menjadi salah satu acuan seseorang dalam menentukan film

yang akan mereka saksikan, karena berisikan sekumpulan opini orang lain terhadap film yang mereka saksikan. Sentimen analisis mempermudah memperoleh kesimpulan dari banyaknya data yang orang lain berikan terhadap sebuah film.

Sentimen analisis dapat diselesaikan dengan beberapa metode *machine learning*, seperti *Naive Bayes* (NB), *K-nearest neighbor* (KNN), *Support vector machine* (SVM), *Random forest*, dan *Maximum Entropy*. Salah satu contoh penelitiannya yaitu *Sentiment Analysis on Twitter Data using* KNN *and* SVM [5]. Pada penelitian ini, metode *machine learning* yang akan digunakan adalah KNN, metode KNN memberikan hasil yang kurang baik pada proses klasifikasi data karena terdapat fitur noise [6], namun terdapat beberapa penelitian yang menyebutkan bahwa performansi pada metode *machine learning* dapat menghasilkan performa yang baik ketika dikombinasikan dengan ekstraksi fitur dan seleksi fitur yang tepat [6]. Lalu, penerapan kombinasi fitur ekstraksi TF-IDF dan *lexicon* SentiwordNet menghasilkan perfomansi yang baik saat diuji menggunakan metode klasifikasi *Naïve Bayes* (NB) yaitu dengan akurasi 84.75% dibandingkan dengan hanya menggunakan TF-IDF saja yaitu 84.62% [7]. *Dataset* yang digunakan dalam penelitian ini adalah ulasan film pada website iMDb sejumlah 2000 diantaranya 1000 label positif dan 1000 label negatif berbahasa inggris. Tahapan awal pada penelitian ini adalah *data pre-processing* bertujuan agar data yang nanti diproses terstruktur, selanjutnya proses fitur ekstraksi menggunakan metode TF-IDF dan *lexicon*, lalu terakhir yaitu tahap klasifikasi.

2. Studi Terkait

Sentimen analisis merupakan proses komputasi untuk mengidentifikasi dan mengkatagorikan opini yang mendefinisikan ungkapan penulis biasanya berupa ungkapan positif ataupun negatif untuk sebuah produk, film, dan bahasan lainnya. Selain itu, implementasi sentiment analisis terhadap suatu ulasan produk, film, dan lainnya itu penting untuk memahami dan mendeskripsikan ungkapan seseorang terhadap sebuah produk baik itu untuk pembuat produk atau film untuk membenahi kesalahan dimasa depan[8].

Pada jurnal [6] didapatkan metode K-Nearest Neighbour (KNN) menghasilkan performansi yang baik menggunakan *dataset* ulasan film, dibandingkan dengan metode lainnya yaitu Naive Bayes (NB) dan Support Machine Vector (SVM) yang dilakukan pada penelitian tersebut. Namun, performansi hasil klasifikasi dapat ditingkatkan dengan pemilihan ekstraksi fitur dan seleksi fitur yang tepat. Selain itu, untuk mengoptimasi hasil klasifikasi pada penelitian [9] diperlukan untuk mencari *nearest neighbor* yang tepat agar hasil klasifikasi menjadi lebih efisien dan efektif.

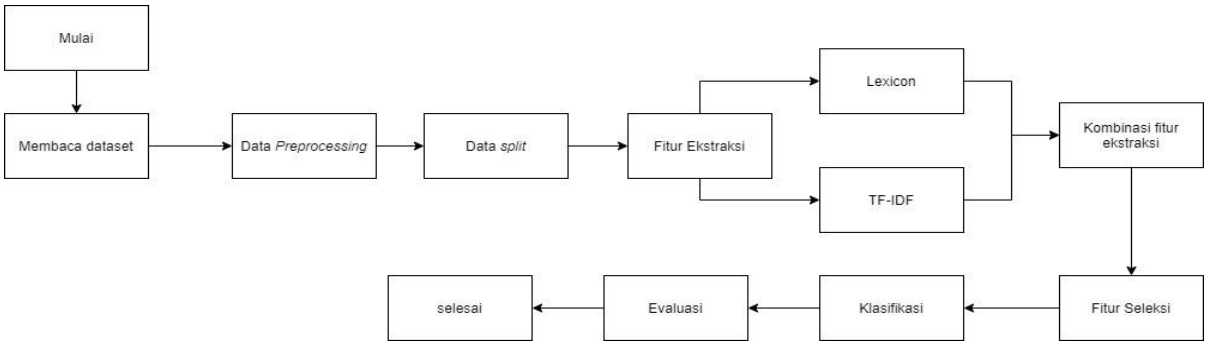
Sudah banyak peneliti yang membuat model sentimen analisis untuk meneliti yang mengidentifikasi antara sentimen positif dan sentimen negatif, pada penelitian [7] kombinasi fitur ekstraksi *Term Frequency — Inverse Document Frequency* (TF-IDF) dan *Lexicon* memberikan hasil yang bagus terhadap akurasi ketika digunakan pada metode machine learning *Naïve Bayes* (NB). Dalam penelitian [10] SentiWordNet digunakan untuk mendapatkan akurasi lebih dari teknik *lexicon* yang lain untuk analisis sentimen untuk ulasan pelanggan dan ulasan perangkat lunak.

Pada penelitian [6] fitur seleksi menggunakan *Information Gain* (IG) dapat mengoptimasi kinerja KNN, dikarenakan IG mampu menghapus fitur yang tidak berguna untuk proses klasifikasi. Disamping kinerja KNN yang cukup sederhana dengan membandingkan jarak antar objek namun hasil klasifikasi KNN bergantung kepada fitur yang digunakan.

3. Sistem yang Dibangun

3.1. Konsep Model

Dalam model penelitian yang akan dibangun, terdapat beberapa proses yang akan dilakukan antara lain, dataset akan masuk pada tahap preprocessing untuk menghilangkan informasi yang tidak diperlukan.
Tahapan pengujian dapat dilihat pada diagram alir berikut:



Gambar 1. Alur diagram model

3.2. *Preprocessing*

Tahap preprocessing digunakan untuk membuat data menjadi lebih terstruktur agar data dapat untuk diproses. Terdapat beberapa proses pada tahap preprocessing yang dilakukan dalam penelitian ini diantaranya adalah:

A. Negation handling
   Proses negation handling bertujuan untuk merubah polaritas kata kedalam kalimat yang beda. Contohnya seperti kata "didn't" dirubah menjadi "did not", "wasn't" dirubah menjadi "was not".

B. Cleansing data
   Proses cleansing data bertujuan untuk menghilangkan tanda baca, link, dan simbol pada dataset agar data lebih mudah digunakan pada tahap processing. Contohnya seperti pada kalimat "alur cerita bagus!!" menjadi "alur cerita bagus".

C. Case folding
   Tahap case folding bertujuan untuk merubah huruf uppercase menjadi lowercase dan hanya menerima semua huruf alfabet. Contohnya seperti pada kalimat "Indonesia Movie" menjadi "indonesia movie"

D. Tokenization
   Proses ini bertujuan untuk memisahkan teks menjadi kata yang dianggap sebagai token berdasarkan tiap spasi yang ada dalam kalimat. Contohnya pada kalimat "this film so good" setelah proses tokenisasi akan menjadi "'this' 'film' 'so' 'good'".

E. Stopword removal
   Proses stopword removal bertujuan untuk menghilangkan kata yang tidak penting, kata yang tidak berbobot, atau kata yang bermakna. Seperti kata "from", "with", "to", "and", dan lainnya.

F. Stemming
   Proses stemming digunakan untuk menghilangkan awalan atau akhiran pada kata seperti pada kata "liked" menjadi "like" dan lainnya.

3.3. *Data Split*

Data split digunakan untuk memisahkan antara data latih dan data tes. Pada penelitian ini perbandingan jumlah data latih dengan data tes yaitu 80:20. 80% untuk data latih dan 20% untuk data tes. Jumlah data latih setelah dipisahkan yaitu 1600 data dengan 800 label positif dan 800 label negatif, sedangkan untuk data tes berjumlah 400 data dengan 200 label positif dan 200 negatif.

3.4. Fitur Ekstraksi

Fitur ekstraksi merupakan faktor penting yang dapat mempengaruhi tingkat akurasi pada tahap klasifikasi. Fitur seleksi yang digunakan dalam penelitian ini yaitu TF-IDF dan Lexicon. TF-IDF merupakan fitur ekstraksi yang populer [2], metode ini bekerja dengan cara menghitung bobot setiap kata yang umum digunakan[11]. Untuk menghitung bobot setiap kata dalam dokumen metode ini akan menghitung kemunculan sebuah kata dalam dokumen. Fitur ekstraksi *lexicon* digunakan untuk menghitung polaritas setiap kata dalam kalimat, kemudian hasil dari polaritas setiap kata diubah menjadi bobot yang ditentukan. Hasil dari metode *lexicon* yang

didapat akan dikalikan dengan nilai TF-IDF[7]. Bobot sendiri merupakan nilai yang dapat dihitung, sedangkan polaritas pada sentimen merupakan representasi emosi dalam kata, yang dibagi menjadi 3 bagian yaitu, positif, negatif, dan netral. Kedua fitur ekstraksi ini akan dikombinasikan bertujuan untuk meningkatkan performansi dari hasil klasifikasi.

TF-IDF

Term Frequency — Inverse Document Frequency (TF-IDF) merupakan algoritma yang digunakan untuk menghitung bobot setiap kata yang umum digunakan. TF-IDF sering digunakan untuk memberikan karakteristik pada dokumen[11]. Metode ini akan menghitung seberapa sering suatu kata muncul dalam sebuah dokumen.

Term Frequency (TF) adalah banyaknya i dalam data j, lalu hasilnya dibagi dengan total *term* yang ada dalam data j. Berikut ini merupakan rumus yang digunakan untuk menghitung nilai TF.

$$tf_{ij} = \frac{f_d(i)}{\max_{j \in d} f_d(j)} \text{ [12]}$$

Inversed Document Frequency (IDF) memiliki tujuan untuk mereduksi bobot pada *term* jika keberadaannya terdapat pada seluruh dokumen. Berikut ini merupakan rumus yang digunakan untuk menghitung nilai IDF.

$$idf(t,D) = \log\left(\frac{N}{\text{df}(i)}\right) \text{ [12]}$$

Lexicon SentiWordNet

Metode *lexicon* digunakan untuk menentukan sentimen kata opini. *Lexicon* SentiWordNet digunakan karena menghasilkan akurasi yang cukup bagus pada penelitian[13]. Nilai sentimen akan diberi bobot 1 jika kata merupakan opini positif, -1 jika kata merupakan sentimen negatif, dan jika terdapat kata yang tidak ada pada kamus SentiWordNet, maka akan diberi bobot 0. Berikut tabel contoh penggunaan metode *lexicon* SentiWordNet:

Tabel 1. Contoh Pelabelan Kata *SentiWordNet*

No	Kata	Kategori
1	Good, interesting, nice, well, accept, funny	Positif
2	Bad, Boring, wrong	Negatif
3	So, film, movie	Netral

TF-IDF dan *Lexicon* SentiWordNet

Pada penelitian ini penulis mencoba menggabungkan ekstraksi fitur TF-IDF dengan *lexicon* SentiWordNet dengan cara mengalikan bobot fitur pada setiap data, Berikut merupakan contoh tabel ilustrasi pembobotan TF-IDF dan *lexicon* SentiWordNet setelah dikalikan:

Tabel 2. Contoh Pembobotan Fitur Ekstraksi

Kalimat	Interesting	boring	Film
TF-IDF	0.5	0.5	0.5
SentiWordNet	1	-1	0
TF-IDF X SentiWordNet	0.5	-0.5	0

3.5. Fitur Seleksi

Setelah melewati tahap preprocessing, Selanjutnya adalah, tahapan fitur seleksi menggunakan Information Gain (IG) digunakan untuk mengurangi fitur yang tidak diperlukan dalam tahapan klasifikasi tujuannya adalah memberikan efisiensi pada algoritma tanpa mengurangi tingkat akurasi meskipun fitur berkurang[2].

Information Gain merupakan salah satu fitur seleksi, metode yang digunakan adalah scoring untuk nominal maupun pembobotan atribut kontinu yang didiskretkan menggunakan maksimal entropy. Suatu entropy digunakan untuk mendefinisikan nilai *Information Gain*[14]. Fitur seleksi ini merupakan salah satu yang sering digunakan dalam sebuah sentimen analisis[15]. Berikut ini merupakan rumus yang digunakan untuk menghitung IG.

$$IG(c,t) = S(c) - \sum_{j \in value(t)}^{n} \frac{|c_j|}{|c|} S(c_j) \text{ [16]}$$

S(c) adalah entropi pada semua fitur c sebelum dipisah, S(c_j) adalah entropi fitur c untuk kelas t = j sesudah dipisah, value(t) adalah himpunan dari nilai yang mungkin untuk kelas t, n adalah total nilai yang mungkin untuk kelas t, $|c_j|$ adalah total sampel dengan nilai yang sama dengan j ,$|c|$ adalah total sampel pada total kelas, dan entropi S(c) memiliki rumus berikut ini.

$$S(c) = \sum_{j \in value(t)}^{n} p(c_j) \log p(c_i) \text{ [16]}$$

Nilai $p(c_i)$ adalah kemungkinan munculnya fitur ke-*i,* dan *n* adalah total fitur maksimal[16].

Pada penelitian ini, penggunaan IG memakai data kontinu dari hasil fitur ekstraksi. Bobot nilai setiap fitur atau atribut yang sudah dikalkulasi pada IG tidak akan digunakan dalam tahap klasifikasi, melainkan hanya digunakan untuk seleksi fitur berdasarkan nilai *threshold* yang digunakan. Berikut merupakan tabel ilustrasi bagaimana IG digunakan untuk seleksi atribut:

Tabel 3. Contoh Penggunaan Fitur Seleksi

Good	Positif	Negatif	Total	Entropy	IG
0.40	2	3	5	0.0970950	0.2467
0.0	4	0	4	0	
0.80	3	2	5	0.0970950	
Total	9	5	14	0.0940825	

Contohnya kata ‘Good’ merupakan salah satu atribut yang ada dalam data. Langkah awal yang dilakukan adalah menghitung banyaknya bobot dari nilai TF-IDF yang ada dalam data, dalam tabel terdapat beberapa bobot yaitu 0.40, 0.0, dan 0.80. Lalu, Langkah kedua yaitu menghitung label pada setiap bobot, selanjutnya langkah ketiga yaitu menghitung nilai ‘entropy’ pada setiap bobot TF-IDF dalam kata ‘good’, setelah semua entropy dihitung, selanjutnya menghitung nilai IG-nya, dan terakhir yaitu, menentukan nilai *threshold* jika bobot IG < *threshold,* maka fitur atau atribut akan terseleksi.

3.6. KNN

Proses klasifikasi dilakukan setelah melakukan fitur ekstraksi menggunakan metode *K-Nearest Neighbor* (KNN). *K-Nearest Neighbor* (KNN) merupakan salah satu metode *machine learning* yang dapat digunakan untuk sentimen analisis. *K-Nearest Neighbor* adalah salah satu algoritma supervised learning dimana hasil dari instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori **k**-tetangga terdekat[17]. Berikut ini merupakan rumus yang digunakan untuk menghitung KNN.

$$Dist = \sqrt{\sum_{j \in value(t)}^{n} |Te_i - Tr_i|^2} \quad [6]$$

Nilai *d* merupakan nilai dimensi atau fitur, Te_i merupakan nilai fitur *i* pada data tes dan Tr_i merupakan fitur *i* pada data latih[6].

.

4. Evaluasi

4.1. Tahap Pengujian

Pada penelitian ini, terdapat beberapa proses pengujian yang dilakukan untuk mengetahui dan menganalisis performansi yang didapatkan ketika menggunakan metode ektraksi fitur TF-IDF dan *Lexicon* SentiWordNet, seleksi fitur *Information Gain* (IG), dan metode klasifikasi *K-Nearest Neighbor* (KNN), berikut ini merupakan tabel tahapan pengujian yang akan dilakukan.

Tabel 4. Tahap Pengujian

Pengujian	Tujuan
Membandingkan penggunaan fitur ekstraksi *lexicon* SentiWordNet dengan tanpa menggunakan *lexicon* SentiWordNet.	Mengetahui pengaruh penggunaan fitur ekstraksi *lexicon* SentiWordNet pada data ulasan film berbahasa inggris.
Membandingkan penggunaan fitur seleksi *Information Gain* (IG) dengan tanpa menggunakan *Information Gain* (IG).	Mengetahui pengaruh penggunaan fitur seleksi *Information Gain* (IG) pada data ulasan film berbahasa inggris
Membandingkan nilai nn dalam rentang 1-100 pada KNN.	Mengetahui nilai nn yang paling optimal untuk KNN terhadap dataset ulasan film berbahasa inggris.

4.2. Hasil Pengujian Skenario 1

Pengujian ini dilakukan untuk mengetahui pengaruh penggunaan fitur ekstraksi *lexicon* SentiWordNet untuk data ulasan film berbahasa inggris. Pada skenario ini terdapat 3 proses pengujian yaitu, pertama, menggunakan fitur ekstraksi kombinasi antara TF-IDF dengan *Lexicon* SentiWordNet, kedua, menggunakan fitur ekstraksi TF-IDF saja dan ketiga menggunakan fitur ekstraksi *Lexicon* SentiWordNet saja. Dalam tahap ini pengujian dilakukan dengan menggunakan metode split *Stratified shuffle split* agar mendapatkan pembagian data dengan label positif dan negatif yang seimbang, lalu menggunakan seleksi fitur *Information Gain* dengan *threshold* = 0.06. Penggunaan fitur ektraksi *Lexicon* SentiWordNet pada kombinasi fitur ekstraksi TF-IDF dan *Lexicon* bobot 0 tidak digunakan, karena terdapat banyak polaritas yang memiliki nilai netral atau kata yang tidak ada dalam kamus sehingga membuat hasil kombinasi ekstraksi fitur menjadi banyak yang memiliki bobot 0 ketika digabungkan dan hasil yang didapatkan ketika memiliki F1 *score* 62.74%.

Tabel 5. Hasil Perbandingan Akurasi Fitur Ekstraksi

Fitur Ekstraksi	Akurasi	F1	Precision	Recall
TF-IDF X *Lexicon* SentiWordNet	73.31	73.31	67.54	92.5
TF-IDF	81.04	81.04	80.90	87.5
Lexicon SentiWordNet	63.84	63.84	64.7	62.5

Hasil pengujian yang telah dilakukan dapat dilihat pada tabel 5, dapat dilihat bahwa fitur ekstraksi TF-IDF memiliki nilai evaluasi yang paling baik dalam setiap tahap pengujian dengan nilai akurasi sebesar 81.04%, F1 score 81.04%, precision 80.90%, dan recall 87.5%, sedangkan nilai terendah, selanjutnya pada tahap menggunakan fitur ekstraksi kombinasi TF-IDF dengan *Lexicon* memberikan hasil yang cukup yaitu, dengan akurasi 73.31%, F1 score 73.31%, precision 67.54%, recall 92.5%. Pada fitur ekstraksi menggunakan *Lexicon SentiWordNet* memberikan hasil yang kurang baik yaitu akurasi 63.84%, F1 score 63.84%, precision 64.7%, dan recall 62.5% . Dari hasil pengujian pada skenario ini, dapat disimpulkan bahwa penggunaan *lexicon* terhadap fitur ekstraksi tidak memberikan hasil yang baik, dikarenakan skenario sistem yang dibangun pada *paper* rujukan melakukan *split data* pada saat sebelum melakukan klasifikasi, oleh karena itu, mesin telah mengetahui isi dari data test terlebih dahulu sehingga memberikan hasil yang baik.

4.3. Hasil Pengujian Skenario 2

Pengujian ini dilakukan untuk mengetahui pengaruh penggunaan fitur seleksi *Information Gain* (IG) pada dataset ulasan film berbahasa inggris. Pengujian ini dilakukan dengan menggunakan *threshold* untuk *Information Gain* (IG) dalam rentang 0 sampai 0.08, fitur ekstraksi menggunakan TF-IDF dan menggunakan nn = 85 untuk klasifikasi KNN.

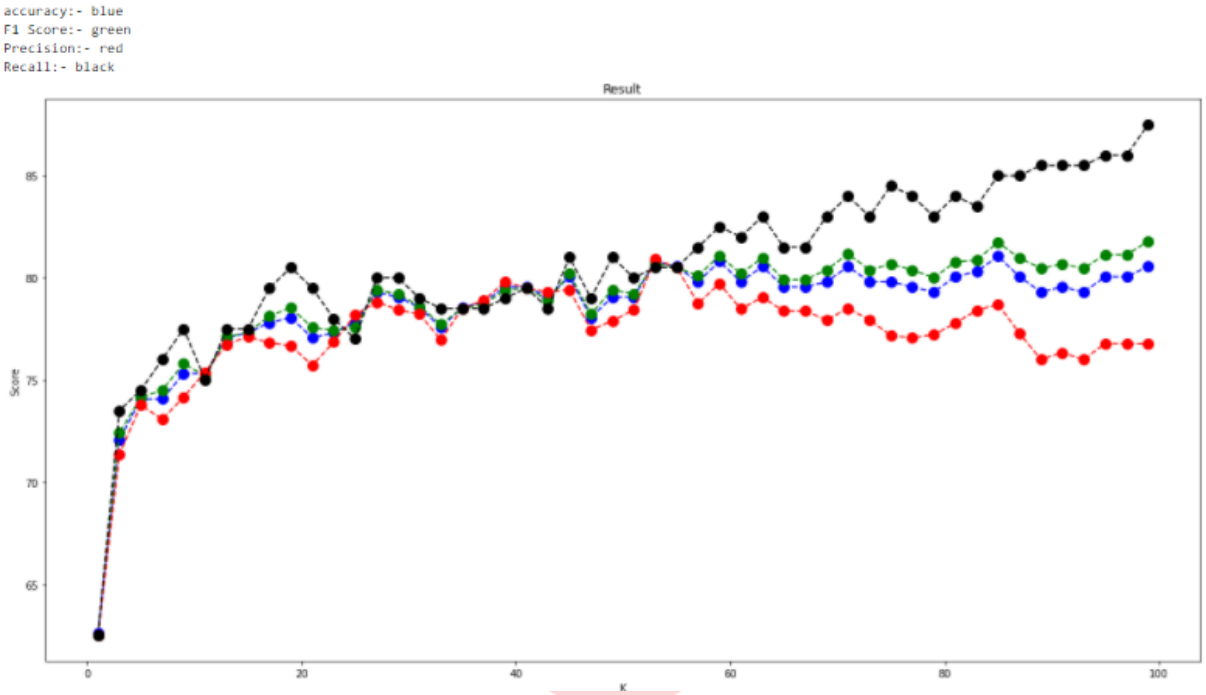
Tabel 6. Hasil Perbandingan Evaluasi Menggunakan Fitur Seleksi *Information Gain* (IG)

Threshold	Jumlah fitur sebelum	Jumlah fitur sesudah	Akurasi	F1	Precision	Recall
Tanpa Threshold	1656	1656	77.80	77.80	72.46	89.5
0.04	1656	1379	73.56	73.56	66.54	94.5
0.05	1656	1203	79.80	79.80	74.89	89.5
0.06	1656	955	**81.04**	**81.04**	**78.70**	85
0.07	1656	804	72.81	72.81	66.19	93
0.08	1656	693	62.59	62.59	57.26	**98.5**

Dari hasil percobaan yang telah dilakukan dapat diketahui bahwa penggunaan fitur seleksi *Information Gain* (IG) dapat mengoptimalkan akurasi terhadap klasifikasi untuk dataset ulasan film berbahasa inggris. Nilai optimum yang didapatkan yaitu pada *threshold* 0.06 dengan nilai akurasi 81.04%, F1 score 81.04%, precision 78.70%, namun nilai recall terbaik terdapat pada *threshold* 0.08 yaitu 98.5%. Pada saat threshold 0.08 memiliki nilai *precision* yang lebih rendah dibandingkan dengan nilai *recall*, nilai *precision* rendah diakibatkan oleh banyaknya label yang *false positive*, artinya sistem banyak memprediksi label yang seharusnya berlabel positif namun salah prediksi, sedangkan nilai *recall* tinggi diakibatkan oleh nilai *false negative* yang sedikit, artinya sistem sedikit memprediksi nilai negatif yang salah. Hal tersebut dapat terjadi karena, semakin tinggi nilai *threshold* maka semakin banyak fitur yang terseleksi termasuk dengan fitur yang bermanfaat sehingga berpengaruh pada saat tahap klasifikasi. Dari hasil pengujian pada skenario ini, fitur seleksi IG dapat membantu mengoptimalkan hasil evaluasi karena menghilangkan fitur yang tidak digunakan. Namun, diperlukan untuk mencari nilai *threshold* yang tepat terhadap data yang digunakan agar dapat menghilangkan fitur yang tidak diperlukan tetapi tidak juga menghilangkan fitur yang berguna.

4.4. Hasil Pengujian Skenario 3

Pengujian ini dilakukan untuk mengetahui nilai *Nearest neighbor* optimum untuk metode klasifikasi dengan KNN dengan menggunakan fitur ekstraksi *lexicon* SentiWordNet dan fitur seleksi *Information Gain* (IG) pada data ulasan film berbahasa inggris. Dalam tahap ini pengujian dilakukan dengan menggunakan metode split *Stratified shuffle split* agar mendapatkan pembagian data dengan label positif dan negatif seimbang, lalu menggunakan seleksi fitur *Information Gain* dengan *threshold* = 0.06.

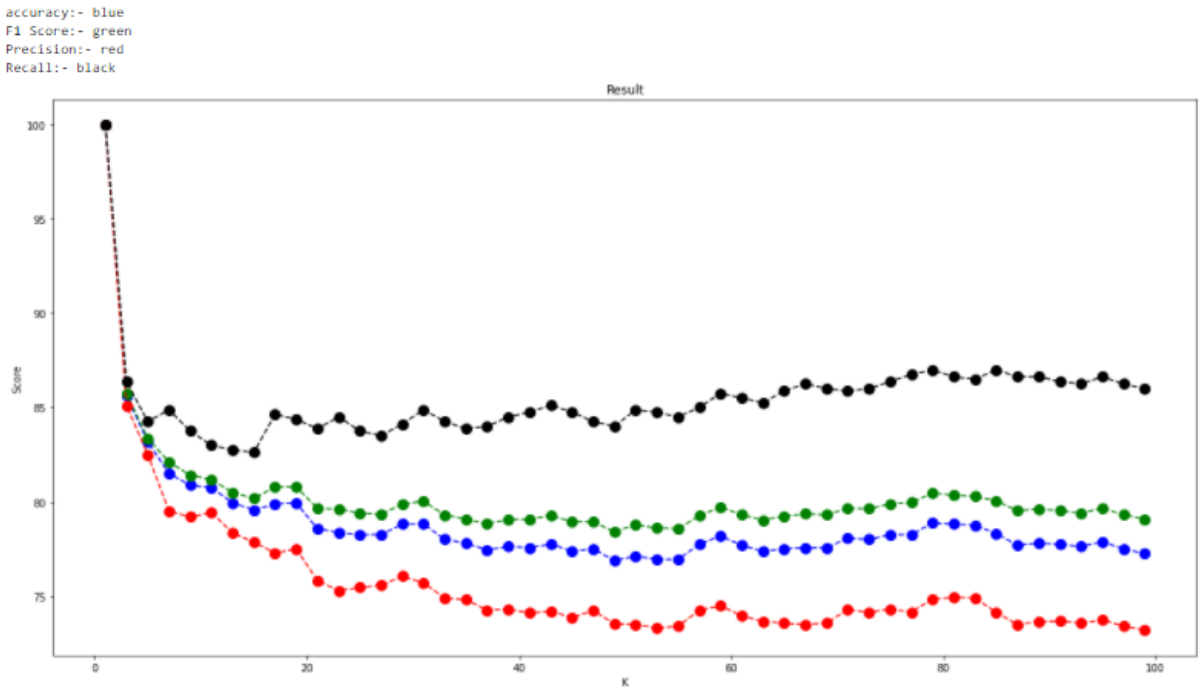


Gambar 2. Hasil Penggunaan dengan Fitur Ekstraksi TF-IDF dan Fitur Seleksi IG Dengan Data Tes

Tabel 7. Hasil Penggunaan dengan Fitur Ekstraksi TF-IDF dan Fitur Seleksi IG Dengan Data Tes

Evaluasi	**Hasil**	***Nearest neighbor***
Akurasi	81.04	85
F1 Score	81.04	85
Precision	80.90	53
Recall	87.5	99

Dalam tahap pengujian ini menggunakan kombinasi fitur ekstraksi TF-IDF dan fitur seleksi *Information Gain* (IG) pada metode klasifikasi KNN dapat dilihat nilai akurasi tertinggi yang diperoleh yaitu 82.29% pada nn = 39, lalu untuk nilai F1 score tertinggi yaitu 82.55% pada nn = 39, kemudian untuk nilai precision tertinggi yaitu 81.21% pada nn = 35, dan recall tertinggi yaitu 84.5% pada nn = 93.



Gambar 3. Hasil Penggunaan dengan Fitur Ekstraksi TF-IDF dan Fitur Seleksi IG Dengan Data Latih

Tabel 8. Hasil Penggunaan dengan Fitur Ekstraksi TF-IDF dan Fitur Seleksi IG Dengan Data Latih

Evaluasi	Hasil	*Nearest neighbor*
Akurasi	85.62	3
F1 Score	85.62	3
Precision	85.09	3
Recall	86.37	3

Selanjutnya pada pengujian ini dilakukan juga percobaan menggunakan data latih, hasil yang didapatkan yaitu nilai akurasi 85.62%, F1 *score* 85.62%, *precision* 85.09%, dan *recall* 86.37% pada *nn* = 3, hal tersebut terjadi karena masih terdapat fitur yang mengurangi hasil klasifikasi atau terdapat kata yang berpengaruh tapi terseleksi.

5. Kesimpulan

Dari hasil pengujian yang telah dilakukan terhadap dataset ulasan film berbahasa inggris mendapat kesimpulan bahwa untuk mengoptimasi performansi penggunaan metode klasifikasi KNN diperlukan untuk menemukan *nearest neighbor* yang tepat, pada pengujian ini didapatkan nn = 85. Pada pengujian penggabungan ekstraksi fitur TF-IDF dengan *Lexicon* SentiWordnet didapatkan bahwa, nilai akurasi penggabungan kedua fitur seleksi tersebut tidak lebih bagus dibandingkan dengan hanya , menggunakan fitur ekstraksi TF-IDF yaitu 73.31%, sedangkan dengan TF-IDF saja mendapatkan 81.04%. Terakhir, dalam pengujian penggunaan fitur seleksi *Information Gain* (IG) diketahui jika, penggunaan IG mampu mengoptimasi hasil performansi pada metode klasifikasi KNN, tetapi diperlukan untuk mencari nilai *threshold* yang tepat terhadap dataset yang digunakan agar hasil optimal, akurasi dengan *threshold* = 0.06 mendapatkan nilai optimal yaitu 81.04%, dibandingkan jika tidak menggunakan IG yaitu 77.80%.

Adapun saran untuk pengembangan selanjutanya, yaitu untuk mencoba menggabungkan metode fitur ekstraksi TF-IDF dengan metode *lexicon* selain SentiWordnet, kemudian menerapkan K-*fold cross validation* pada tahapan split data supaya mengetahui performansi suatu model dengan percobaan sebanyak *k* kali dan untuk memvalidasi keakuratan suatu sistem yang dibangun.

REFERENSI

[1] V. Uma Ramya and K. Thirupathi Rao, "Sentiment analysis of movie review using machine learning techniques," *Int. J. Eng. Technol.*, vol. 7, no. 16, pp. 676–681, 2018, doi: 10.14419/ijet.v7i2.7.10921.

[2] K. Kumar, B. S. Harish, and H. K. Darshan, "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 5, p. 109, 2019, doi: 10.9781/ijimai.2018.12.005.

[3] O. Kolchyna, P. C. Treleaven, and T. Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination."

[4] B. Liu, *Sentiment Sentiment Analysis Analysis and and Opinion Opinion Mining Mining*. .

[5] M. Rezwanul, A. Ali, and A. Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, pp. 19–25, 2017, doi: 10.14569/ijacsa.2017.080603.

[6] 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, doi: 10.34818/JDSA.2020.3.22.

[7] W. P. Ali, Y. Sibaroni, and S. Si, "Analisis Sentimen Masyarakat Terhadap Kinerja Presiden Indonesia Dalam Aspek Ekonomi , Kesehatan , dan Pembangunan Berdasarkan Opini dari Twitter," *e-Proceeding Eng.*, vol. 6, no. 2, pp. 8637–8649, 2019.

[8] E. Cambria, A. Valdivia, M. V. Luzón, and F. Herrera, "AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS Sentiment Analysis in TripAdvisor," 2017, [Online]. Available: www.computer.org/intelligent.

[9] S. Zhang *et al.*, "IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 1 Efficient kNN Classification With Different Numbers of Nearest Neighbors," *Ieee Trans. Neural Networks Learn. Syst.*, pp. 1–12, 2017, [Online]. Available: http://ieeexplore.ieee.org.

[10] A. Khan, B. Baharudin, and K. Khan, "Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews," *Trends Appl. Sci. Res.*, vol. 6, no. 10, pp. 1141–1157, 2011, doi: 10.3923/tasr.2011.1141.1157.

[11] K. Sugiyama, "Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages," pp. 198–207.

[12] M. Yunus, "TF-IDF (Term Frequency-Inverse Document Frequency) : Representasi Vector Data Text," *medium*, 2020. https://medium.com/@yunusmuhammad007/tf-idf-term-frequency-inverse-document-frequency-representasi-vector-data-text-2a4eff56cda (accessed Nov. 18, 2020).

[13] A. M. Ismail, "Cara Kerja Algoritma k-Nearest Neighbor (k-NN)," *medium*, 2018. https://medium.com/bee-solution-partners/cara-kerja-algoritma-k-nearest-neighbor-k-nn-389297de543e (accessed Nov. 16, 2020).

[14] I. Maulida, A. Suyatno, and H. R. Hatta, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," vol. 17, no. 2, pp. 249–258, 2016.

[15] J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization."

[16] M. Yunus, "Feature Selection menggunakan Information Gain," *medium*, 2020. https://medium.com/@yunusmuhammad007/feature-selection-menggunakan-information-gain-ba94ca66f658 (accessed Nov. 18, 2020).

[17] A. M. Ismail, "Cara Kerja Algoritma k-Nearest Neighbor (k-NN)," *medium*, 2018. https://medium.com/bee-solution-partners/cara-kerja-algoritma-k-nearest-neighbor-k-nn-389297de543e (accessed Nov. 14, 2020).