Analisis Sentimen terhadap Ulasan Film Berbahasa Inggris Menggunakan Metode *Support Vector Machine* dengan Feature Selection *Information Gain*

Nauffan Muti Hibattullah[1], Said Al Faraby[2], Mahendra Dwifebri Purbolaksono[3]

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

Abstrak

Analisis sentimen adalah suatu bidang yang menganalisis opini, sikap, dan emosi dari banyak orang terhadap suatu produk, jasa atau entitas lain, bidang penelitian ini cukup popular. Teks ulasan merupakan sebuah teks yang isinya berupa ulasan, *review* atau penilaian pada suatu karya seperti film, drama, dan juga buku. Teks ulasan memiliki fungsi yaitu untuk menilai, menimbang, serta mengajukan kritik pada karya maupun peristiwa yang diulas. Bahasa Inggris adalah bahasa yang paling banyak digunakan di dunia. Ada 400 juta penutur asli dan 2 milyar orang mempelajarinya sebagai bahasa kedua. *Support Vector Machine* (SVM) merupakan salah satu metode dalam *supervised learning* yang biasanya digunakan untuk klasifikasi (*Support Vector Classification*) dan regresi (*Support Vector Regression*). Metode *Information Gain* adalah metode yang menggunakan teknik *scoring* untuk pembobotan sebuah fitur dengan menggunakan maksimal *entropy*. Fitur yang dipilih adalah fitur dengan nilai *Information Gain* yang lebih besar atau sama dengan nilai *threshold* tertentu. Penelitian ini menunjukan proses dengan dilakukan kombinasi *Stopword* dan *Stemming* maka akurasi yang dihasilkan akan lebih maksimal, karena proses *preprocessing* lebih lengkap sebesar 86,12%. Selain itu didapatkan bahwa seleksi fitur *Information Gain* (IG) pada penelitian ini membuat akurasi menjadi rendah, tetapi dapat menjadi sebuah solusi yang cukup baik untuk mengatasi masalah *overfitting* pada pengujian analisis sentimen ini. Dan klasifikasi analisis sentimen *movie review* sangat tepat menggunakan algoritma *Support Vector Machine* (SVM) kernel *Liniear*. Dikarenakan *Linear* berfokus dengan fitur-fitur yang mengandung nilai biner yang didapatkan dari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas.

Kata kunci: Analisis Sentimen, Ulasan, Film, Berbahasa Inggris, *Support Vector Machine, Information Gain*

Abstract

Sentiment analysis is a field that analyzes the opinions, attitudes, and emotions of many people towards a product, service or other entity, this field of research is quite popular. Review text is a text whose contents are in the form of reviews, reviews or ratings on a work such as films, dramas, and books. The review text has the function of assessing, weighing, and criticizing the works or events being reviewed. English is the most widely spoken language in the world. There are 400 million native speakers and 2 billion people learn it as a second language. Support Vector Machine (SVM) is one of the methods in supervised learning that's usually used for classification (such as Support Vector Classification) and regression (Support Vector Regression). The Information Gain method is a method that uses a scoring technique to weight a feature by using maximum entropy. The selected feature is a feature with an Information Gain value that is greater than or equal to a certain threshold value. This study shows the process by using a combination of Stopword and Stemming, the resulting accuracy will be maximized, because the preprocessing process is more complete at 86.12%. In addition, it was found that the Information Gain (IG) feature selection in this study made the accuracy low, but it could be a good enough solution to overcome the overfitting problem in this sentiment analysis test. And the classification of movie review sentiment analysis is very precise using the Linear kernel Support Vector Machine (SVM) algorithm. Because Linear focuses on features that contain binary values obtained from the best hyperplane by maximizing the distance between classes.

Keywords: Sentiment Analysis, Reviews, Film, English Language, Support Vector Machine, Information Gain

1. Pendahuluan

Latar Belakang

Perkembangan teknologi informasi saat ini sudah semakin pesat, ditandai dengan luar biasa jumlah kontribusi pengguna di internet yang mengungkapkan opini tentang segala macam subjek, masalah, acara dan produk. *Blog* merupakan saluran yang biasa digunakan untuk mengekspresikan opini salah satunya tentang film. Salah satu

ulasan tersebut adalah ulasan film yang mempengaruhi semua orang [1]. Film merupakan seni visual yang terus berkembang biak dari tahun ke tahun. Melalui ulasan film, penonton bisa mencari tahu film mana yang berkualitas baik. Semakin banyak film yang diproduksi akan membuat banyak ulasan yang akan dihasilkan. Pendapat yang diungkapkan dalam ulasan film sangat memberikan kesan cerminan sebenarnya dari emosi yang sedang disampaikan. Itu adanya penggunaan kata-kata sentimen yang begitu hebat untuk mengekspresikan ulasan [2].

Sehingga diperlukan banyak tenaga bagi penonton untuk membaca banyak ulasan film, sehingga mereka dapat memperoleh informasi tentang film [3]. Berdasarkan pengujian yang telah dilakukan oleh Fitri Eka Cahyanti, memberikan kesimpulan bahwa analisis sentimen dengan kombinasi ekstraksi ciri TF-IDF dan LDA dan SVM sebagai metode klasifikasi mendapatkan hasil kinerja yang sangat baik [4]. Dalam penerapan LDA, dokumen direpresentasikan sebagai campuran acak untuk setiap topik yang dihasilkan, sedangkan topik itu sendiri didapatkan dari olahan kata-kata. LDA menjadi probabilistik pemodelan teks yang paling sering digunakan [5]. Pada studi ini, klasifikasi dokumen berbasis LDA hasilnya menunjukkan akurasi terbaik sekitar 60% sebagai akurasi rata-rata semua lipatan dan akurasi terbaik sekitar 80% kali lipat 6 dan 7 [6]. Hasil penelitian kombinasi metode *Information Gain* dan *Naïve Bayes Classifier* berhasil mendapatkan tingkat akurasi 82,19%. Netral menjadi kelas dengan presisi terendah yaitu 63,76%. Penelitian ini masih memiliki kelemahan yaitu memiliki beberapa salah klasifikasi. Ketepatan kelas pada data netral menjadi masalah utama [29].

Jadi berdasarkan kondisi tersebut, penulis tertarik untuk melakukan analisis sentimen dalam ulasan film dengan menggunakan Metode *Support Vector Machine* (SVM) dengan *Feature Selecetion Information Gain*. Metode *Support Vector Machine* (SVM) dengan pendekatan *machine learning* adalah algoritma populer yang berkinerja baik [7]. Analisis sentimen juga dikenal sebagai penambangan opini [8]. Dari studi ini, kami menemukan bahwa SVM mengungguli CRF. Pekerjaan ini akan membantu penggalian perasaan dan komentar pengguna tentang film baru dan pada gilirannya memeringkat film berdasarkan ulasan ini [9]. *Information Gain* adalah salah satu dari pemilihan fitur terbaik [10]. Metode fitur seleksi ini dijadikan untuk membantu meningkatkan performansi dari algoritma SVM yang digunakan dalam sentiment analisis ulasan film berbahasa inggris.

Topik dan Batasan

Pada penelitian ini penulis akan membangun model untuk analisis sentimen mengenai *movie review* Penelitian ini akan berfokus pada proses *prepocessing*, seleksi fitur, dan klasifikasi. Pada proses *prepocessing* penulis membandingkan sistem tanpa menggunakan proses *stemming* dan tanpa menggunakan proses *stopwords.* Pada proses seleksi fitur, penulis akan membandingkan penggunaan seleksi fitur *Information Gain* (IG) dengan yang tidak menggunakan seleksi fitur *Information Gain* (IG). Kemudian pada proses klasifikasi menggunakan *Support Vector Machine* (SVM). Pada penelitian ini terdapat beberapa batasan masalah, yaitu dataset *movie review* yang digunakan bersumber dari IMDB dengan jumlah 6000 *review* dan hanya berfokus pada *review* berbahasa Inggris. Dataset dilabelkan secara manual kedalam dua kelas, yaitu positif dan negatif. Proses klasifikasi menggunakan *kernel linear, RBF* dan *Polynomial.*

Tujuan

Penelitian ini dilakukan dengan tujuan untuk membandingkan pengaruh sistem pengklasifikasian penggunaan metode *Support Vector Machine* (SVM) dan fitur seleksi *Information Gain* dengan tanpa penggunakan seleksi fitur *Information Gain* (IG) dalam pemilihan fitur. Serta menganalisis pengaruh penggunaan proses *prepocessing* sistem tanpa menggunakan proses *Stemming* dan tanpa menggunakan proses S*topwords* terhadap sistem. Dan membandingkan pengaruh dari sistem menggunakan fungsi klasifikasi menggunakan *kernel linear, RBF* dan *Polynomial.*

Organisasi Tulisan

Bagian selanjutnya pada penelitian ini adalah bagian 2 yang membahas studi terkait dengan penelitian yang telah dilakukan, bagian 3 membahas rancangan sistem yang dibangun, bagian 4 membahas evaluasi dari hasil pengujian, dan bagian 5 membahas kesimpulan dari penelitian ini dan saran untuk penelitian selanjutnya.

2. Studi Terkait

Penelitian yang dilakukan oleh Novelty Octaviani Faomasi Daeli [3] pada tahun 2020 dengan judul *Sentiment analysis on movie reviews using Information gain and K-nearest neighbor* bahwa pada penelitian ini, Polaritas v2.0 dari *dataset review film Cornell* akan dipakai sebagai penguji metode KNN dan pemilihan fitur *Information Gain* untuk mendapatkan kinerja yang lebih baik. Tujuan penelitian ini yaitu menemukan K yang optimal pada KNN berdasarkan ambang *Information Gain* (IG), dan mendapatkan ambang *Information Gain* (IG) terbaik.

Penelitian yang dilakukan oleh Asriyanti Indah Pratiwi dan Adiwijaya [10] pada tahun 2018 dengan judul *On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis* bahwa pada penelitian ini, metode yang digunakan mengurangi lebih dari 90% fitur yang tidak digunakan sedangkan skema

klasifikasi yang digunakan mencapai 96% akurasi klasifikasi sentimen. Dari hasil pengujian, dapat disimpulkan bahwa kombinasi seleksi fitur yang digunakan dan klasifikasi mencapai kinerja terbaik selama ini.

Penelitian yang dilakukan oleh Zheng, Wenying danYe, Qiang [11] pada tahun 2009 dengan judul *Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm* bahwa pada pekerjaan ini melakukan eksplorasi analisis sentimen terhadap ulasan wisatawan China dengan algoritma *Support Vector Machine* (SVM). Hasil pengujian menunjukkan bahwa, dibandingkan dengan penelitian sebelumnya tentang ulasan bahasa Inggris, algoritma SVM dapat menghasilkan performa klasifikasi sentimen yang sangat baik untuk ulasan wisatawan dalam bahasa China.

Penelitian yang dilakukan oleh Hanif Salaf [12] pada tahun 2019 dengan judul Analisis Pengaruh Seleksi Fitur *Information Gain* dan *Mutual Information* pada Klasifikasi Sentimen Ulasan Film Menggunakan *Support Vector Machine* bahwa dengan menggunakan *Information Gain* sebagai seleksi fitur pada algoritma klasifikasi SVM memiliki hasil yang sama dengan *Mutual Information*. Tujuan analisis sentimen yaitu untuk menentukan komentar positif atau negatif dalam suatu kalimat atau dokumen. Masing-masing dari kedua metode pemilihan fitur tersebut dijadikan sebagai seleksi fitur untuk membantu meningkatkan performansi algoritma klasifikasi *Support Vector Machine* (SVM). Kemudian ketika kedua seleksi fitur tersebut dibandingkan, *Information Gain* memiliki hasil yang sama dengan *Mutual Information* dengan nilai akurasi tertinggi sebesar 89.05%.

Penelitian yang dilakukan oleh Fitri Eka Cahyanti, Adiwijaya, dan Said Al Faraby [4] pada tahun 2019 dengan judul *On The Feature Extraction For Sentiment Analysis of Movie Reviews Based on* SVM bahwa Penelitian ini menghasilkan performansi terbaik pada kombinasi TF-IDF dan LDA, menggunakan 240 topik memiliki 29792 fitur yaitu 82,16%. Menonton film salah satu aktivitas untuk mengurangi rasa bosan, maka perlu dicarilah informasi mengenai film yang dikemas dengan bentuk ulasan film untuk mengetahui apakah film tersebut layak untuk ditonton atau tidak. Oleh sebab itu, diperlukan analisis sentimen untuk mengklasifikasikan ulasan film menjadi sentimen positif dan negatif.

2.1. Analisis Sentimen

Analisis sentimen merupakan sebuah bidang penelitian yang dapat mengelola bahasa natural, komputasi *linguistic*, dan *text mining*. Analisis sentiment atau yang sering disebut dengan *opinion mining* merupakan studi komputasional dari opini yang diberikan orang lain yang terdapat dalam entitas, *event*, dan atribut yang dimiliki. Tujuan dari analisis sentiment ini adalah untuk mengelompokkan polaritas pada suatu teks apakah pendapat yang dikemukakan bersifat positif, negatif, atau netral [13]. Sebuah sentiment dapat ditemukan pada *document level*, *sentence level*, maupun *entity level*. *Document level sentiment analysis* bertujuan untuk mengetahui apakah suatu dokumen itu menunjukkan sentiment positif ataupun negatif. *Sentence and phase-level sentiment analysis* bertujuan untuk mengetahui apakah suatu kalimat itu menunjukkan sentiment positif ataupun negatif. *Enty and aspect- level opinions* bertujuan untuk mengetahui apakah suatu pendapat atau opini itu menunjukkan sentimen positif ataupun negatif [13].

2.2. Ulasan Film Berbahasa Inggris

Film merupakan media komunikasi yang pernah digunakan sejak perang dunia pertama untuk untuk menyampaikan informasi, opini, dan juga hiburan [14]. Teks ulasan adalah sebuah teks yang berisi tinjauan, penilaian atau ulasan terhadap suatu seni seperti film, drama, atau sebuah buku untuk mengetahui kelebihan dan kekurangan dari suatu karya tersebut. Teks ulasan dapat dikatakan dengan resensi. Ketika mengolah suatu karya, pengolah harus berprilaku kritis supaya hasil ulasanya bisa menghasilkan kontribusi bagi perkembangan karya tersebut [15]. Ada beberapa pekerjaan sebelumnya yang menerapkan analisis sentimen bahasa dengan sumber bahasa relevan yang langka. Pengecualian penting adalah pekerjaan untuk menghasilkan sumber daya analisis subjektivitas lingual dari data bahasa Inggris [16].

2.3. Metode *Support Vector Machine*

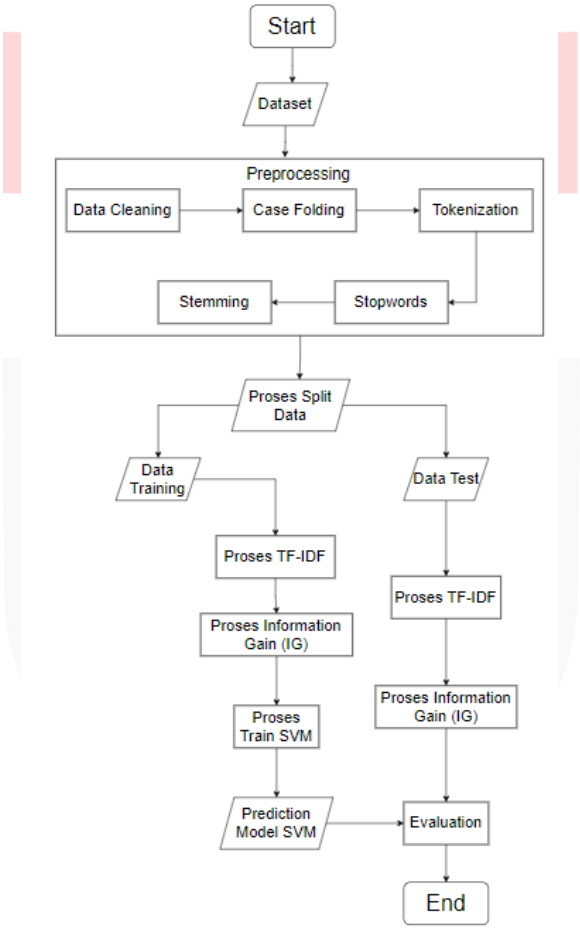
Support Vector Machine (SVM) merupakan motode algoritma pembelajaran mesin yang baru. SVM memiliki klasifikasi yang kuat dan kemampuan generalisasi, diterapkan secara luas. SVM banyak digunakan untuk melakukan klasifikasi otomatis. Banyak penelitian menerapkan SVM sebagai metodenya untuk membuktikan bahwa metode ini merupakan metode yang efesien, diantaranya yaitu pengenalan citra, analisis medik, dan melakukan prediksi [17]. SVM memiliki beberapa karakteristik yaitu hasil *support vector* yang diperoleh disimpan untuk digunakan kembali pada proses testing, model yang dihasilkan SVM selalu sama pada tiap *testing* dengan *margin* yang maksimal, dapat memisahkan distribusi kelas *linear* atau *non linear*, SVM tidak memerlukan reduksi dimensi, dan banyak data memengaruhi memori yang digunakan dalam SVM. SVM memiliki beberapa kelebihan diantaranya yaitu pertama SVM mampu untuk melakukan generalisasi baik data yang digunakan ataupun yang tidak. Kedua *Curse or Demensionality* dimana jumlah sampel data lebih minim dibandingkan dengan dimensi ruang vector data. Ketiga landasan teori SVM dapat dianalisis dengan jelas karena tidak bersifat *blackbox* [18].

2.4. Fitur Seleksi *Information Gain*

Fitur Seleksi adalah proses menghapus fitur berlebihan dan tidak relevan dari dataset yang digunakan. Sehingga waktu yang digunakan untuk pengklasifikasi, dan dapat meningkatkan akurasi juga karena fitur yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif [19]. Dengan fitur seleksi dapat meningkatkan pengetahuan dan biaya penanganan data menjadi lebih kecil [19]. *Information Gain* (IG) merupakan ukuran efektifitas suatu atribut dalam mengklasifikasi data. *Information Gain* (IG) digunakan untuk menentukan urutan atribut dimana atribut yang memilki nilai IG terbesar yang akan dipilih. *Information Gain* (IG) dengan pendekatan *machine learning* digunakan untuk memilih suatu fitur yang memiliki relevansi yang baik dengan fitur IG yang buruk akan dihilangkan [11]. *Information Gain* (IG) mengukur tingkat relevansi dari setiap kelas. Fitur yang baik adalah fitur yang memiliki relevansi tinggi dengan kelas tertentu [10].

3. Sistem yang Dibangun

Bagian ini, dibangun sebuah sistem yang dapat melakukan klasifikasi sentimen pada dataset ulasan film yang menjelaskan tentang *dataset, preprocessing, feature selection, classification*, dan *evaluation*. Rancangan sistem yang dibangun dapat dilihat pada Gambar1 [1].



Gambar 1 Gambaran Sistem yang dibangun

3.1. *Dataset*

Pada penelitian yang ditulis oleh Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts berjudul *Learning Word Vectors for Sentiment Analysis* yaitu kaggle.com *website* yang berasal dari India yang berisi kumpulan-kumpulan *dataset* yang dirancang oleh Lakshmipathi N., seorang *engineer* berasal dari India. *Dataset* tersebut telah diberi label dengan prediksi jumlah ulasan positif dan negatif menggunakan algoritma klasifikasi atau pembelajaran mendalam. Kemudian Ada data lain yang tidak berlabel untuk diproses juga dan teks mentah lalu diproses oleh *format bag of words* [28].

Pada tahap awal yang harus dilakukan berdasarkan rancangan sistem pada Gambar 1 adalah *dataset*. *Dataset* yang digunakan dalam penelitian ini berjumlah 6000 ulasan film berbahasa inggris dari IMDB yang diunduh dari *website* kaggle.com. *Dataset* ini terdapat 3000 ulasan film positif (1) dan 3000 ulasan film negatif (-1). *Dataset* tersebut nantinya pada proses *split data* akan dipisah menjadi 2 data yaitu data latih dengan distribusi 4500 ulasan

film (positif dan negatif) dan data tes dengan 1500 ulasan film (positif dan negatif). Berikut pada Tabel 1 merupakan contoh ulasan film dari dataset penelitian ini :

Tabel 1 Contoh dari dataset penelitian ini

1	*From this Dilan film, we know that to be box office and number one, sometimes you don't need excessive imagination, complex stories, and sophisticated editing technology. It's enough to tell the story of 2 human children that touch everyone's daily life, with beautiful puns combined with a simple storyline. A fantastic blend of Indonesian literature and film!*
-1	*This country, and the way it sees the world, we have a fear of outsiders. No one really wants to see their own faults, their guilt, the devil within them*

3.2. *Preprocessing*

Untuk memilih sebuah kata agar dapat diolah dengan sebuah algoritma maka diperlukan proses *prepocessing*. Pada tahap ini, *dataset* dalam bentuk teks akan melalui tahapan *preprocessing* yaitu sebagai berikut:

1. *Data Cleaning*
   Dalam tahap ini menghapus simbol-simbol yang ada pada dokumen dengan tujuan membersihkan dokumen.
2. *Case Folding*
   Pada tahap ini *dataset* diolah menjadi huruf kecil atau *lowercase*.
3. *Tokenization*
   Tahap ini memisahkan kata-kata dalam kumpulan data menjadi token [4].
4. *Stopwords*
   Pada tahap ini beberapa kata yang dianggap tidak diperlukan akan dihilangkan dalam teks [4].
   Penelitian ini menggunakan *stopwords* bahasa inggris dari NLTK. Peneliti menghapus kata-kata yang diaggap tidak penting pada *list* NLTK seperti 'doesn', ‘doesn't’, 'hadn', ‘hadn't’, 'hasn', ‘hasn't’, 'needn', ‘needn't’, ‘do’, ‘does’, 'shan', dan ‘shan't’. Dan kemudian sistem pada tahap ini nantinya akan menghapus kata – kata yang terkandung didalam NLTK seperti ‘i’, ‘me’, ‘my’, ‘myself’, ‘we’, ‘our’, ‘ours’, ‘him’, ‘his’, ‘himself’, ‘being’, ‘have’, ‘has’, ‘having’, 'if', 'or', 'because', 'as', ‘because', 'as', dan 'until'.
5. *Stemming*
   Proses ini mengurangi infleksi dalam kata-kata (misalnya bermasalah, kesulitan) ke bentuk dasarnya (misalnya masalah). "Akar" dalam hal ini mungkin bukan akar kata yang sebenarnya, tetapi hanya bentuk kanonik dari kata aslinya.

3.3. Data Split

Pada tahap ini *dataset* yang telah dilakukan *Preprocessing* (*Data cleaning, Case Folding, Tokenization, Stopwords, dan Stemming*) *Dataset* yang digunakan dalam penelitian ini berjumlah 6000 ulasan dengan jumlah data yang berlabel positif 3000 (50%) dan negatif 3000 (50%), kemudian dipecah menjadi 2 bagian dengan karateristik ukuran data sama dengan 0.25 dan *random state* sama dengan 42, yaitu data latih dengan distribusi 4500 (positif (1) sama dengan 2258 dan negatif (-1) sama dengan 2242) dan data tes dengan 1500 (positif (1) sama dengan 742 dan negatif (-1) sama dengan 758). Data uji akan disimpan dahulu untuk digunakan pada proses evaluasi. Kemudian Data latih akan diproses pada selanjutnya.

3.4. Ekstraksi Fitur N-Gram

Proses *feature extraction* sangat diperlukan dalam klasifikasi. Tujuan dari *feature extraction* adalah untuk melakukan pembobotan kata dengan menghitung jumlah bobot setiap kata dan berapa kali kata tersebut muncul dalam suatu kalimat [20].

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

Keterangan :

TF * IDF(d,t) : Pembobotan TF-IDF.

TF(d,t) : Frekuensi munculnya term t pada dokumen d.

N : Jumlah dari semua kumpulan dokumen.

df(t) : Jumlah dari dokumen yang mengandung term t.

Pada TF-IDF terdapat salah satu parameter yaitu *N-gram*. *N-gram* pada TF-IDF bertugas untuk mengelompokkan fitur pada dokumen. Terdapat pembagian secara *Unigram, Bigram*, dan *Trigram*. *Unigram* mengelompokkan fitur kata per satu kata, *Bigram* mengelompokkan fitur kata per dua kata, dan *Trigram* mengelompokkan fitur kata per tiga kata [21].

3.5. *Features Selection*

Pada tahap ini *dataset* yang sudah bersih akan dilakukan pemilihan fitur terbaik dengan menggunakan algoritma *Information Gain* (IG). Sebelum melakukan proses seleksi fitur, *dataset* diolah terlebih dahulu untuk mendapatkan daftar fitur yang akan diseleksi. Setiap fitur akan dimasukkan ke dalam sebuah kamus kata baru, yang di dalamnya tidak boleh ada kata yang sama. Selanjutnya semua fitur yang berada di dalam kamus kata tersebut, akan dilakukan perhitungan nilai IG, sehingga dari proses ini dapat diketahui apakah fitur IG relevan atau kurang relevan. *Information Gain* (IG) merupakan metode fitur seleksi yang menilai sebuah kata terhadap ada atau tidaknya kata tersebut dalam sebuah dokumen [7]. *Information Gain* (IG) menggunakan *entropy* sebagai parameter dalam menghitung nilai *gain* yang terdapat pada suatu kata. *Entropy* merupakan parameter yang digunakan untuk mengukur tingkat keberagaman sebuah data. Semakin beragam data, nilai *entropy* akan semakin besar. Untuk menghitung nilai *entropy* menggunakan persamaan 1.

$$Entropy\ (S) = -p_{(+)}log_2 p_{(+)} - p_{(-)}log_2 p_{(-)} \qquad (2)$$

Dimana S adalah ruang atau data sampel yang digunakan untuk latih, *p(+)* adalah probabilitas fitur tertentu yang bernilai positif pada data sampel, sedangkan *p(-)* adalah probabilitas fitur tertentu yang bernilai negatif pada data sampel. Setelah hasil perhitungan *entropy* didapatkan, maka selanjutnya nilai *entropy* tersebut digunakan dalam penghitungan gain. Nilai *gain* dari suatu fitur dapat dilakukan dengan menggunakan persamaan 2 [12].

$$Gain(S, A) = Entropy(S) - \sum_{v\ \in Values(A)} \frac{|S_v|}{|S|} Entropy\ (S_v) \qquad (3)$$

Dimana *A* adalah sebuah atribut, *v* merupakan suatu nilai yang mungkin untuk atribut *A*, *Values* (*A*) merupakan himpunan nilai yang mungkin untuk atribut *A* dan $|\mathcal{S}v|$ merupakan jumlah sampel untuk nilai *v*, $|\mathcal{S}|$ merupakan jumlah seluruh sampel data, dan *Entropy* $(\mathcal{S}v)$ merupakan *entropy* untuk sampel yang memiliki nilai *v* [12].

Tabel 2 continuous variable

Range Nilai Gain	Main
Nilai Gain < 0.01	Tidak
Nilai Gain > 0.01	Ya

Dimana pada Tabel 2 *continuous variable Information Gain* (IG) terdapat kelas *Main* (ya) dan *Main* (tidak), dimana dengan *range* nilai kecil dari 0.01 masuk kelas *Main* (tidak) dan nilai yang besar dari 0.01 masuk kelas *Main* (ya).

Untuk menentukan probabilitas penelitian ini menggunakan kNN-based Estimators, dimana pada bagian ini kami menyarankan kelas estimasi berbasis kNN yang membiaskan asumsi keseragaman lokal [30].

$$\hat{I}_{LNC}(x) = \hat{I}(x) - \frac{1}{N}\sum_{i=1}^{N} log\ \frac{\bar{V}(i)}{V(i)} \qquad (4)$$

Di sekitar titik *x(i)* yang berisi k tetangga terdekat, dinyatakan wilayah ruang ini yaitu *V(i)* $\subset \mathbb{R}^{\mathrm{d}}$, yang volumenya adalah *V(i)*. Asumsikan bahwa kerapatan seragam di dalam *V(i)* di sekitar titik *x(i)* , kita asumsikan bahwa ada beberapa subset, $\bar{V}(i) \subseteq V(i)$ dengan volume$\bar{V}(i) \leq V(i)$ dimana kerapatannya konstan, yaitu, $\hat{p} = \frac{II[x \in \bar{v}(i)]}{\bar{v}(i)}$. Kami sekarang mengulangi derivasi di atas menggunakan asumsi yang diubah tentang kerapatan lokal di sekitar setiap titik untuk $\hat{H}$(x). Penelitian ini tidak melakukan perubahan pada perkiraan entropi pada marginal subspaces [30].

3.6. Classification

Tahap selanjutnya adalah *classification* untuk melakukan klasifikasi sentimen dengan menggunakan algoritma *Support Vector Machine* (SVM). Secara garis besar, SVM bekerja dengan memetakan variabel-variabel data dalam beberapa dimensi dan mencoba membagi kelas menggunakan *hyperplane* yang memaksimalkan jarak antara setiap kelas dengan *hyperplane* [22]. *Hyperplane* yang optimal adalah yang memiliki *margin* terbesar, yaitu jarak antara *hyperplane* dengan *support vector* [23]. *Hyperplane* adalah garis batas pemisah data antar kelas, sedangkan *support vector* adalah data yang memiliki jarak terdekat dengan *hyperplane*. Proses *training* pada SVM merupakan proses pencarian bobot untuk setiap vektor di dalam data latih. Vektor-vektor dengan bobot lebih dari nol maka akan menjadi vektor yang mendefinisikan *hyperplane* pemisah. Vektor-vektor inilah yang kemudian disebut sebagai *support vector*. Untuk menentukan nilai *hyperplane*, pertama kita harus memaksimalkan nilai *margin* [24]. Dengan rumus berikut:

$$\frac{1}{2} \| w \|^2 \quad (5)$$
$$f = w.x_i + b = 0 \quad (6)$$

Dimana w menyatakan nilai parameter *hyperplane* yang dieksplorasi untuk mendapatkan garis tegak lurus yaitu antara garis *hyperplane* dan titik *support vector*, x_i menyatakan atribut x pada dokumen i, sedangkan b sebagai bias. Dari persamaan (5), kelas *hyperplane* dibagi menjadi dua yaitu kelas positif (+1) dan kelas negatif (-1), kemudian datanya diprediksi menggunakan (6) dan (7).

$$w.x_i + b \leq -1 \quad (7)$$
$$w.x_i + b \geq +1 \quad (8)$$

Persamaan (7) merupakan persamaan *hyperplane* untuk kelas negatif sedangkan (8) merupakan persamaan *hyperplane* untuk kelas positif, dengan menggunakan data *kernel linier* yang telah melewati proses sebelumnya yang dipisahkan dengan *hyperplane linier* dengan pelatihan proses yang dilakukan dengan menggunakan metode SVM. Penggunaan SVM dengan *karnel linier* dapat menghasilkan kinerja yang baik [25].

3.7. Evaluation

Evaluation merupakan tahap akhir dari seluruh rangkaian proses klasifikasi. Pada klasifikasi, untuk menghindari hasil yang *overfitting* dan mendapatkan hasil yang akurat, dalam pembagian set data ke sejumlah partisi, wajib menjalankan proses *stratification.* Peneltian ini menggunakan *Cross Validiation* yang bertujuan untuk mendapatkan performa yang lebih baik. Metode ini menjalankan percobaan sebanyak k kali pada sebuah model menggunakan parameter yang tidak berbeda dari sebelumnya [26].

Tabel 3 Tabel *Cross Validation*

Fold 1	Test	Train	Train	Train	Train
Fold 2	Train	Test	Train	Train	Train
Fold 3	Train	Train	Test	Train	Train
Fold 4	Train	Train	Train	Test	Train
Fold 5	Train	Train	Train	Train	Test

Tabel 3 merupakan gambaran dari 5 - *fold cross validation* merupakan memproses *fold* sebanyak 5 kali. *Fold* 1, membuat bagian pembatas awal menjadi data *testing* dan pembatas lainnya menjadi data *training*. *Fold* 2, membuat bagian pembatas kedua menjadi data *testing* dan pembatas lainnya menjadi data *training* dan begitu seterusnya. Hasil dari 5 *fold* ini, lalu akan diproses hasil performa dengan menerapkan metode *confusion matrix.* Dengan bantuan *confusion matrix*, akan dilakukan perhitungan akurasi yang bisa diprediksi dengan baik dan total data keseluruhan. Lalu dicari hasil rata-rata pada semua percobaan. Oleh karena itu didapatkan percobaan yang bisa menjadi acuan penggunaan sebuah model algoritma yang dipilih, berikut tabel *confusion matrix*.

Tabel 4 Tabel *Confusion Matrix*

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

Berdasarkan tabel *confusion matrix*, keempat istilah tersebut dijelaskan sebagai berikut:

1. *True Positive* (TP) yaitu data yang diprediksi positif dan klasifikasi aktualnya juga positif.
2. *False Positive* (FP) yaitu data yang diprediksi positif tetapi klasifikasi aktualnya negatif.
3. *False Negative* (FN) yaitu data yang diprediksi negatif tetapi klasifikasi aktualnya positif.
4. *True Negative* (TN) yaitu data yang diprediksi negatif dan klasifikasi aktualnya juga negatif.

Proses Evaluasi ini dinyatakan berdasarkan rumus berikut [27]]:

1. Recall

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

2. Precision

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

3. F-Measure

$$F - Measure = \frac{2 * recall * precision}{recall + precision} \quad (11)$$

4. Accuracy

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (12)$$

4. Evaluasi

Dalam penelitian ini dilakukan tiga kombinasi skenario pengujian untuk mendapatkan model yang memiliki performansi paling baik. Skenario pertama yaitu membandingkan *Prepocessing Stopwords* dan *Stemming*. Tujuan skenario ini untuk mengetahui pengaruh *Prepocessing* terhadap performa sistem analisis sentimen pada *movie review* berbahasa inggris menggunakan *Support Vector Machine* (SVM) dan *Information Gain* (IG). Skenario kedua yaitu membandingkan proses *Feature Selection* yaitu dengan melakukan proses klasifikasi *Support Vector Machine* (SVM) dan seleksi fitur *Information Gain* (IG) dengan *Support Vector Machine* (SVM) tanpa seleksi fitur *Information Gain* (IG). Tujuan dari skenario ini untuk mengetahui pengaruh performa menggunakan seleksi fitur *Information Gain* (IG). Sedangkan skenario ketiga yaitu fungsi *classifier* yang digunakan pada proses klasifikasi, yaitu *Linear, RBF*, dan *Polynomia*. Tujuan skenario ini untuk mengetahui pengaruh fungsi klasifikasi terhadap performa sistem.

4.1. Hasil Pengujian

4.1.1 Hasil Analisis Perbandingan *Prepocessing*

Pada skenario pertama dilakukan 3 kali pengujian dengan menerapkan 3 macam kombinasi teknik *preprocessing*, yaitu pertama proses *prepocessing* C*leaning, Case Folding, Tokenizing, Stopwords*, dan *Stemming*, kedua *prepocessing* C*leaning, Case Folding, Tokenizing*, dan *Stemming* (tanpa *Stopwords*), dan ketiga *prepocessing Cleaning, Case Folding, Tokenizing*, dan *Stopwords* (tanpa *Stemming*). Pengujian ini menggunakan SVM *linear* karena pada pengujian menggunakan SVM *linear* mendapatkan akurasi yang sangat bagus dibandingkan *kernel classifier* yang lain yaitu 86,12%, dengan batas *threeshold* 0,01, dan menggunakan seleksi fitur *Information Gain* (IG) . Pada hasil pengujian yang dilakukan menunjukkan proses *Stopwords* dapat menyebabkan akurasi menurun. Proses *Stopwords* merupakan proses membuang kata yang termasuk di dalam kamus NLTK berbahasa inggris. Sebagai contoh, kalimat yang seharusnya "don't make" jika dilakukan *Stopwords* akan menjadi "make", dan kata-kata tersebut akan menjadikan kalimat yang berbeda arti. Selain proses *Stopwords* juga dilakukan proses *Stemming* dimana proses ini mengubah kata-kata yang memiliki imbuhan menjadi bentuk kata dasar. Sebagai contoh yaitu, kata "moving" menjadi "move", pada proses *Stemming* akurasi yang didapatkan juga menurun, hal ini dikarenakan jika suatu kata tidak dilakukan proses *Stemming*, maka kata yang berimbuhan dan kata dasar akan mengandung makna yang berbeda, walaupun sebenarnya memiliki arti yang sama.

Tabel 5 Hasil Akurasi Perbandingan *Prepocessing*

Stopword	Stemming	Macro Precision	Macro Recall	Macro F1 Score	**Max Macro Score (%)**
√	√	86,97%	85,48%	86,12%	86,12%
-	√	85,56%	85,22%	85,38%	
√	-	85,31%	84,30%	84,80%	

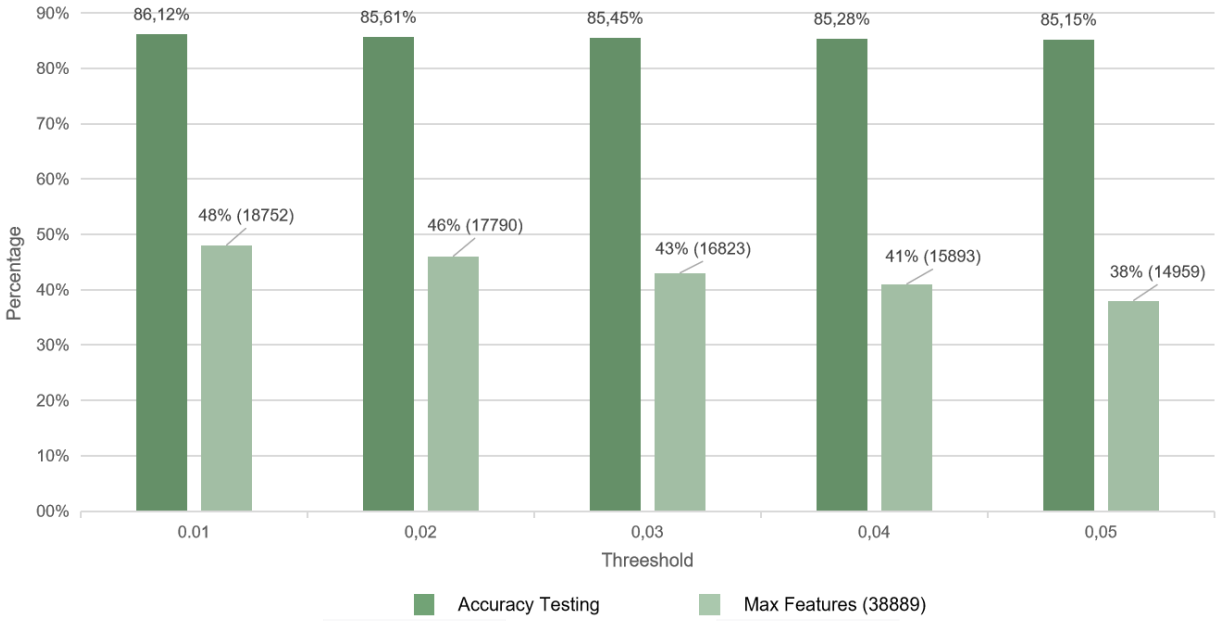
Berdasarkan tabel 5 bahwa hasil akurasi dengan dilakukannya kombinasi *Stopwords* dan *Stemming* mendapatkan hasil akurasi yang lebih tinggi yaitu 86,12%, dari pada dilakukan proses tanpa *Stopwords* yaitu 85,38%. Sedangkan hasil akurasi dengan tidak melakukan *Stemming* juga menurun yaitu 84,80%. Dapat kita ketahui bahwa data yang digunakan pada penelitian ini bernilai biner dengan keterangan positif (1) dan negatif (-1), karena didalam sebuah data terdapat banyak fitur sehingga pada kedua percobaan sistem tanpa *Stopwords* dan tanpa *Stemming* memilki rata - rata hasil yang hampir sama pada penjumlahan data yang dihasilkan sehingga perbedaan akurasi yang didapatkan tidak terlalu jauh. Oleh karena itu dapat disimpulkan bahwa jika dilakukan sebuah proses dengan menggunakan proses *Stopwords* tanpa *Stemming* mendapatkan akurasi terendah dan dengan dilakukan kombinasi *Stopword* dan *Stemming* maka akurasi yang dihasilkan akan lebih maksimal, karena proses *preprocessing* lebih lengkap.

4.1.2 Hasil Analisis Seleksi Fitur *Information Gain*

Pada skenario kedua dilakukan 2 kali pengujian menggunakan total 38889 fitur melakukan klasifikasi menggunakan fitur seleksi *Information Gain* (IG) dengan yang tidak menggunakan seleksi fitur *Informatioan Gain*

(IG), bertujuan untuk mengetahui pengaruhnya terhadap performa sistem. Pada pengujian menggunakan fitur seleksi *Information Gain* (IG), menggunakan batas *threeshold* 0,01 karena dapat kita lihat pada gambar 2 bahwa *threeshold* dengan menggunakan nilai 0,01 menghasilkan akurasi yang paling bagus, menggunakan seleksi fitur *Information Gain* (IG), *C* sama dengan 1.0 dan menggunakan klasifikasi SVM *linear*. Sedangkan dengan yang tidak menggunakan seleksi fitur *Informatioan Gain* (IG) menggunakan batas *threeshold* 0,01, *float* SVM sama dengan 1.0 dan menggunakan klasifikasi SVM *linear* (tanpa menggunakan seleksi fitur *Information Gain* (IG)).

Gambar 2 Jumlah fitur yang terambil terhadap Threeshold



Tabel 6 IG untuk top 10 term dengan nilai terbesar

Terms	Rank	IG
movi	101.047046	0.757243
film	91.020048	0.300472
one	88.351601	0.226903
like	83.312549	0.593962
watch	77.624936	0.261538
good	75.886700	0.210690
time	74.707352	0.000000
see	73.622371	0.319398
make	70.890644	0.166061
get	69.994565	0.434128

Pada tabel 6 kita dapat mengetahui top 10 pada IG, terdapat IG diurutkan berdasarkan *TF-IDF* yaitu nilai *Rank* terbesar 101.047046 dengan *Terms* sama dengan "movi". Jika kita lihat kembali pada tabel 6, ada *terms* dengan nilai IG = 0.0 namun nilai TF-IDF nya besar. Ini disebutkan keadaan normal, karena *Information Gain* (IG) akan mengkalkulasi derajat keterkaitan antara *term* sebagai *feature* (*input*) data dengan label sebagai *class* /(*output*) data.

Pada penelitian ini dilakukan proses validasi menggunakan *k-fold Cross-Validation* dengn k sebanyak 5 kali kemudian didapatkan hasil masing-masing k untuk sistem tanpa menggunakan *Information Gain* (IG) yaitu 85,6%, 86,2%, 86,0%, 86,4%, 85,8% dengan akurasi rata rata yaitu 86,0% sedangkan untuk sistem menggunakan *Information Gain* (IG) mendapatkan akurasi sebesar 84,0%, 85,0%, 84,2%, 86,2%, 82,6% dengan akurasi rata-rata yaitu 84,0%.

Tabel 7 Perbandingan akurasi menggunakan IG dan tanpa menggunakan IG

	Train	Validasi	Test
IG	94,09%	84,00%	86,12%

No IG	98,22%	86,00%	86,34%

Setelah didapatkan nilai validasi menggunakan *k-fold Cross-Validation* diketahui yaitu sistem yang dibangun ketika tidak melakukan seleksi fitur mengalami *overfitting*. Dari hasil pengujian menggunakan data validasi didapatkan selisih nilai akurasi *train* yang jauh lebih besar dari pada nilai akurasi validasi. Ini terjadi kemungkinan karena model yang dibuat terlalu fokus pada *training dataset*, sehingga tidak bisa melakukan prediksi dengan tepat jika diberikan *dataset* lain. Hal ini kemungkinan disebabkan juga karena jumlah data yang digunakan masih sedikit dan fitur yang digunakan pada data validasi dan data test berbeda dengan fitur pada data *train*, dan kemungkinan ada fitur-fitur yang tersedia pada data validasi dan data test tetapi tidak tersedia pada data *train* yang mengakibatkan terganggunya proses klasifikasi.

Terlihat Pada tabel 7 sistem yang telah dibangun ketika menggunakan seleksi fitur *Information Gain* (IG) memiliki akurasi yang hampir sama dengan sistem yang dibangun tanpa menggunakan seleksi fitur *Information Gain* (IG). Akurasi yang didapatkan pada data *test* 86,12% saat menggunakan seleksi fitur *Information Gain* (IG) dan tanpa seleksi fitur *Information Gain* (IG) mendapatkan akurasi sebesar 86,34%, seperti yang kita ketahui bahwa sistem tidak mengalami peningkatan akurasi pada data *test*. Hal ini karena sistem yang menggunakan *Information Gain* mengalami pengurangan fitur ketika memiliki nilai yang rendah atau mendekati 0, sehingga pada proses perhitungan *Gain* mendapatkan jumlah fitur (18752 fitur) menghasilkan performa yang menurun yaitu 0,22%.

Dan pada tabel 7, terlihat bahwa ketika menggunakan *Information Gain* (IG) akurasi yang diperoleh yaitu data *test* sebesar 86,12%, data validasi 84,00%, dan data *train* 94,09%. Sedangkan tanpa menggunakan IG, akurasi data *test* sebesar 86,34%, data validasi 86,00%, dan data *train* 98,22%. Sistem yang dibangun ketika menggunakan seleksi fitur mengalami dampak yang baik dalam mengurangi masalah *overfitting*, diketahui dengan adanya penurunan pada akurasi *train* yang mampu mendekati akurasi *test*. Oleh karena itu dapat disimpulkan bahwa seleksi fitur *Information Gain* (IG) merupakan sebuah solusi yang cukup baik untuk mengatasi masalah *overfitting* pada pengujian analisis sentimen ini, tetapi tidak cukup baik untuk mendapatkan performa yang lebih baik dari sistem yang hanya menggunakan SVM, karena dapat menyebabkan hilangnya beberapa informasi yang diperlukan untuk menentukan emosi. Dan sebagai konsekuensinya, penurunan dalam performa sistem.

Gambar 3 Perbandingan 20 data klasifikasi yang salah pada sistem dengan Information Gain dan tanpa Information Gain

Klasifikasi dengan Information Gain			
No	Review	Sebelum Klasifikasi	Setelah Klasifikasi
1	one found father regio	-1	1
2	thoughtless ignor illcor	-1	1
3	time time ive state pe	1	-1
4	mild spoiler basic plot	1	-1
5	dont ruin ill brief there	1	-1
6	huge rupert everett fa	1	-1
7	aussi shakespear 1824	-1	1
8	prussic ga murder don	1	-1
9	respons previou comm	-1	1
10	comment movi 1 thou	-1	1
11	look forward guardian	1	-1
12	movi inhabit famili fou	1	-1
13	good except end huge	1	-1
14	mostli routin factbas t	1	-1
15	yall hatin fact youd pr	1	-1
16	everyon name may so	1	-1
17	movi without doubt pe	1	-1
18	well movi terribl whon	-1	1
19	video guid masterpiec	-1	1
20	movi maker alway go	-1	1

Klasifikasi dengan tanpa Information Gain			
No	Review	Sebelum Klasifikasi	Setelah Klasifikasi
1	one found father region	-1	1
2	anita seem littl excus me	-1	1
3	want like one situat rich	-1	1
4	time time ive state peopl	1	-1
5	mild spoiler basic plot ou	1	-1
6	dont ruin ill brief there gr	1	-1
7	ordinarili anthoni mann r	1	1
8	aussi shakespear 1824 se	-1	1
9	respons previou commer	-1	1
10	comment movi 1 thought	-1	1
11	look forward guardian w	-1	1
12	good except end huge dis	-1	1
13	rich tycoon kill plane cras	-1	1
14	yall hatin fact youd proba	-1	1
15	movi without doubt perf	-1	1
16	well movi terribl whomev	1	-1
17	wonder quirki romant ita	-1	1
18	video guid masterpiec ye	1	-1
19	say doctor might conjur i	1	-1
20	harrison ford play play co	1	-1

Kemudian telah dilakukan analisis terhadap data pada gambar 3 yang masih salah dalam klasifikasi. Diambil 20 data dari masing-masing rancangan sistem yang menggunakan IG dan tidak menggunakan IG. Didapatkan 13 (kolom bewarna hijau) dari 20 data tersebut memiliki salah klasifikasi yang sama. Ditemukan bahwa 7 data lagi kemungkinan karena penyebab salah klasifikasi ketika sistem menggunakan *Feature Selection Information Gain* (IG) yaitu ketika suatu kata terdapat pada data *review* didalam perangkingan dan bernilai 0 kemungkinan menyebabkan data tersebut salah karena dapat membuat sistem tidak stabil dalam memprediksi. Sedangkan sistem tanpa menggunakan *Feature Selection Information Gain* (IG), ditemukan bahwa penyebab salah dalam klasifikasi

ketika suatu data *review* memilki banyak kata yang memiliki jumlah bobot yang besar sehingga menyebabkan salah dalam klasifikasi.

4.1.3 Hasil Analisis Perbandingan *Classifier*

Hasil dari proses analisis yang dilakukan pada skenario ini adalah pendugaan fungsi *kernel* terbaik, akurasi ketepatan klasifikasi pada data pelatihan dan pengujian menggunakan tiga fungsi *kernel* yaitu *linear*, *RBF*, dan *Polynomial*. Pada skenario ini dilakukan 3 kali pengujian dengan melakukan perbandingan klasifikasi menggunakan tiga fungsi *kernel* berbeda yaitu *linear*, *RBF*, dan *Polynomial*. Setiap percobaan ketiga kernel klasifikasi menggunakan batas *threeshold* 0,01, *C* sama dengan 1.0, dan menggunakan seleksi fitur *Information Gain* (IG). Fungsi *kernel* terbaik dan akurasi ketepatan klasifikasi dapat dilihat pada tabel 8.

Tabel 8 Hasil Akurasi Perbandingan *Classifier*

Classifier	***Macro Precision***	***Macro Recall***	***Macro F1 Score***	***Max Macro F1 Score (%)***
Linear	86,97%	85,48%	86,12%	86,12%
RBF	87,32%	84,56%	85,91%	
Polynomial	88,67%	60,94%	72,17%	

Berdasarkan pada Tabel 8, hasil yang didapatkan dari ketiga *classifier* tersebut adalah *Linear* dan *classifier RBF* memiliki akurasi yang besar dibandingkan *Polynomial*. *Linear* mendapatkan akurasi 86,12%, 85,48% *recall*, 86,97% *precision* dan *RBF* mendapatkan akurasi 85,91%, 84,56% *recall*, dan 87,32% *precision* hanya saja *Linear* memiliki akurasi sedikit lebih tinggi yaitu 86,12% sedangkan *RBF* 85,91% dibandingkan fungsi kernel *Polynomial* yang mendapatkan akurasi yang rendah yaitu 72,17%, *precision* 88,67%, dan *recall* 60,94%.

Hal ini menunjukan bahwa klasifikasi analisis sentimen *movie review* sangat tepat diklasifikasikan menggunakan algoritma *Support Vector Machine* (SVM) *kernel Liniear* dengan *default* parameter. Dikarenakan *Linear* berfokus dengan fitur-fitur yang mengandung nilai biner yang didapatkan dari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. Seperti pada penelitian ini menggunakan nilai dari masing-masing label yang menggunakan notasi 1 dan -1, sehingga *classifier* yang tepat adalah *Linear* pada penelitian ini.

5. Kesimpulan

Berdasarkan penelitian yang telah dilakukan untuk analisis sentimen pada *movie review* berbahasa inggris, dengan melakukan proses *Prepocessing* (C*leaning, Case Folding, Tokenizing, Stopwords, dan Stemming*), pembobotan kata menggunakan TF-IDF, seleksi fitur *Information Gain* (IG), dengan *Split Data* menggunakan metode *Train/Test Split*, proses klasifikasi menggunakan *Support Vector Machine* (SVM) menggunakan batas *threeshold* 0,01, *C* sama dengan 1.0, dan pengukuran performansi *Confusion matrix*.

Berdasarkan pengujian yang telah dilakukan bahwa hasil akurasi dengan dilakukannya kombinasi *Stopwords* dan *Stemming* mendapatkan hasil akurasi yang lebih tinggi yaitu 86,12%, dari pada dilakukan proses tanpa *Stopwords* yaitu 85,38%. Sedangkan hasil akurasi dengan tidak melakukan *Stemming* juga menurun yaitu 84,80%. Dapat kita ketahui bahwa data yang digunakan pada penelitian ini bernilai biner dengan keterangan positif (1) dan negatif (-1), karena didalam sebuah data terdapat banyak fitur sehingga pada kedua percobaan sistem tanpa *Stopwords* dan tanpa *Stemming* memilki rata - rata hasil yang hampir sama pada penjumlahan data yang dihasilkan sehingga perbedaan akurasi yang didapatkan tidak terlalu jauh. Oleh karena itu dapat disimpulkan bahwa jika dilakukan sebuah proses dengan menggunakan proses *Stopwords* tanpa *Stemming* mendapatkan akurasi terendah dan dengan dilakukan kombinasi *Stopword* dan *Stemming* maka akurasi yang dihasilkan akan lebih maksimal, karena proses *preprocessing* lebih lengkap. Selain itu pada pengujian kedua, terlihat bahwa ketika menggunakan *Information Gain* (IG) akurasi yang diperoleh yaitu data *test* sebesar 86,12%, data validasi 84,00%, dan data *train* 94,09%. Sedangkan tanpa menggunakan IG, akurasi data *test* sebesar 86,34%, data validasi 86,00%, dan data *train* 98,22%. Sistem yang dibangun ketika menggunakan seleksi fitur mengalami dampak yang baik dalam mengurangi masalah *overfitting*, diketahui dengan adanya penurunan pada akurasi *train* yang mampu mendekati akurasi *test*. Oleh karena itu dapat disimpulkan bahwa seleksi fitur *Information Gain* (IG) merupakan sebuah solusi yang cukup baik untuk mengatasi masalah *overfitting* pada pengujian analisis sentimen ini, tetapi tidak cukup baik untuk mendapatkan performa yang lebih baik dari sistem yang hanya menggunakan SVM, karena dapat menyebabkan hilangnya beberapa informasi yang diperlukan untuk menentukan emosi. Dan sebagai konsekuensinya, penurunan dalam performa sistem.

Kemudian pada proses *classifier* diperoleh bahwa fungsi *kernel* yang tepat digunakan dalam menentukan ketepatan klasifikasi. Hasil yang didapatkan dari ketiga *classifier* tersebut adalah *Linear* dan *classifier RBF* memiliki akurasi yang besar dibandingkan *Polynomial*. *Linear* mendapatkan akurasi 86,12%, 85,48% *recall*, 86,97% *precision* dan *RBF* mendapatkan akurasi 85,91%, 84,56% *recall*, dan 87,32% *precision* hanya saja *Linear* memiliki akurasi sedikit lebih tinggi yaitu 86,12% sedangkan *RBF* 85,91% dibandingkan fungsi kernel *Polynomial*

yang mendapatkan akurasi yang rendah yaitu 72,17%, *precision* 88,67%, dan *recall* 60,94%. Hal ini menunjukan bahwa klasifikasi analisis sentimen *movie review* sangat tepat diklasifikasikan menggunakan algoritma *Support Vector Machine* (SVM) *kernel Liniear* dengan *default* parameter. Dikarenakan *Linear* berfokus dengan fitur-fitur yang mengandung nilai biner yang didapatkan dari *hyperplane* terbaik dengan memaksimalkan jarak antar kelas. Seperti pada penelitian ini menggunakan nilai dari masing-masing label yang menggunakan notasi 1 dan -1, sehingga *classifier* yang tepat adalah *Linear* pada penelitian ini.

Saran untuk penelitian selannjutnya supaya memperoleh hasil terbaik pada penelitian selanjutnya, untuk menggunakan suatu metode yang bisa mengklasifikasikan data komentar menjadi lebih *detail* sesuai dengan kriteria. Hal tersebut untuk memeroleh kalimat sentimen pada masing-masing komentar berdasarkan kriteria. Dan juga mencoba melakukan proses seleksi fitur yang lain selain *Information gain* (IG) sehingga menghasilkan performa yang lebih maksimal.

REFERENSI

[1] J. C. Na, T. T. Thet, and C. S. G. Khoo, "Comparing sentiment expression in movie reviews from four online genres," *Online Inf. Rev.*, vol. 34, no. 2, pp. 317–338, 2010, doi: 10.1108/14684521011037016.

[2] T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection and classification algorithms," *Int. Conf. Microelectron. Comput. Commun. MicroCom 2016*, 2016, doi: 10.1109/MicroCom.2016.7522583.

[3] 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.

[4] F. E. Cahyanti, Adiwijaya, and S. Al Faraby, "On the Feature Extraction for Sentiment Analysis of Movie Reviews Based on SVM," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166397.

[5] M. Cindo, "Studi Komparatif Metode Ekstraksi Fitur pada Analisis Sentimen," vol. 1, no. 10, pp. 9–12, 2021.

[6] S. N. Kane, A. Mishra, and A. K. Dutta, "Preface: International Conference on Recent Trends in Physics (ICRTP 2016)," *J. Phys. Conf. Ser.*, vol. 755, no. 1, 2016, doi: 10.1088/1742-6596/755/1/011001.

[7] R. xin Nie, Z. peng Tian, J. qiang Wang, and K. S. Chin, "Hotel selection driven by online textual reviews: Applying a semantic partitioned sentiment dictionary and evidence theory," *Int. J. Hosp. Manag.*, vol. 88, no. March, p. 102495, 2020, doi: 10.1016/j.ijhm.2020.102495.

[8] C. Nanda, M. Dua, and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," *Proc. 2018 IEEE Int. Conf. Commun. Signal Process. ICCSP 2018*, pp. 1069–1072, 2018, doi: 10.1109/ICCSP.2018.8524223.

[9] S. S. and P. K.V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, no. xxxx, pp. 2–7, 2020, doi: 10.1016/j.icte.2020.04.003.

[10] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," *Appl. Comput. Intell. Soft Comput.*, vol. 2018, 2018, doi: 10.1155/2018/1407817.

[11] W. Zheng and Q. Ye, "Sentiment classification of Chinese traveler reviews by support vector machine algorithm," *3rd Int. Symp. Intell. Inf. Technol. Appl. IITA 2009*, vol. 3, pp. 335–338, 2009, doi: 10.1109/IITA.2009.457.

[12] 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.

[13] R. Fikri Aldi Nugraha, Nisa Hanum Harani, *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning*. Kreatif, 2020.

[14] H. S. S. B, *Politik Komunikasi*. Grasindo.

[15] T. M. Eduka, *Top Fokus Ulangan & Ujian SMP: Ulangan & Ujian SMP*. Genta Group Production, 2020.

[16] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," *ACL 2007 - Proc. 45th Annu. Meet. Assoc. Comput. Linguist.*, no. June, pp. 976–983, 2007.

[17] M. Y. H. S. S. K. M. K. R. H. S. K. M. T. Irfan Mayendra Putra, *PANDUAN LENGKAP KLASIFIKASI DOKUMEN ARSIP PROGRAM STUDI MENGUNAKAN SUPPORT VECTOR MACHINE*. CV. Kreatif Industri Nusantara, 2020.

[18] I. Werdiningsih and B. Nuqoba, *Data Mining Menggunakan Android, Weka, Dan Spss*. Airlangga University Press, 2020.

[19] A. Rahmansyah, O. Dewi, P. Andini, T. Hastuti, P. Ningrum, and M. E. Suryana, "Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine," *Semin. Nas. Apl. Teknol. Inf.*, pp. 1–7, 2018.

[20] M. S. Park and J. Y. Choi, "Theoretical analysis on feature extraction capability of class-augmented PCA," *Pattern Recognit.*, vol. 42, no. 11, pp. 2353–2362, 2009, doi: 10.1016/j.patcog.2009.04.011.

[21] A. Hamzah, "Deteksi Bahasa Untuk Dokumen Teks," *Semin. Nas. Inform.*, vol. 22, no. semnasIF, pp. 5–13, 2010.

[22] F. R. S. Rangkuti, M. A. Fauzi, Y. A. Sari, and E. D. L. Sari, "Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection," *3rd Int. Conf. Sustain. Inf. Eng. Technol. SIET 2018 - Proc.*, pp. 88–91, 2018, doi: 10.1109/SIET.2018.8693211.

[23] Y. X. Chu, X. G. Liu, and C. H. Gao, "Multiscale models on time series of silicon content in blast furnace hot metal based on Hilbert-Huang transform," *Proc. 2011 Chinese Control Decis. Conf. CCDC 2011*, pp. 842–847, 2011, doi: 10.1109/CCDC.2011.5968300.

[24] "Analisis Sentimen Menggunakan Support Vector Machine dan Maximum Entropy Sentiment," pp. 1–7, 2017, [Online].

[25] S. Al Faraby, E. R. R. Jasin, A. Kusumaningrum, and Adiwijaya, "Classification of hadith into positive suggestion, negative suggestion, and information," *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012046.

[26] B. Santosa and A. Umam, *Data Mining dan Big Data Analytics: Teori dan Implementasi Menggunakan Python & Apache Spark*. Penebar Media Pustaka.

[27] R. P. Nawangsari, R. Kusumaningrum, and A. Wibowo, "Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study," *Procedia Comput. Sci.*, vol. 157, pp. 360–366, 2019, doi: 10.1016/j.procs.2019.08.178.

[28] Maas A L, Daly R E, Pham P T, Huang D, Ng A Y, dan Potts C. 2011. Learning Word Vectors for Sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 142– 150

[29] 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.

[30] A. Kraskov, H. Stogbauer and P. Grassberger, "Estimating mutual information". Phys. Rev. E 69, 2004.