Klasifikasi Sentimen Ulasan Film Menggunakan Support Vector Machine, Information Gain, dan N-Grams

1st Rizky Hilman Faturrahman
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
rizkyhilmanf@student.telkomuniversity.ac.id

2nd Widi Astuti
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
widiwdu@telkomuniversity.ac.id

3rd Mahendra Dwifebri Purbolaksono
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
mahendradp@telkomuniversity.ac.id

Abstrak

Saat ini, orang menggunakan Internet untuk menulis opini mereka di blog, jejaring sosial, dan situs web. Situs review seperti IMDb adalah salah satu situs yang sering dikunjungi oleh pengguna Internet yang menyediakan informasi yang lengkap tentang aktor, kru, peringkat dan ulasan film yang diberikan oleh orang lain. ulasan dan peringkat film tersebut dapat mempengaruhi perilaku pembelian mereka. Pendekatan *machine learning* digunakan untuk memecahkan permasalahan ambiguitas opini dengan mengklasifikasikan ulasan film tersebut ke dalam sebuah sentimen. Pada penelitian ini, 50.000 dataset dari IMDb akan digunakan untuk melakukan pengujian klasifikasi *Support Vector Machine* dengan seleksi fitur *Information Gain* untuk membantu performansi dari klasifikasi SVM. Dari hasil pengujian, didapat nilai akurasi tertinggi sebesar 86% untuk unigram dan 76,4% untuk bigram pada klasifikasi SVM dengan menggunakan Information Gain sebagai seleksi fiturnya.

Kata kunci : SVM, analisis sentimen, Information Gain, ulasan ilm, n-grams

Abstract

Nowadays, people use the Internet to write their opinions on blogs, social networks, and websites. Review sites like IMDb are one of the most frequently visited sites by Internet users that provide comprehensive information about actors, crew, ratings and reviews of films given by others. The reviews and ratings of these films can influence their buying behavior. Machine learning approach is used to solve the problem of ambiguity of opinion by classifying the film review into a sentiment. In this study, 50.000 datasets from IMDb will be used to test the Support Vector Machine classification with Information Gain feature selection to help the performance of the SVM classification. From the test results, obtained the highest accuracy value of 86.1% for unigram and 76,2% for bigram in SVM classification using Information Gain as the feature selection.

Keywords: SVM, sentiment analysis, Information Gain, movie review, n-grams

I. PENDAHULUAN

A. Latar Belakang

Saat ini orang menggunakan *Internet* untuk menulis opini mereka di *blog*, jejaring sosial, dan situs *web*. Analisis sentimen atau *opinion mining* menjadi sangat penting bagi pengguna dan juga untuk bisnis [1]. Situs *review* seperti IMDb adalah salah satu situs yang sering dikunjungi oleh pengguna *Internet*, situs tersebut menyediakan informasi yang sangat lengkap mengenai perfilman mulai dari aktor, kru, sinopsis, peringkat dan ulasan film yang diberikan oleh pengguna lain [2][3]. Ulasan dan peringkat dari film tersebut dapat mempengaruhi perilaku orang-orang untuk pergi ke teater/bioskop atau untuk membeli sebuah produk.

Klasifikasi sentimen adalah sub divisi baru dari klasifikasi teks yang memperhatikan topik dan pendapat yang diungkapkan pada sebuah dokumen. Analisis statistik dari sentimen dapat memberikan gambaran yang kuat terhadap berita baru yang mempengaruhi sebuah entitas yang penting [4]. Dalam memecahkan permasalahan yang ada pada klasifikasi sentimen, pendekatan *machine learning* memiliki performansi yang bagus akan tetapi memakan banyak waktu dalam memproses model agar sesuai dengan data latih. Penggunaan kombinasi dari *machine learning* dan *feature selection* dapat mengurangi waktu yang diperlukan dan meningkatkan performansi dari machine learning [5].

Tim O'Keefe dan Irena Koprinska meneliti pengaruh *feature selection* dan *weighting method* pada analisis sentimen menggunakan klasifikasi *Naïve Bayes* (NB) dan *Support Vector machine* (SVM) [6]. Dari penelitian tersebut didapatkan hasil akurasi tertinggi 87.15% dengan menggunakan *Categorical Feature Difference* (PD) sebagai *feature selection* dan SVM sebagai metode klasifikasi. Penelitian selanjutnya, meneliti penggunaan *Information Gain* (IG) dengan 4 metode klasifikasi yaitu NB, SVM, *Random Forest* (RF) *dan K-Nearest Neighbour* (KNN). Dari keempat metode klasifikasi tersebut, SVM mendapatkan akurasi terendah tanpa menggunakan IG [5]. Namun dengan menggunakan IG dapat terlihat peningkatan akurasinya.

Metode klasifikasi SVM terbukti mengungguli sebagian besar metode klasifikasi lain pada data text [7]. Namun, *performance* dari *classifier* bersangkutan dengan *training corpus*-nya, dan *training corpus* yang baik dapat memperoleh classifier dengan performance yang baik [8]. Oleh karena itu, penelitian ini berfokus pada seleksi fitur Information Gain dan n-grams menggunakan klasifikasi SVM terhadap klasifikasi *sentiment movie review* untuk melihat pengaruh penggunaan pemilihan fitur dalam menentukan sentimen dari *movie review*.

B. Topik dan Batasannya

Topik penelitian yang diambil penulis adalah pengaruh Information Gain terhadap model evaluasi Support Vector Machine dan Naïve Bayes yang menggunakan N-Grams TF-IDF pada dataset movie review untuk mendapatkan nilai akurasi dari hasil evaluasi. Batasan yang terdapat pada penelitian ini yaitu: Pertama, feature yang akan digunakan pada tahap ekstraksi dan seleksi di batasi menjadi 3000 feature. Dikarenakan keterbatasan hardware ataupun cloud service yang digunakan penulis. Kedua, N-grams yang akan digunakan yaitu unigram dan bigram, untuk trigram dan seterusnya tidak dapat diaplikasikan karena dibutuhkan waktu yang cukup lama untuk melakukan proses feature seleksinya.

C. Tujuan

Tujuan yang akan dicapai dalam penelitian tugas akhir ini adalah menerapkan *Support Vector Machine* dalam mengklasifikasikan teks ulasan film serta mengukur performansi dari model yang dibangun menggunakan *Information Gain dan n-grams.*

D. Organisasi Tulisan

Urutan penulisan selanjutnya adalah studi terkait yang merupakan penjelasan singkat mengenai istilah-istilah yang digunakan. Berikutnya, adalah sistem yang dibangun berisi mengenai detail penjelasan metode dan tahapan-tahapan sistem yang dibangun. Selanjutnya, pada evaluasi berisi hasil dari pengujian sistem dan analisis dari sistem tersebut. Pada kesimpulan berisi rangkuman hasil yang diperoleh dan saran untuk penelitian selanjutnya. Daftar Pustaka berisikan sumber literatur penunjang dari penelitian tugas akhir ini.

II. KAJIAN TEORI

A. Penelitian Terdahulu

Banyak penelitian terkait analisis sentimen ulasan film menggunakan pendekatan yang bervariasi, mulai dari metode klasifikasinya, seleksi fiturnya, dan *preprocessing*.

Pada penelitian [3], dibandingkan beberapa metode klasifikasi yaitu NB, SVM, dan KNN dengan SVM+IG. Hasil yang didapatkan SVM+IG memiliki akurasi tertinggi yaitu 85,65% dan memiliki selisih 2,6% dengan SVM yang tidak menggunakan IG dan dapat dikatakan bahwa IG dapat mempengaruhi peningkatan performance dari metode klasifikasi SVM.

Penelitian berikutnya [5], menggunakan pendekatan IG namun lebih berfokus kepada klasifikasi KNN. Penelitian tersebut membahas pengaruh IG terhadap KNN dan juga ke 3 metode lain yaitu NB, SVM, dan RF. Hasil yang didapatkan dari penelitian IG dengan threshold 0.1, 0.2,0.3,0.4 dan 0.5 tersebut adalah NB memiliki nilai akurasi rata-rata yang sangat tinggi yaitu 93%, lalu KNN mendapatkan peningkatan yang drastis pada threshold 0.3, 0.4 dan 0.5 dengan akurasi rata-rata 96%. Dengan meningkatnya nilai threshold dari IG, maka semakin sedikit fitur yang dipilih. Pada kasus SVM, nilai akurasi yang didapat pada saat menggunakan IG menurun seiring bertambahnya nilai threshold IG, hal tersebut terjadi karena kualitas dari term dataset seperti SVM perlu dipisahkan dan akan sulit untuk membangun model jika fiturnya tidak dipisahkan secara linier.

Penelitian [9] membahas perbandingan feature selection dan teknik machine learning secara menyeluruh. Feature selection yang digunakan adalah Document Frequency (DF), IG, Gain Ratio (GR), CHI statistic (CHI), dan Relief-F Algorithm. Dan metode klasifikasi yang digunakan adalah NB, SVM, Max Entropy (MaxEnt), Decision Tree (DT), dan Adaboost classifier. Semua seleksi fitur dan metode klasifikasinya diujikan dalam penelitian tersebut dan didapatkan hasil bahwa NB dengan GR memiliki akurasi tertinggi yaitu 90,90% sementara SVM memiliki akurasi 90,15%. Namun pada saat menggunakan IG, SVM memiliki akurasi tertinggi yaitu 89,20% sementara NB memiliki akurasi 88,85%.

B. *Sentiment Classification*

Klasifikasi sentimen atau dikenal juga dengan klasifikasi polarity telah menjadi disiplin ilmu yang penting di bidang Natural Language Processing (NLP), penambangan teks, dan pencarian informasi. Klasifikasi dilakukan pada beberapa tingkatan unit teks. Unit teks bisa berupa kata, frase, kalimat atau dokumen [10]. Opini pada sebuah review yang ada di Internet seringkali tidak terstruktur atau semi-terstruktur sehingga penggunaan seleksi fitur yang tepat menjadi masalah yang penting. Dan juga fitur sentimen tidak diekspresikan secara objektif dan eksplisit sehingga klasifikasi sentimen memerlukan analisis dan pemahaman yang lebih dalam tentang fitur tekstual [3].

C. Support Vektor Machine

Support vector machine (SVM) sangat efektif dalam kategorisasi teks tradisional, dan dapat mengungguli Naive Bayes. SVM mencari hyper-plane yang diwakili oleh vektor yang memisahkan vektor pelatihan positif dan negatif dari dokumen dengan margin maksimum.

SVM mencari decision surface untuk memisahkan poin data latih menjadi dua kelas dan membuat keputusan berdasarkan support vector yang dipilih sebagai satu-satunya elemen efektif dalam training set-nya. Optimalisasi SVM (bentuk ganda) adalah untuk meminimalkan [8].

$$a^{\rightarrow} = \arg\min\{-\sum_{i=1}^{n} a + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j \langle x_i , x_j \rangle\} \quad (1)$$

Subject terhadap:

$$\sum_{i=1}^{n} a_i y_i = 0;\ 0 \le a_i \le C \quad (2)$$

D. Information Gain

Information gain (IG) adalah salah satu teknik pemilihan fitur untuk klasifikasi sentimen. IG digunakan untuk memilih fitur yang berhubungan dengan atribut kelas. Pemilihan fitur diukur dengan pengurangan ketidakpastian dalam mengidentifikasi atribut kelas ketika nilai fitur diketahui. Fitur dengan peringkat teratas dipilih untuk mengurangi ukuran vektor fitur yang akan menghasilkan hasil klasifikasi yang lebih baik [11].

$$\text{IG (t)} = |- \sum_{J=1}^{K} P(C_J)\log(P(C_J)) + P(w)\sum_{J=1}^{K} P(C_J \mid w) + p(\bar{w})\sum_{J=1}^{K} P(C_J \mid \bar{w})\log P(C_J \mid \bar{w}) \quad (3)$$

Pada rumus diatas, $P(C_J)$ adalah pecahan dari jumlah dokumen yang dimiliki kelas C_J dari total dokumen dan $P(w)$ adalah pecahan dari dokumen dimana muncul term

w. $P(C_j \mid w)$ dihitung sebagai pecahan dokumen dari kelas C_J yang memiliki term w.

E. TF-IDF

TF-IDF adalah intuisi heuristik bahwa query term yang muncul di banyak dokumen bukanlah pembeda yang baik, dan harus diberi bobot yang lebih kecil daripada yang muncul di beberapa dokumen. Dibawah ini adalah rumus TF-IDF yang digunakan untuk pembobotan term [11].

$$W_{ij} = tf_{ij} \; x \log \left(\frac{N}{df_i}\right) \quad (4)$$

Term frequency (tf) adalah jumlah kemunculan kata dalam sebuah dokumen. N adalah jumlah total kumpulan data. df adalah jumlah dokumen yang berisi fitur-fitur terkait, dan ij adalah atribut yang dibobot. Kata-kata yang telah melalui preprocessing akan dibobot menggunakan rumus TF-IDF. Hasil akhir dari proses ekstraksi ciri ini adalah kata-kata unik yang telah diberi bobot.

F. N-Grams Model

Model N-Gram adalah metode pengecekan kata 'n' kontinu atau suara dari urutan teks atau ucapan tertentu. Model n-gram membantu menganalisis sentimen teks atau dokumen. Unigram mengacu pada n-gram ukuran 1, Bigram mengacu pada n-gram ukuran 2, Trigram mengacu pada n-gram ukuran 3 dan seterusnya [12].

G. Confusion Matrix

Confusion Matrix mengandung informasi klasifikasi sebenarnya dan klasifikasi prediksi hasil dari sistem klasifikasi. Peforma dari sistem yang dibangun biasanya dievaluasi dengan data yang ada didalam matrix. Dengan Confussion Matrix, nilai akurasi, precission, recall, dan f1-score bisa diperoleh. Terdapat empat kemungkinan yang akan terjadi [9].

1. True Positive (TP) : merupakan jumlah dari data positif yang benar terprediksi oleh sistem klasifikasi.
2. True Negative (TN) : merupakan jumlah dari data negatif yang benar terprediksi oleh sistem klasifikasi.
3. False Negative (FN) : merupakan jumlah dari data positif yang terklasifikasi sebagai data negatif oleh sistem klasifikasi.
4. False Postive (FB) : merupakan jumlah data negatif yang terklasifikasi sebagai data positif oleh sistem klasifikasi

Akurasi, presentasi dari hasil yang diklasifikasikan benar. Akurasi dapat dihitung menggunakan persamaan 5.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision, nilai dari presentasi hal yang relevan. Precision dapat dihitung menggunakan persamaan 6.

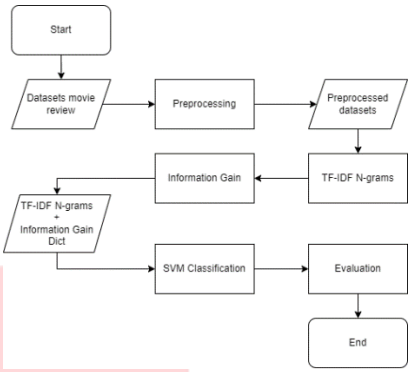
$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall, nilai dari presentasi total hasil relevan yang juga diklasifikasikan benar oleh sistem yang dibangun. Recall dapat dihitung menggunakan persamaan 7.

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

III. METODE

Sistem yang dibangun bertujuan untuk menentukan klasifikasi dari movie review dengan tahapan-tahapan yang digambarkan agar mempermudah ilustrasi pembaca. Pada Gambar 1 dapat dilihat gambaran umum dari sistem yang akan dibangun dalam tugas akhir ini.



Gambar 1: Gambaran Umum Sistem

A. Dataset

Dataset yang akan digunakan dalam penelitian ini diambil dari dataset IMDb pada website "*https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews*". Terdapat 50000 data yang dapat diolah dan terbagi menjadi 25000 review positif dan 25000 review negatif. Selanjutnya akan dilakukan preprocessing terhadap dataset tersebut untuk digunakan pada tahap selanjutnya yaitu *feature extraction*.

B. Preprocessing

Tujuan dari preprocessing adalah mengubah data menjadi terstruktur sehingga dataset dapat diolah. Pertama, setiap kata akan diubah menjadi huruf kecil untuk menyamakan struktur setiap kata. Kemudian proses penghilangan *non-alphabetical*, *stopword* dan tanda baca. Proses selanjutnya adalah stemming untuk menyederhanakan kata dengan imbuhan. Stemming adalah proses mereduksi kata-kata yang diubah ke bentuk kata dasar atau akar kata, tujuannya adalah untuk menyatukan berbagai bentuk kata. Setelah itu dataset akan melalui proses tokenisasi yaitu setiap data ulasan akan dibagi menjadi urutan kata. Hasil dari proses preprocessing ini berupa kamus kata unik dan akan digunakan sebagai fitur.

C. N-Grams TF-IDF

Feature extraction adalah proses mengubah input data ke dalam serangkaian fitur. Kinerja dari proses machine learning sangat bergantung pada fitur-fiturnya sehingga sangat penting untuk memilih fitur. Tujuannya adalah untuk meringkas dan mengubah data input menjadi satu set fitur representasi yang bekerja dengan tepat untuk pengklasifikasian. Pada tulisan ini ekstraksi fitur yang digunakan yaitu TF-IDF dengan bentuk unigram dan bigram.

D. Information Gain

Metode fitur seleksi yang digunakan adalah Information gain. Information gain adalah metode yang menunjukkan pentingnya fitur untuk kelas atribut. Dalam penelitian ini threshold untuk nilai IG yang digunakan yaitu 0.01, 0.03, 0.05, dan 0.1 setelah itu dilakukan juga normalisasi *l1*.

E. SVM Classification

Pada tahap klasifikasi, metode klasifikasi support vectore machine akan digunakan fitur-fitur yang telah dipilih pada tahap feature extraction dan feature selection. Support vector machine adalah metode klasifikasi data linier, dengan menggunakan pemetaan non linier untuk mengubah data latih menjadi dimensi yang lebih tinggi. setiap item data diplot sebagai titik dalam ruang n-dimensi (di mana 'n' adalah jumlah fitur) dengan nilai setiap fitur menjadi nilai koordinat tertentu. Kemudian dilakukan klasifikasi dengan mencari hyper-plane yang membedakan kedua kelas tersebut dengan sangat baik sehingga contoh dari kategori yang terpisah tersebut terbagi dengan celah yang jelas selebar mungkin.

IV. HASIL DAN PEMBAHASAN

Pada penelitian ini akan dilakukan 3 skenario uji, yang pertama adalah melakukan evaluasi terhadap SVM tanpa menggunakan Information gain dan membandingkan nilai akurasi dengan Naive bayes, data di split sebanyak 70% training data dan 30% test data menggunakan 5-Fold Cross Validation (5-FCV) agar jumlah data pada setiap kelas semirip mungkin. Skenario test yang kedua, membandingkan nilai akurasi SVM dengan Naive bayes menggunakan Information gain. Threshold yang digunakan untuk information gain yaitu 0.01, 0.03, 0.05 dan 0.1 dengan 5-Fold Cross Validation. Lalu skenario ketiga yaitu evaluasi terhadap klasifikasi bigram tanpa menggunakan stopword removal. Semua skenario pengujian menggunakan library NLTK, Skicit-Learn, Scipy, Pandas dan Numpy [13][14][15][16][17].

Table 1 Skenario pengujian klasifikasi

No	Stop word Removal	TF-IDF	Fitur Seleksi IG	Treshold	N-grams	Klasifikasi	5 Fold-Cross Validation
1	Y	3000 fitur	-	-	Unigram, Bigram (1,2)	SVM, Naïve Bayes	Y
2	Y	3000 fitur	Y	0.01, 0.03, 0.05, dan 0,1	Unigram, Bigram (1,2)	SVM, Naïve Bayes	Y
3	-	3000 fitur	Y	0.01, 0.03, 0.05, dan 0,1	Unigram, Bigram (1,2)	SVM, Naïve Bayes	Y
4	-	3000 fitur	-	-	Unigram, Bigram (1,2)	SVM, Naïve Bayes	Y

A. Hasil Pengujian scenario 1 tanpa menggunakan IG

Table 2. Hasil klasifikasi tanpa IG dan stopword removal

Classifier	N-grams	Akurasi rata-rata dari set validasi	Tanpa Stopword
SVM	Unigram(%)	86.8	87.4
	Bigram(%)	77.5	82.5
Naive bayes	Unigram(%)	85.1	85.2
	Bigram(%)	80	83

Tabel 2 menunjukkan akurasi rata-rata dengan menggunakan 5-FCV dan 3000 feature TFIDF, pada klasifikasi Support Vector Machine nilai akurasi unigam dan bigram didapatkan 86.7% dan 76.2% secara berurutan. Dan untuk klasifikasi Naïve bayes didapatkan nilai akurasi 84.9% dan 78.8% untuk unigram dan bigram. Nilai akurasi Unigram pada SVM lebih unggul 1.7% daripada Naïve bayes dan nilai akurasi Bigram pada Naïve bayes lebih unggul 2.6% daripada SVM. Penggunaan stopword removal pada tahap klasifikasi mempengaruhi penurunan nilai akurasi pada klasifikasi bigram, nilai akurasi yang di dapat tanpa stopword removal yaitu 82.5% untuk SVM dan 83% untuk NB. Tabel 3 merupakan hasil fitur seleksi yang akan digunakan untuk klasifikasi pada skenario selanjutnya.

Table 3 Hasil klasifikasi IG dengan threshold 0,01

Classifier	N-grams	Akurasi rata-rata dari set validasi
SVM	Unigram(%)	86.1
	Bigram(%)	76.2
Naive bayes	Unigram(%)	84.4
	Bigram(%)	77.7

Fitur yang di seleksi oleh IG dengan threshold 0.01 untuk unigram adalah 1382 dan bigram 1528 setelah itu fitur tersebut akan digunakan pada tahap klasifikasi. Hasil akurasi rata-rata dari 5-FCV unigram dan bigram untuk klasifikasi SVM 86,1% dan 76,2%. Rata-rata akurasi unigram dan bigram pada klasifikasi NB yaitu 84,4% dan 77,7%. Berdasarkan hasil tersebut terdapat penurunan akurasi rata-rata untuk SVM+unigram 0,7%, SVM+bigram 1,3%, NB+unigram 0,7% dan NB+bigram 2,3%. Penurunan nilai akurasi tersebut terjadi karena IG hanya memilih fitur yang memiliki nilai melebihi threshold sehingga dapat menyebabkan hilangnya beberapa informasi yang dibutuhkan untuk menentukan sentimennya.

B. Hasil Pengujian klasifikasi IG dengan stopword removal

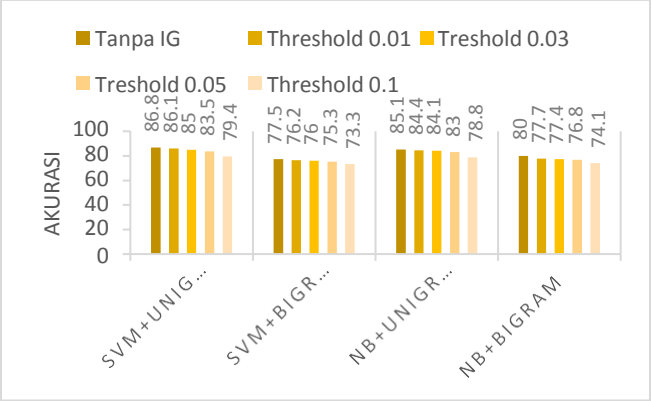
Table 4 Perbandingan jumlah fitur yang diseleksi oleh IG

	N-grams	Tanpa IG	Threshold 0.01	Threshold 0.03	Threshold 0.05	Threshold 0.1
Dengan stopword removal	Unigram	3000	1382	688	285	40
	Bigram	3000	1528	1332	1144	730
Tanpa stopword removal	Unigram	3000	1514	1051	659	150
	Bigram	3000	1543	1244	974	382

Jumlah Fitur

Jumlah fitur yang digunakan pada model klasifikasi dapat mempengaruhi nilai akurasi, pada skenario kedua fitur yang di dapat pada Tabel 4 akan digunakan untuk mencari nilai akurasi pada klasifikasi SVM dan NB sesuai dengan

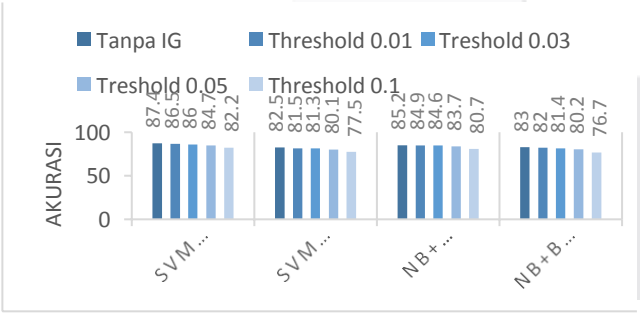
threshold yang digunakan. dapat dilihat bahwa fitur seleksi IG dapat mereduksi jumlah fitur sebanyak ±55% dari fitur awal yang digunakan pada threshold 0.01 dan ±99% untuk threshold 0,1 pada unigram. Sementara bigram tidak mengalami pengurangan fitur yang signifikan, artinya hanya sedikit fitur yang tidak relevan diseleksi pada setiap thresholdnya.



Gambar 2 : Perbandingan akurasi terhadap nilai threshold pada klasifikasi SVM dan Naive bayes menggunakan stopword removal

Gambar 2 merupakan hasil akurasi yang didapat dari klasifikasi SVM dan NB dengan nilai threshold yang digunakan pada fitur seleksi IG, semakin tinggi nilai threshold yang digunakan maka akan semakin banyak fitur yang di reduksi. Hasil akurasi tertinggi dari SVM+unigram yaitu 86,1%, SVM+bigram 76,2% pada threshold 0.01. Untuk NB+unigram didapatkan hasil tertinggi 84,4%, dan NB+bigram 77,7% pada threshold 0.01. Klasifikasi SVM mengungguli kinerja NB pada unigram sementara NB unggul pada bigram.

C. Hasil pengujian klasifikasi IG tanpa stopword removal



Gambar 3 : Perbandingan akurasi terhadap nilai threshold pada SVM dan NB tanpa stopword removal

Percobaan berikutnya terkait pengaruh penggunaan stopword removal pada tahap preprocessing, pada Tabel 4 terlihat peningkatan nilai akurasi untuk klasifikasi SVM dan NB+bigram oleh karna itu Gambar 3 menampilkan hasil dari klasifikasi tanpa penggunaan stopword removal untuk di Analisa lebih lanjut. Hasil klasifikasi SVM didapat nilai akurasi 87,4% untuk unigram dan 82,5% untuk bigram. Klasifikasi NB didapat nilai akurasi 85.2% untuk bigram dan 83% untuk bigram. Jika dibandingkan dengan nilai akurasi dari klasifikasi dengan penggunaan stopword removal, akurasi dari klasifikasi SVM+unigram naik sebesar 0,6% dan SVM+bigram naik sebesar 5%, NB+unigram naik sebesar 0,1% dan NB+bigram naik sebesar 3%. Peningkatan nilai akurasi pada klasifikasi bigram terjadi karena penggunaan stopword removal dapat mereduksi kombinasi kata bigram yang memiliki nilai TF-IDF dan IG yang besar untuk digunakan pada metode klasifikasi.

V. KESIMPULAN

Berdasarkan hasil skenario penelitian yang telah dilakukan dan dianalisa, akurasi tertinggi yang di dapatkan pada penelitian ini yaitu 86,8% pada klasifikasi SVM+unigram tanpa IG dengan stopword removal. Sementara klasifikasi SVM+unigram dengan IG dan stopword removal mendapatkan akurasi tertinggi sebesar 86.1% selisih 0,7%. Klasifikasi NB+bigram tanpa IG dengan stopword removal mendapat akurasi tertinggi 80% dan klasifikasi NB+Bigram dengan IG dan stopword removal mendapatkan nilai akurasi 77,7%. Hal tersebut membuktikan bahwa penggunaan IG tidak memastikan peningkatan kinerja dari klasifikasi menjadi lebih baik. Namun perlu diketahui bahwa fitur awal dibatasi sebanyak 3000 fitur dan proses fitur seleksi dilakukan setelah pembobotan fitur TF-IDF, sehingga fitur seleksi IG (dengan threshold 0.01, 0.03, 0.05, dan 0.1) dapat mereduksi fitur yang seharusnya digunakan pada proses klasifikasi.

Saran untuk penelitian selanjutnya, sebaiknya fitur awal pada tahap pembobotan TF-IDF dicari dengan menggunakan metode sehingga pada saat melakukan fitur seleksi dapat terlihat jelas perbandingannya jika menggunakan dataset yang besar. Dan mendapatkan threshold pada seleksi fitur dengan menggunakan metode untuk mendapatkan hasil seleksi fitur untuk di ujikan pada tahap klasifikasinya.

REFERENSI

[1] B. Agarwal and N. Mittal, “Sentiment Classification using Rough Set based Hybrid Feature Selection,” *Proc. 4th Work. Comput. Approaches to Subj. Sentim. Soc. Media Anal. (WASSA 2013)*, 2013.

[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] R. Maulana, P. A. Rahayuningsih, W. Irmayani, D. Saputra, and W. E. Jayanti, “Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain,” *J. Phys. Conf. Ser.*, vol. 1641, no. 1, pp. 0–6, 2020, doi: 10.1088/1742-6596/1641/1/012060.

[4] T. Yu and K. T. Nwet, “Sentiment analysis system for myanmar news using k nearest neighbor and naïve bayes,” 2020, doi: 10.18178/wcse.2020.02.001.

[5] N. Octaviani Faomasi Daeli, “Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor,” *J. Data Sci. Its Appl.*, 2020.

[6] T. O’Keefe and I. Koprinska, “Feature selection and weighting methods in sentiment analysis,” *ADCS 2009 - Proc. Fourteenth Australas. Doc. Comput. Symp.*, 2009.

[7] J. Brank, M. Grobelnik, N. Milić-Frayling, and D. Mladenić, "Feature selection using support vector machines," 2002, doi: 10.1142/9789812794710_0004.

[8] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Trans. Comput.*, 2005, doi: 10.11499/sicejl1962.38.456.

[9] A. Sharma and S. Dey, "A comparative study of selection and machine learning techniques for sentiment analysis," 2012, doi: 10.1145/2401603.2401605.

[10] K. Sarvabhotla, P. Pingali, and V. Varma, "Sentiment classification: A lexical similarity based approach for extracting subjectivity in documents," *Information Retrieval*. 2011, doi: 10.1007/s10791-010-9161-5.

[11] R. M. Elawady, S. Barakat, and N. M. Elrashidy, "Different Feature Selection for Sentiment Classification," *Int. J. Inf. Sci. Intell. Syst.*, 2014.

[12] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, 2016, doi: 10.1016/j.eswa.2016.03.028.

[13] S. Bird, "NLTK: The natural language toolkit," *COLING/ACL 2006 - 21st Int. Conf. Comput. Linguist. 44th Annu. Meet. Assoc. Comput. Linguist. Proc. Interact. Present. Sess.*, pp. 69–72, 2006, doi: 10.3115/1225403.1225421.

[14] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[15] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach," *Environ. Health Perspect.*, vol. 127, no. 9, pp. 2825–2830, 2019, doi: 10.1289/EHP4713.

[16] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, vol. 1, no. Scipy, pp. 56–61, 2010, doi: 10.25080/majora-92bf1922-00a.

[17] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.