

PERBANDINGAN METODE SUPPORT VECTOR MACHINE DENGAN METODE LEXICON DALAM ANALISIS SENTIMEN BAHASA INDONESIA

Muhammad Resa Arif Yudianto¹, Arham Rahim², Pristi Sukmasetya³, Rofi Abul Hasani⁴

^{1,3,4} Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Magelang

Jl. Mayjen Bambang Soegeng, Mertoyudan, Magelang

² Program Studi Teknologi Pendidikan, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Madako Tolitoli

Jl. Madako No.1, Toli-Toli, Sulawesi Tengah

resamuhammad96@unimma.ac.id, arham.rahim@umada.ac.id, pristi.sukmasetya@ummgl.ac.id, rofiabul@unimma.ac.id

Abstract - Sharing information in the current era is more accessible to social media; a popular one is Twitter. This media is specifically designed to express opinions and express one's feelings with a limited number of 280 characters. The uploads in the media have contents that describe the problems/feelings of someone who contains hidden knowledge. Therefore to find out the meaning of these sentences, sentiment analysis must be done. Several methods can be used to analyze a sentence's sentiments, such as a lexicon-based approach and the knowledge-based approach with the Support Vector Machine (SVM) algorithm. Both of these methods have different working principles in conducting sentiment analysis. This study compares the ability of SVM and lexicon-based approaches using the Indonesian sentiment dataset. From several experiments scenarios of 4000 datasets, it was found that the SVM method was superior in classifying positive and negative sentiments with an accuracy of 98.5% with unigram feature extraction parameters with a dataset ratio of 80:20. The lexicon-based approach is not good enough in conducting sentiment analysis with the highest accuracy of 78.43%. That is due to the lack of positive word dictionaries whose number is half of the number of negative word dictionaries, so words that have positive values cannot be appropriately recognized. The word dictionary that has a value in each word is more accurate in analyzing sentiments than the word dictionary with no value/score.

Abstrak - Berbagi informasi di era saat ini semakin mudah dengan adanya media sosial, salah satu yang populer adalah twitter. Media ini memang didesain khusus untuk mengutarakan pendapat dan mengekspresikan perasaan seseorang dengan jumlah karakter yang terbatas yaitu 280 karakter. Unggahan-unggahan dalam media tersebut memiliki isi yang menggambarkan permasalahan/perasaan seseorang yang mengandung pengetahuan yang tersembunyi. Oleh karena itu untuk mengetahui makna dari kalimat-kalimat tersebut harus dilakukan analisis sentimen. Terdapat beberapa metode yang dapat digunakan untuk menganalisis sentimen suatu kalimat seperti pendekatan lexicon based dan pendekatan knowledge based dengan algoritma Support Vector Machine (SVM). Kedua metode tersebut memiliki prinsip kerja yang berbeda dalam melakukan sentimen analisis. Penelitian ini membandingkan kemampuan dari pendekatan metode SVM dan lexicon based menggunakan dataset sentimen berbahasa Indonesia. Dari beberapa skenario percobaan terhadap 4000 dataset, didapatkan bahwa metode SVM lebih diunggulkan dalam mengklasifikasikan sentimen positif dan negatif dengan akurasi sebesar 98,5% dengan parameter ekstraksi fitur unigram dengan rasio dataset 80:20. Pendekatan berbasis lexicon based kurang baik dalam melakukan analisis sentimen dengan akurasi tertinggi sebesar 78,43%. Hal tersebut dikarenakan minimnya kamus kata positif yang jumlahnya adalah setengah dari jumlah kamus kata negatif, sehingga kata yang bernilai positif tidak dapat dikenali dengan baik. Kamus kata yang memiliki nilai pada setiap kata lebih akurat dalam menganalisis sentimen dibandingkan dengan kamus kata yang tidak memiliki nilai/skor.

Kata Kunci - Lexicon Based, Support Vector Machine, Analisis Sentimen, Big Data.

I. PENDAHULUAN

Semakin berkembangnya teknologi dan informasi memudahkan orang untuk saling berkomunikasi dan mengekspresikan apa yang dirasakan melalui media sosial. Salah satu media sosial yang populer dan banyak digunakan adalah twitter, media sosial ini memang didesain khusus untuk mengutarakan pendapat (*opinion*) dan gagasan

seseorang dengan jumlah karakter maksimal 280 karakter dalam 1 unggahan. Karena terbatasnya jumlah karakter yang disediakan oleh twitter membuat unggahan-unggahan yang dibuat di twitter sangat ekspresif dan menggambarkan inti dari permasalahan. Oleh sebab itu konten-konten yang ada di twitter mengandung arti/sentimen positif dan negatif yang menarik untuk dianalisis melalui proses penambangan teks.

Penambahan teks pada twitter, telah dilakukan oleh [1] untuk mengetahui sentimen masyarakat terhadap produk tertentu. Penelitian ini, menguji dan menganalisa kinerjanya algoritma SVM dalam melakukan klasifikasi. Penelitian lain, [2] juga menganalisis sentimen masyarakat dan menyimpulkan bahwa algoritma SVM memiliki nilai akurasi yang lebih tinggi dari algoritma *Recurrent Neural Network* (RNN).

Selain menggunakan algoritma *machine learning*, analisis sentimen juga dapat dilakukan dengan pendekatan berbasis kamus (*lexicon based*) seperti beberapa penelitian sebelumnya [3][4][5]. Penelitian yang dilakukan oleh [5] melakukan analisis sentimen terdapat teks pada media sosial dengan membandingkan pendekatan *lexicon based* dan pendekatan metode klasifikasi. Penelitian ini menyimpulkan bahwa metode klasifikasi kurang efektif dalam mengetahui sentimen yang berasal dari media sosial karena dataset yang tersedia terdapat banyak *noise*, tidak terstruktur dan memiliki banyak kata yang tidak baku. Pendekatan *lexicon* lebih efektif dalam melakukan analisis sentimen pada media sosial, karena berbasis kata per kata yang dicocokkan dengan kamus kata.

Penelitian sebelumnya tentang analisis sentimen dengan pendekatan *machine learning* dan *lexicon based* akan dijadikan sebagai acuan pada penelitian ini. Dalam penelitiannya [6] bertujuan untuk melakukan analisis sentimen dari *tweet* bahasa arab mesir dengan mengkombinasikan *semantic orientation* (*Cosine similarity and ISRI Arabic stemmer*) dengan *machine learning* (SVM dan NB), menyimpulkan, SVM memiliki tingkat akurasi yang lebih tinggi dengan 92,98% dibandingkan Naïve Bayes (NB) 62,41% menggunakan 1520 data *training*.

Penelitian yang dilakukan [1] bertujuan untuk menguji dan menganalisis kinerja dari SVM dalam melakukan analisis sentimen masyarakat terhadap mobil yang dapat mengendarai otomatis dan produk *apple*. Penelitian ini menyimpulkan kinerja algoritma SVM sangat bergantung terhadap jumlah dataset yang digunakan sebagai masukan dalam proses pembelajaran serta jumlah data pada masing-masing kelas dari dua jenis dataset SVM menghasilkan nilai akurasi yang berbeda di mana dataset 1 menghasilkan nilai akurasi 59,91% sedangkan dataset 2 menghasilkan nilai akurasi 71,2%.

Dalam penelitiannya [2] dilakukan analisis sentimen masyarakat terhadap hotel melalui teks dari hasil *review*. Penelitian ini, membandingkan akurasi algoritma RNN dengan SVM, menyimpulkan bahwa akurasi algoritma SVM lebih unggul dengan nilai akurasi 93,4%, 89,8% dan 95,4% dibanding algoritma RNN dengan nilai akurasi 48%, 48% dan 87%. Penelitian yang lain [3] yang bertujuan untuk menganalisis sentimen terhadap penggemar kegiatan olahraga (turnamen Cricket) dari data twitter menggunakan *lexicon based*, menyimpulkan NRC

Emotion Lexicon dapat menunjukkan perubahan sentimen terhadap pertandingan dan pemain dari 24 November - 24 desember 2017.

Penelitian yang dilakukan oleh [7] menganalisis sentimen terhadap program televisi dari komentar twitter dengan mengkombinasikan *lexicon based* dan SVM. Hasil penelitian menunjukkan bahwa metode kombinasi ini dapat diimplementasikan dalam menganalisis sentimen pada program televisi dengan tingkat akurasi yang mencapai 80%. Selanjutnya peneliti [8] membandingkan performa kombinasi *machine learning* dan *lexicon based* dengan metode sentimen analisis yang lain, disimpulkan bahwa kombinasi antara *machine learning* dan *lexicon based* memberikan hasil yang paling baik dibandingkan dengan pendekatan analisis yang lain seperti *Alyien Text Api*, *Alchemy Language Api* dan *Text Analytics Api*.

Berdasarkan dua jenis pendekatan di atas, penelitian ini akan membandingkan pendekatan *machine learning* menggunakan algoritma SVM dan pendekatan *lexicon based* pada dataset bahasa Indonesia. Proses perbandingan dilakukan dengan menganalisis sentimen positif dan negatif dari twitter berbahasa Indonesia yang diperoleh dari [9]. Dataset diperoleh dalam format .txt yang terdiri dari 2000 *tweet* dengan label negatif dan 2000 *tweet* dengan label positif. Dataset telah melalui proses pelabelan sehingga diharapkan dapat memberikan kemudahan dalam proses pengukuran.

II. METODE PENELITIAN

Penelitian ini akan membandingkan pendekatan dengan algoritma SVM dan pendekatan berbasis kamus (*lexicon based*) terhadap dataset yang diperoleh melalui github untuk dilakukan analisis sentimen.

A. Akuisisi Data

1. Pengumpulan Data

Dataset yang digunakan pada penelitian ini dari sumber internet yaitu *github* yang sebelumnya telah digunakan dalam penelitian serupa [9]. Dataset ini berjumlah 4000 yang terbagi sama rata menjadi dua sentimen yaitu 2000 baris data dengan label positif dan 2000 data dengan label negatif. Pada pendekatan berbasis kamus, digunakan kamus kata berbahasa Indonesia yang didapatkan dari <https://github.com/fajri91/InSet> dengan jumlah 10216 kata yang terbagi menjadi 6607 kata negatif dan 3609 kata positif.

2. Persiapan Data

Dataset sentimen yang didapatkan telah dilakukan *preprocessing* dan telah diberikan label oleh peneliti sebelumnya [9] mulai dari proses *stopping word* dan *steeming*. Pada kamus data didapatkan kamus yang memiliki nilai/skor 1-5 untuk kamus kata positif dan -1 hingga -5 untuk

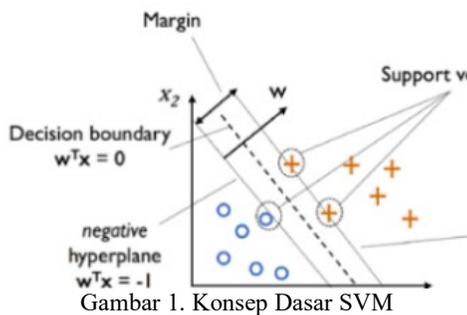
kamus kata negatif. Persiapan data pada kamus data berbahasa Indonesia ini diperlukan proses konversi file dari format .tsv menjadi .csv untuk digunakan pada proses selanjutnya.

B. Text Preprocessing

Proses perlakuan terhadap data sebelum diolah pada pembelajaran mesin termasuk di dalamnya proses pembersihan dan penormalan data untuk memudahkan proses selanjutnya [10]. Beberapa tahapan *preprocessing* pada penelitian [9] diantaranya adalah *retweet removal*, *text cleansing*, *hashtag handling*, *spell correction*, *negation handling*, *stopword removal* dan pelabelan secara otomatis menggunakan metode *lexicon* sehingga didapatkan dataset yang digunakan pada penelitian ini. Adapun *preprocessing* yang dilakukan pada penelitian ini adalah pembersihan terhadap tanda baca, *stopword removal*, *steeming* menggunakan *library* python sastrawi dan penghilangan fitur tag yang merupakan hasil dari proses *hashtag handling* pada penelitian sebelumnya. Karena pada penelitian ini tidak berfokus terhadap fitur tag yang terdapat dalam kalimat sentimen.

C. Support Vector Machine

Penelitian ini memilih algoritma SVM ini sebagai algoritma klasifikasi pada pendekatan pembelajaran mesin pada analisis sentimen ini. Algoritma ini termasuk ke dalam algoritma *supervised learning* yang pertama kali dikenalkan oleh [11]. Algoritma ini didesain khusus untuk menangani klasifikasi 2 kelas, akan tetapi telah dikembangkan juga untuk kasus multi kelas.



Gambar 1. Konsep Dasar SVM

Hyperplane yang ditemukan SVM diilustrasikan seperti Gambar 1 posisinya berada di tengah-tengah antara dua kelas, artinya jarak antara *hyperplane* dengan objek-objek data berbeda dengan kelas yang berdekatan (terluar) yang diberi tanda bulat kosong dan positif. Dalam SVM objek data terluar yang paling dekat dengan *hyperplane* disebut *support vector*. Objek yang disebut *support vector* paling sulit diklasifikasikan dikarenakan posisi yang hampir tumpang tindih (*overlap*) dengan kelas lain.

$$\tau(w) = \frac{1}{2} \|\vec{w}\|^2 \tag{1}$$

$$\frac{1}{2} \|\vec{w}\|^2 = \frac{1}{2} (w_1^2 + w_2^2) \tag{2}$$

Dengan syarat :

$$y_i(x_i \cdot w + b) - 1 \geq 0, i = 1,2,3,\dots,n$$

$$y_i(x_1 \cdot w_1 + x_2 \cdot w_2 + b) \geq 1$$

Rumus perhitungan algoritma SVM dalam mengklasifikasikan objek ke dalam kelas-kelasnya ditunjukkan pada persamaan (1) dan (2). Pada penelitian ini nilai masukan berupa teks/kalimat sentimen dengan label 1 untuk sentimen positif dan label 0 untuk sentimen negatif. Kalimat ini telah dikonversi ke bentuk numerik agar dapat diproses menggunakan rumus perhitungan SVM seperti pada persamaan (1) dan (2) hingga menghasilkan keluaran berupa nilai 1 atau 0.

D. Lexicon Based Method

Pendekatan berbasis kamus (*lexicon based approach*) merupakan metode pengklasifikasian kalimat ke dalam sentimen positif atau negatif. Konsep dasar dari metode ini adalah penggunaan kamus kata positif dan negatif sebagai bahan pembanding dan perhitungan untuk menentukan tingkat polaritas suatu kalimat untuk ditentukan sebagai sentimen negatif atau positif [12]. Dalam berbagai kamus kata *lexicon* ada 2 jenis kamus yang digunakan, yaitu kamus kata tanpa adanya nilai polaritas dari setiap kata, dan kamus kata dengan nilai/skor dari tiap kata tergantung dari tingkat polaritasnya. Pada penelitian ini peneliti menggunakan *lexicon* berbasis kamus (*dictionary*) yang terdiri dari 10216 kata positif dan negatif yang diperoleh dari <https://github.com/fajri91/InSet>.

Kamus leksikon sentimen bahasa Indonesia ini dibuat pada penelitian sebelumnya oleh [13]. Pada penelitian tersebut, dilakukan beberapa tahapan seperti top 5000 n-gram, melakukan pembobotan secara manual oleh 2 pakar, proses *steeming*, penambahan kata sinonim, dan pembersihan data. Pembobotan kata berada pada rentang -5 hingga 5 untuk kata negatif dan positif. Pada penelitian [13], kamus sentimen ini diuji dan dibandingkan dengan kamus sentimen yang lain dan didapatkan bahwa kamus sentimen leksikon ini mendapatkan akurasi tertinggi sebesar 65,78% karena pada kamus ini selain kata-kata normal juga terdapat kata-kata informal.

Oleh karena itu, penelitian ini menggunakan leksikon sentimen ini karena berdasarkan pengujian pada penelitian sebelumnya, memperoleh hasil yang terbaik. Pada penelitian ini dilakukan analisis sentimen menggunakan kamus kata yang memiliki skor maupun kamus kata sederhana dengan menghilangkan skor kata pada kamus kata. Skenario ini dilakukan untuk mengetahui hasil akurasi terbaik pada proses klasifikasi kalimat sentimen dari kedua kamus kata yang digunakan.

E. Evaluasi Model

Pengujian model hasil proses pembelajaran terbimbing (*supervised learning*) dapat dilakukan

dengan *confusion matrix*. Pengujian ini [14] memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh model dengan hasil klasifikasi yang sesungguhnya. Dalam metode *confusion matrix* ini terdapat beberapa variabel pengujian, yaitu akurasi, presisi, *recall* dan *f1-score*.

Untuk dapat menentukan nilai akurasi, presisi dan *recall* diperlukan bantuan tabel hasil prediksi seperti pada Tabel 1.

Tabel 1. *Confusion Matrix*

		Prediction Class	
		Positive	Negative
True Class	Positive	TP	FN
	Negative	FP	TN

Nilai *True Negative* (TN) adalah data yang diklasifikasi dengan tepat sebagai keluaran negatif atau salah. *True Positive* (TP) adalah data yang diklasifikasi dengan tepat sebagai keluaran positif atau benar. *False Positive* (FP) adalah data yang diklasifikasi dengan kurang tepat apabila keluaran berupa positif atau benar. *False Negative* (FN) adalah data yang diklasifikasi dengan kurang tepat [15].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100\% \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (6)$$

Pengukuran akurasi melakukan pengujian seberapa akurat model dapat mengklasifikasikan dengan benar seperti pada persamaan (3). Presisi menggambarkan rasio prediksi benar positif terhadap keseluruhan hasil yang diprediksi sebagai positif, yang dapat dicari dengan persamaan (4). Kemudian untuk variabel *recall* merepresentasikan keberhasilan model dalam menemukan kembali informasi seperti pada persamaan (5). Antara hasil presisi dan *recall* akan menghasilkan nilai *f1-score* yang merupakan perbandingan presisi dengan nilai *recall*. Penilaian *f1-score* ini dapat dicari melalui persamaan (6) [16].

F. Skenario Percobaan

Pada penelitian ini dilakukan analisis sentimen menggunakan dua pendekatan yang akan dibandingkan tingkat kemampuan dalam mengklasifikasikan kalimat sentimen positif dan negatif. Dalam melakukan perbandingan kedua pendekatan ini digunakan beberapa skenario pengujian.

Tabel 2. Skenario percobaan SVM

Skenario	Alias
70:30 => TF-IDF, max_df = 0.5, min_df = 5	S1

70:30 => TF-IDF, max_df = 0.6, min_df = 5	S2
70:30 => TF-IDF, max_df = 0.7, min_df = 5	S3
80:20 => TF-IDF, max_df = 0.5, min_df = 5	S4
80:20 => TF-IDF, max_df = 0.6, min_df = 5	S5
80:20 => TF-IDF, max_df = 0.7, min_df = 5	S6

Tabel 2 merepresentasikan beberapa skenario percobaan melalui pendekatan metode SVM yang terdiri dari skenario pembagian jumlah dataset sebagai data latih dan data uji, serta parameter filtering data. Sebagai contoh max_df = 0.5 akan mengabaikan kata/istilah yang muncul lebih dari 50% dalam keseluruhan dokumen, kemudian min_df = 5 yang berarti akan mengabaikan kata/istilah yang muncul kurang dari 5 kata dalam keseluruhan dokumen. Skenario tersebut akan dikombinasikan dengan N-gram mulai dari *unigram*, *bigram* dan *trigram*. Skenario ini dipilih karena pada dataset terdapat kalimat sentimen yang memiliki kecenderungan pengulangan kata bahkan kalimat yang sama.

Tabel 3. Skenario percobaan *Lexicon Based*

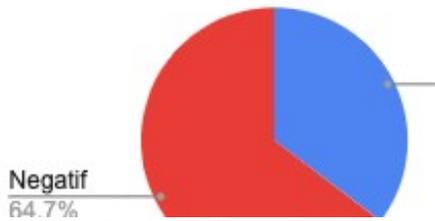
Skenario	Alias
Kamus data dengan poin	L1
Kamus data dengan poin tanpa steeming	L2
Kamus data dengan poin tanpa steeming dan stopword	L3
Kamus data tanpa poin	L4
Kamus data tanpa poin tanpa steeming	L5
Kamus data tanpa poin tanpa steeming dan stopword	L6

Pada Tabel 3 ditunjukkan skenario percobaan pada pendekatan metode *lexicon based*. Terdapat dua skenario utama yaitu pengujian menggunakan kamus yang memiliki poin dan kamus yang tidak memiliki poin pada setiap kata nya. Skenario tersebut dipilih untuk mengetahui pengaruh penggunaan nilai/skor kata terhadap akurasi proses klasifikasi kalimat positif dan negatif. Selain itu, dari masing-masing skenario utama akan dikombinasikan dengan beberapa perlakuan pada proses *preprocessing* karena pada metode *lexicon based* ini sangat tergantung dengan kamus kata yang digunakan.

III. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset sebanyak 4000 data domain politik yang telah diberikan label positif sebanyak 2000 kalimat dan 2000 label negatif. Dataset tersebut telah dilakukan proses *preprocessing* oleh peneliti sebelumnya. Sedangkan dalam pengolahan kalimat sentimen menggunakan pendekatan *lexicon based* kamus kata berbahasa Indonesia yang digunakan sebanyak 10216 kata yang terbagi terbagi menjadi 6607 kata negatif dan 3609 kata positif seperti ditunjukkan pada Gambar 2.

Leksikon Sentimen Indone



Gambar 2. Kamus Data

A. Text Preprocessing

Dari dataset yang diperoleh ternyata terdapat banyak kalimat yang masih mengandung tanda baca (*illegal string*) dan tag yang digunakan untuk keperluan penelitian sebelumnya, sehingga diperlukan penghilangan tanda baca dan tag tersebut seperti ditunjukkan pada Tabel 4.

Tabel 4. Text preprocessing

Sebelum	satu putaran langsung jungkal apa ya ? makin tidak sehat nalar . penjara ahok \$\$ HT_NEG
Sesudah	satu putaran langsung jungkal apa ya makin tidak sehat nalar penjara ahok
Sebelum	solidaritas dan adil harus tegak ! tunjuk yang benar dan jujur pasti menang kami ahok \$\$ HT_POS
Sesudah	solidaritas dan adil harus tegak tunjuk yang benar dan jujur pasti menang kami ahok
Sebelum	satu putaran langsung jungkal apa ya ? makin tidak sehat nalar . penjara ahok \$\$ HT_NEG

B. Metode SVM

Setelah dataset dilakukan penghilangan tanda baca dan tag tahap berikutnya adalah dilakukan proses pembuatan model menggunakan algoritma SVM dan dilakukan evaluasi untuk mengetahui keunggulan dari model yang dibuat. Pada penelitian ini dilakukan beberapa skenario dalam pembuatan model klasifikasi menggunakan algoritma SVM, untuk lebih jelasnya akan dijelaskan secara lebih rinci di bawah ini.

Tabel 5. SVM dengan *Unigram*

skenario	Accuracy	Precision	Recall	F1-Score
S1	0.9458	0.9518	0.9484	0.9522
S2	0.9558	0.9558	0.9587	0.9559
S3	0.9575	0.9574	0.9574	0.9574
S4	0.9850	0.9847	0.9855	0.9850
S5	0.9850	0.9847	0.9855	0.9850
S6	0.9850	0.9847	0.9855	0.9850

Pada Tabel 5 merupakan hasil evaluasi terhadap fitur ekstraksi unigram. Dari beberapa skenario, didapatkan akurasi terbaik dengan rasio dataset 80:20 pada seluruh kondisi pada parameter max_df. Akurasi

terburuk pada skenario 70:30 dengan mengabaikan istilah yang muncul lebih dari 50% dari keseluruhan dokumen.

Tabel 6. SVM dengan *bigram*

skenario	Accuracy	Precision	Recall	F1-Score
S1	0.9717	0.9716	0.9721	0.9975
S2	0.9975	0.9975	0.9976	0.9975
S3	0.9975	0.9975	0.9976	0.9975
S4	0.9988	0.9987	0.9987	0.9988
S5	0.9988	0.9987	0.9987	0.9988
S6	0.9988	0.9987	0.9987	0.9988

Hasil evaluasi terhadap fitur ekstraksi *bigram* ditunjukkan pada Tabel 6. Didapatkan hasil akurasi terbaik dengan rasio dataset 80:20 dengan mengabaikan istilah yang muncul lebih dari 50% pada keseluruhan dokumen, dan akurasi terburuk pada skenario 70:30 dengan pengabaian istilah yang muncul di atas 50%.

Tabel 7. SVM dengan *trigram*

skenario	Accuracy	Precision	Recall	F1-Score
S1	0.9725	0.9724	0.9729	0.9725
S2	0.9975	0.9975	0.9976	0.9975
S3	0.9975	0.9975	0.9976	0.9975
S4	0.9988	0.9988	0.9987	0.9988
S5	0.9988	0.9988	0.9987	0.9988
S6	0.9988	0.9988	0.9987	0.9988

Tabel 7 di atas merepresentasikan hasil evaluasi terhadap fitur ekstraksi trigram. Dari beberapa skenario yang dilakukan, didapatkan akurasi terbaik dengan rasio dataset 80:20 dengan mengabaikan istilah yang muncul lebih dari 50% pada keseluruhan dokumen, dan akurasi terburuk pada skenario 70:30 dengan pengabaian istilah yang muncul di atas 50%.

Dari ketiga parameter N-gram yang digunakan kemudian dilakukan pengujian model dengan 10 kalimat baru yang berbeda dengan data latih maupun data uji, yang terdiri dari 5 kalimat positif dan 5 kalimat negatif. Dari proses pengujian diperoleh bahwa skenario terbaik dengan kesalahan prediksi terkecil yaitu kombinasi antara rasio dataset 80:20 dengan fitur ekstraksi *unigram* yaitu dengan nilai akurasi 98,5%. Pada ekstraksi bigram dan trigram meskipun memiliki nilai akurasi yang tertinggi, akan tetapi saat dilakukan pengujian dengan data baru mengalami *overfitting* dengan melakukan kesalahan prediksi yang tinggi. Hasil akurasi pada penelitian ini lebih baik dibandingkan pada penelitian sebelumnya

[9], di mana pada penelitian tersebut hasil akurasi tertinggi adalah 95% dengan penambahan ekstraksi fitur hashtag pada kalimat menggunakan metode sentiHT.

C. Metode Lexicon Based

Dari beberapa skenario yang telah ditentukan seperti pada Tabel 3, digunakan 4000 dataset yang terdiri dari 2000 kalimat dengan label negatif, dan 2000 kalimat positif. Kamus kata yang digunakan sebanyak 10216 kata yang terbagi terbagi menjadi 6607 kata negatif dan 3609 kata positif. Kamus kata tersebut masing-masing kalimat memiliki nilai/skor mulai dari 1-5 untuk kata positif dan -1 sampai -5 untuk kata negatif. Percobaan dilakukan juga dengan menghilangkan nilai/skor yang terdapat pada kamus kata tersebut. Dari hasil percobaan didapatkan hasil prediksi yang ditunjukkan pada Tabel 8.

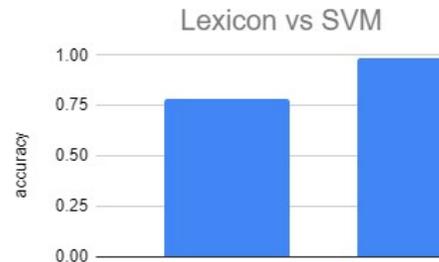
Tabel 8. Hasil prediksi sentimen dengan *lexicon*

Skenario	Predicted			Accuracy
	Positif	Negatif	Netral	
L1	1132	2561	307	0.7830
L2	1137	2557	306	0.7843
L3	796	2962	242	0.6990
L4	1069	2069	862	0.7673
L5	1069	2071	860	0.7673
L6	770	2613	617	0.6925

Pada Tabel 8 merepresentasikan hasil analisis sentimen menggunakan metode kamus. Dari hasil percobaan terhadap 6 skenario, diperoleh hasil bahwa penggunaan kamus kata yang memiliki nilai untuk setiap kata memberikan hasil yang lebih baik daripada kamus kata yang tidak memiliki nilai pada semua kondisi. Hasil terbaik diperoleh pada skenario tanpa proses steeming, karena pada kamus kata yang digunakan kebanyakan kata nya masih memiliki imbuhan. Metode *lexicon* lebih banyak memprediksi kalimat dengan hasil sentimen negatif dibanding sentimen positif, karena jumlah kamus kata negatif jumlahnya 2 kali lebih banyak dari kamus positif, sehingga kata-kata yang kemungkinan memiliki polaritas sentimen positif akan dikenali sebagai kamus netral karena kurangnya kamus kata positif.

D. Perbandingan SVM dan Lexicon Method

Berdasarkan hasil percobaan menggunakan algoritma SVM dan *lexicon based*, didapatkan hasil terbaik dari kedua pendekatan seperti pada Gambar 3.



Dari Gambar 3 didapatkan hasil bahwa pendekatan dengan algoritma SVM lebih unggul dalam penentuan sentimen kalimat berbahasa Indonesia dengan nilai akurasi 98,5%. Sedangkan pada pendekatan *lexicon based* hanya mencapai akurasi 78% dikarenakan kurangnya kamus kata positif. Kecenderungan metode *lexicon* terhadap kamus kata yang digunakan mengakibatkan kesalahan dalam melakukan klasifikasi kalimat sentimen, karena dataset yang digunakan merupakan dataset dengan domain politik sedangkan kamus kata yang digunakan adalah kamus kata bahasa Indonesia secara umum. Algoritma SVM dapat mencapai hasil yang memuaskan karena sangat ideal dan cocok untuk melakukan klasifikasi 2 kelas. Selain itu algoritma ini tidak memerlukan jumlah data yang banyak.

IV. KESIMPULAN

Temuan yang didapatkan dari penelitian ini adalah penggunaan *unigram* dalam proses ekstraksi fitur memiliki tingkat akurasi yang terbaik yaitu 98,5%. Model terbaik didapatkan menggunakan kombinasi antara rasio dataset 80:20 dan penggunaan *unigram*. Fitur *bigram* dan *trigram* menghasilkan akurasi yang tertinggi, akan tetapi setelah diuji dengan data baru tidak dapat memprediksi data tersebut dengan baik, model mengalami *overfitting*. Penggunaan kamus kata yang tidak seimbang antara jumlah kata negatif dan positif pada pendekatan *lexicon* mengakibatkan akurasi klasifikasi sentimen kalimat positif dan negatif yang rendah yaitu hanya didapatkan akurasi tertinggi 78,43%. Pada penelitian ini metode *lexicon* tidak dapat mengenali kalimat yang cenderung memiliki sentimen positif, karena kamus kata positif jumlahnya setengah dari kamus kata negatif.

Sehingga kalimat yang seharusnya diprediksi sebagai sentimen positif tetapi justru diprediksi menjadi sentimen netral karena kurangnya kamus kata positif. Penggunaan kamus kata yang memiliki nilai/skor pada setiap kata nya menghasilkan nilai akurasi yang lebih baik daripada kamus kata yang tidak menggunakan nilai/skor pada setiap kata. Metode SVM mendapatkan hasil akurasi yang jauh lebih baik dibandingkan dengan pendekatan berbasis *lexicon*, karena pada pendekatan *lexicon* terdapat banyak kesalahan prediksi pada kalimat sentimen positif yang tidak dapat dikenali oleh kamus kata positif karena

kurangnya kamus kata positif. Selain itu karena dataset yang digunakan memiliki domain politik, sedangkan kamus kata yang digunakan cenderung pada kamus kata umum.

DAFTAR PUSTAKA

- [1] AHMAD, M., AFTAB, S., AND ALI, I., 2017. Sentiment Analysis of Tweets using SVM. *Int. J. Comput. Appl.*, vol. 177, no. 5, hal. 25–29.
- [2] AL-SMADI, M., QAWASMEH, O., AL-AYYOUB, M., JARARWEH, Y., AND GUPTA, B., 2017. Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels Reviews. *J. Comput. Sci.*
- [3] AGARWAL, A. AND TOSHNIWAL, D., 2018. Application of Lexicon Based Approach in Sentiment Analysis for short Tweets. *Int. Conf. Adv. Comput. Commun. Eng.*, no. June, hal. 189–193.
- [4] AUNG, K. Z., AND MYO, N. N., 2017. Sentiment Analysis of Student's Comment Using Lexicon Based Approach. *IEEE/ACIS 16th Int. Conf. Comput. Inf. Sci.*, hal. 149–154.
- [5] AKTER, S. AND AZIZ, M. T., 2016. Sentiment Analysis On Facebook Group Using Lexicon Based Approach. *3rd Int. Conf. Electr. Eng. Inf. Commun. Technol.*, hal. 8–11.
- [6] ABUELENIN, S., ELMOUGY, S., AND NAGUIB, E., 2017. Twitter Sentiment Analysis for Arabic Tweets. in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics*, hal. 467–476.
- [7] TIARA, SABARIAH, M. K., AND EFFENDY, V., 2015. Sentiment Analysis on Twitter Using the Combination of Lexicon-Based and Support Vector Machine for Assessing the Performance of a Television Program. *3rd Int. Conf. Inf. Commun. Technol. Sentim.*, hal. 386–390.
- [8] NASIM, Z., RAJPUT, Q., AND HAIDER, S., 2017. Sentiment Analysis of Student Feedback Using Machine Learning and Lexicon Based Approaches. *Int. Conf. Res. Innov. Inf. Syst.*, hal. 1–6.
- [9] ALFINA, I., SIGMAWATY, D., NURHIDAYATI, F., AND A. HIDAYANTO, N., 2017. Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain. *ICMLC*, hal. 43–47.
- [10] ANANDARAJAN, M., HILL, C., AND NOLAN, T., 2019. Practical Text Analytics: Maximizing the Value of Text Data. *Springer*, vol. 2.
- [11] CORTES, C., AND VAPNIK, V., 1995. Support vector machine. *Machine learning*, 20(3), 273-297.
- [12] ALMATARNEH, S. AND GAMALLO, P., 2018. A lexicon based method to search for extreme opinions. *PLoS One*, vol. 13, no. 5, hal. 1–19.
- [13] KOTO, F., RAHMANINGTYAS, G.Y., 2018. Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, 2018-January(December), hal. 391–394. doi: 10.1109/IALP.2017.8300625.
- [14] SAMMUT, C., WEBB, G.I., 2011. Encyclopedia of Machine Learning, Springer.
- [15] SOKOLOVA, M., & LAPALME, G., 2009. Information Processing and Management. A systematic analysis of performance measures for classification tasks, 427-437.
- [16] PATIL, N. M. AND NEMADE, M. U., 2017. Music Genre Classification Using MFCC , K-NN and SVM Classifier. *Int. J. Comput. Appl.*, vol. 4, no. 2, hal. 43–47, 2017.