Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM

by Muhammad Fernanda Naufal Fathoni

Submission date: 01-Jul-2024 08:25PM (UTC+0700)
Submission ID: 2411204211
File name: MODEM_-_VOL.1,_NO.3_JULI_2024_HAL_62-76.pdf (405.12K)
Word count: 4371
Character count: 27093

Modem : Jurnal Informatika dan Sains Teknologi
Vol. 1 No. 3 Juli 2024
E-ISSN : 3046-7209 , P-ISSN : 3046-7217, Hal 62-76
DOI: https://doi.org/10.62951/modem.v1i3.112

Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM

Muhammad Fernanda Naufal Fathoni
Universitas Pembangunan Nasional Veteran Jawa Timur

Eva Yulia Puspaningrum
Universitas Pembangunan Nasional Veteran Jawa Timur

Andreas Nugroho Sihananto
Universitas Pembangunan Nasional Veteran Jawa Timur

Alamat: Jl. Rungkut Madya No.1, Gn. Anyar. Kec. Gn.Anyar, Surabaya, Jawa Timur 60294
Korespondensi penulis: evapuspaningrum.if@upnjatim.ac.id

Abstract. *Rohingya in Indonesia has become trending conversation on social media. Sentiment analysis can get public responds. Big data makes the problem time efficiency labeling process, therefore the lexicon dictionary is needed for the labeling process. Data is growing and circulating very rapidly so it takes a fast and efficient time. Although it is fast and makes it easier to solve problems, it is still necessary to question the accuracy produced when using the lexicon labeling. A comparison of the labeling process between the InSet lexicon and the VADER lexicon was conducted to determine the accuracy of the labeling. It was done by combining lexicon with machine learning method of support vector machine and TF-IDF weighting and accuracy result calculated using confusion marix. Data from social media X as many as 9117 lines and labeled with InSet lexicon result 5241 negative sentiments, 1369 positive, and 521 neutral. Then the labeling results with VADER produced 2749 positive, 2523 negative, and 1881 neutral. After labeled, processed SVM and calculated accuracy with results of InSet lexicon accuracy having an average of 85.8% while the VADER SVM lexicon has an average of 82.65%.*

Keywords: *Sentiment, rohingya, Lexicon, InSet VADER, SVM*

Abstrak. *Masuknya rohingya menjadi bahan pembincangan di sosial media. Cara melihat bagaimana respon masyarakat terhadap hal tersebut dibutuhkan analisa sentimen. Banyaknya data menjadikan masalah efisiensi waktu proses pelabelan. Maka dari itu dibutuhkan kamus lexicon untuk pelabelan tersebut. Adanya kamus lexicon menjadi lebih cepat dalam pengolahan data. Meskipun cepat dan mempermudah dalam penyelesaian masalah masih perlu dipertanyakan akurasi yang dihasilkan dengan kamus lexicon tersebut. Dilakukan perbandingan proses pelabelan antara lexicon InSet dan VADER untuk mengetahui manakah kamus lexicon yang memiliki akurasi lebih baik dari pelabelan tersebut. Hal itu dikombinasikan dengan machine learning support vector machine dan pembobotan TF-IDF lalu hasil akhir akan dievaluasi dengan menggunakan confusion matrix. Hasil yang didapatkan lexicon InSet melabeli 5241 sentimen negatif, 1369 sentimen positif, 521 sentimen netral. Lalu hasil pelabelan dengan VADER menghasilkan 2749 sentimen positif, 2523 sentimen negatif, dan 1881 sentimen netral. Setelah terlabeli dilakukan training data dan dihasilkan lexicon InSet dengan SVM didapatkan rata rata akurasi 87,83% dan VADER dengan akurasi rata rata 87,66%*

Kata kunci: *Sentimen, rohingya, lexicon, inset, vader, SVM*

LATAR BELAKANG

Pada desember 2023 terdapat kapal yang mengangkut 315 pengungsi etnis Rohingya di Kabupaten Aceh Besar dan Kabupaten Pidie Provinsi Aceh (Zulkarnaini, 2023). Munculnya hal tersebut menjadikan banyak timbulnya pro dan kontra terkait masuknya Rohingya di Indonesia, serta terciptanya berbagai opini yang beredar di Indonesia.

Received Juni 03, 2024; Accepted Juli 01, 2024; Published Juli 31, 2024
* Muhammad Fernanda Naufal Fathoni, evapuspaningrum.if@upnjatim.ac.id

Saat ini media sosial menjadi bagian dari masyarakat untuk menyampaikan tanggapan, keluhan serta opini dari berbagai peristiwa. Opini yang berkembang di media sosial sangat cepat mengenai masuknya Rohingya, salah satunya di media sosial X. X merupakan media sosial yang banyak digunakan masyarakat dalam mengeluarkan opini berupa teks. Oleh karena itu, banyak peneliti menggunakan data penelitian dari sosial media X tersebut(Giovani et al., 2020).

Analisa sentimen merupakan kegiatan melakukan analisa terkait sensasi, emosi, sentimen, opini, sudut pandang, sikap, yang dimiliki seseorang terhadap berbagai hal berkaitan dengan peristiwa, produk, layanan atau organisasi. Sentimen berupa opini publik, opini tersebut dapat berupa teks, video, dan audio. Analisa sentimen dapat mendeteksi bagaimana prespektif publik terhadap opini-opini yang beredar di masyarakat (D'Aniello et al., 2022). Padangan dan opini masyarakat di media sosial X dapat digunakan sebagai sumber daya dalam merangkum opini publik tersebut.

Sebelum dilakukan analisa sentimen terdapat proses pre-processing data, salah satunya pelabelan data. Proses pelabelan data secara menual membutuhkan proses yang lama, sehingga keterbatasan waktu, kemampuan energi serta biaya yang besar dalam pelabelan tersebut (Chan et al., 2023). Dengan adanya hal tersebut dibutuhkan proses labeling otomatis untuk memberikan label positif, netral atau negatif pada data. Metode yang dapat digunkaan dalam pelabelan otomatis tersebut adalah lexicon.(Biswas et al., 2023) Lexicon merupakan kamus untuk mencari bobot dari setiap kata pada kalimat dan menghitung gabungan kata lalu menentukan sentimen berdasarkan pembobotan kalimat yang telah dihitung (D'Aniello et al., 2022)

Penelitian sebelumnya dilakuakn penggabungan metode untuk mengefisiensi dari segi waktu dan biaya untuk pelabelan. Penelitian dilakukan oleh (Isnan et al., 2023) dengan menggabungkan pelabelan lexicon jenis VADER dengan SVM (Support Vector Machine) menghasilkan F1 sebesar 80%. Penelitian lain dilakukan (Muhammadi et al., 2022) dengan melakukan penggabungan lexicon InSet dan SVM menghasilkan akurasi sebesar 78.86%

Dengan adanya hal tersebut terjadi perbedaan akurasi antara pelabelan lexicon InSet dan VADER sehingga menjadikan peneliti tertarik membuat perbandingan kombinasi lexicon InSet dan VADER dengan SVM pada analisa sentimen Rohingya sosial media X untuk mengetahui manakah performa yang lebih baik.

KAJIAN TEORITIS

Analisa Sentimen

Analisa sentimen merupakan proses mengekstraksi opini dan emosi dari teks yang menghasilkan sebuah kesimpulan dari teks tersebut (Borg & Boldt, 2020). Sentimen dapat berupa opini publik, opini tersebut dapat berupa teks, video, audio. Analisa sentimen akan menghasilkan bagaimana perspektif publik terhadap peristiwa yang beredar di masyarakat apakah positif, negatif atau netral terhadap suatu hal (D'Aniello et al., 2022). Selain itu analisa sentimen dapat membantu pembuat kebijakan publik untuk merumuskan sebuah kebijakan terkait peristiwa yang terjadi (Gupta & Agrawal, 2020)& Agrawal, 2020). Terdapat pendekatan untuk melakukan analisa sentimen antara lain lexicon, machine learning dan hybrid (Yadav & Vishwakarma, 2020)

Machine Learning

Machine Learning adalah sebuah proses algoritma komputasi yang digunakan untuk membuat sebuah program dengan inputan lalu menghasilkan sebuah output tertentu (Musfiroh et al., 2021). Machine learning mampu melakukan komputasi yang dapat meniru kecerdasan manusia, proses dalam penerapan kecerdasan buatan machine learning menerapkan berbagai disiplin ilmu mulai dari probabilitas, statistika, ilmu komputer, dan psikologi (El et al., n.d.). Pengaplikasian machine learning dilakukan dalam beberapa disiplin seperti finance, biologi, computasi, entertain, marketing, dan banyak hal yang bisa dilakukan oleh machine learning. Machine learning akan mempelajari berdasarkan data yang diberikan sebelumnya kemudian ia akan mengingat pembelajaran tersebut untuk memprediksi inputan yang diberikan (El Naqa & Murphy, 2015). Machine learning dapat digunakan untuk pendekatan analisa sentimen dengan mengklasifikasikan teks(Wankhade et al., 2022).

Aplikasi X

Aplikasi X adalah sebuah media sosial yang sebelumnya bernama twitter. X merupakan sosial media yang penggunanya dapat membuat postingan dengan nama tweet. Selain itu X dapat mengomentari, menyukai dan membagikan tweet yang diposting orang lain sehingga mengakibatkan tweet mendapatkan respon banyak atas tweet tersebut (Karami et al., 2020). Saat ini X memiliki 336 juta pengguna dan menghasilkan 500 juta per hari (Geofany & Liza, n.d.). Indonesia memiliki jumlah pengguna X terbesar ke-3 di dunia dan dilaporkan dengan 77% pengguna aktif yang produktif dalam menuliskan tweet (Borg & Boldt, 2020)

Text Preprocessing

Teks preprocessing adalah sebuah tahapan dalam membersihkan text dari karaktek, kata-kata, dan tanda baca yang tidak diinginkan (Tabassum & Patil, 2020) Teks preprocessing merupakan tahapan yang penting dalam NLP (Natural Language Processing), untuk mengubah data mentah menjadi data yang lebih dipahami strukturnya mencakup keyword yang dapat membantu proses klasifikasi maupun klastering.

Lexicon

Lexicon adalah sebuah metode yang dapat digunakan untuk menganalisa sentimen dengan pendekatan kamus. Kamus lexicon dapat mengolah bobot kata berdasarkan kamus sehingga memudahkan dalam melakukan klasifikasi secara otomatis terhadap kalimat yang akan di klasifikasi. Maka dari itu, lexicon dapat digunakan sebagai labeling data latih sebelum dilakukan pelatihan model (Machová et al., 2020)

Support Vector Machine

SVM (Support Vector Machine) adalah sebuah fungsi untuk mencari hyperplane yang optimal untuk mengklasifikasi kelas dalam sebuah data. SVM termasuk machine learning dengan jenis supervised learning yang membutuhkan data terklasifikasi sebagai data latih. Hyperplane (bidang pemisah) tersebut akan menentukan label dari data yang akan diklasifikasi serta memaksimalkan margin dari jarak kelas yang berbeda. Setiap data diwakili sebagai titik dan memaksimalkan jarak/margin antara vektor dari kedua kelas (Muhammadi et al., 2022).

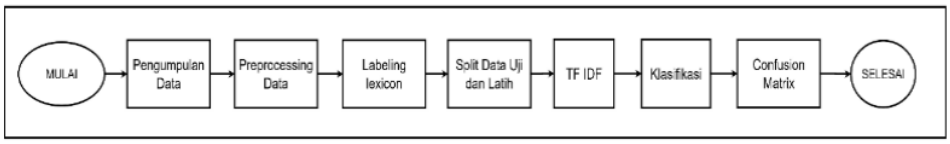
Confusion Matrix

Confusion matrix adalah tahapan mengukur akurasi dari sebuah model svm yang sudah dibuat. Dalam menggunakannya akan dilalkukan perhitungan berdasarkan data uji. Data uji yang sesuai prediksi dan tidak akan dilakukan perhitungan bagaimana presentase hasil data uji dengan data aktual.

METODE PENELITIAN

Perancangan sistem merupakan rancangan metode dari penyelesaian masalah. Dapat dilihat pada Gambar 1 merupakan perancangan sistem yang dilakukan mulai dari pengumpulan

data, preprocessing data, labeling lexicon, split data uji, TF-IDF, Klasifikasi dan confusion matrix.



Gambar 1

Pengumpulan Data

Pengumpulan data dilakukan dengan scraping data twitter menggunakan API. Data tersebut diambil selama 14 hari dan data yang didapatkan berupa data kotor sehingga diperlukan proses preprocessing data. Proses pengambilan tersebut diambil dengan kata kunci "rohingya" dan disimpan dengan bentuk csv.

Preprocessing Data

Setelah data terkumpul dilanjutkan proses preprocessing data. Tahapan ini merupakan tahapan untuk mengelola data mentah dari data yang telah dikumpulkan. Preprocessing dilakukan dengan menggabungkan data yang telah dikumpulkan. Setelah data digabung dilakukan data cleaning yang meliputi penghapusan data kosong, data duplikat, dan isi data yang tidak digunakan. Setelah cleaning, dilanjutkan dengan proses tokenizing yaitu proses memisahkan kalimat perkalimat, Setelah itu dilakukan normalisasi data untuk mengubah kata data yang tidak baku, dilanjutkan proses translate data karena pada bagian labeling lexicon VADER menggunakan bahasa Inggris. Setelah dilakukan translate dilakukan proses Filtering untuk membuang kata yang tidak dipelukan dalam kalimat, filtering dalam bahasa Inggris dan Indonesia berbeda karena perbedaan penggunaan kamus. Setelah itu dilakukan stemming data untuk mengubah kata menjadi kata dasar dan pengubahan kata menjadi kata dasar dalam bahasa Inggris dan Bahasa Indonesia berbeda. Ketika semua proses selesai dilakukan penyimpanan data yang sudah dipreprocessing dan dilanjutkan untuk labeling sentimen dengan kamus lexicon.

Labeling Lexicon

Labeling dilakukan setelah data telah dilakukan preprocessing, dan dilakukan labeling dengan lexicon based. Lexicon based adalah sebuah metode unutk menganalisa dengan menggunakan kamus lexicon. Kamus lexicon berisikan bobot dari sebuah kata yang digunakan

sebagai acuan untuk menghitung bobot dari sebuah kata perhitungan (D'Aniello et al., 2022). Proses pelabelan data menggunakan kamus lexicon InSet dan lexicon VADER.

Lexicon InSet adalah pengembangan kamus lexicon dengan bahasa indonesia dengan tujuan untuk menganalisa sentimen. Lexicon InSet dibuat pada tahun 2018 dengan menggunakan data dari X dengan isi kamus 3609 kata positif dan 6609 negatif. Pembobotan kata lexicon InSet diantara -5 sampai 5 yang diberikan secara manual (Koto & Rahma.Penentuan jenis sentimen dilakukan dengan total bobot pada sebuah kata, jika total bobot lebih dari 0 maka masuk kedalam label positif, jika total bobot kurang dari 0 maka masuk kedalam sentimen negatif, dan apabila bobot sama dengan 0 maka termasuk dalam sentimen netral. (Musfiroh et al., 2021)

Lexicon VADER merupakan sebuah kamus lexicon yang dibuat oleh CJ Hutto dan Eric Gilbert dari Institute Technologi Georgia. Lexicon VADER secara spesifik digunakan untuk menganalisis microblog dan penentuan sentimen ditentukan dengan bobot sebuah kalimat, apabila bobot tersebut lebih dari 0.05 akan masuk kedalam sentimen positif, jika total bobot kurang dari -0.05 masuk kedalam sentimen negatif, dan apabila bobot diantara -0.05 sampai 0.05 akan masuk kedalam sentimen netral (Hutto & Gilbert, 2014). Lexicon VADER dibangun dengan data bahasa Inggris sehingga dalam penggunaannya mengharuskan proses translate data terlebih dahulu kedalam bahasa Inggris. Pada data yang memiliki nilai netral akan dilakukan penghapusan, hal tersebut dikarenakan karena nilai netral dinilai tidak memihak apakah setuju atau tidak terdapat sentimen tersebut. (Arya et al., 2022)

TF-IDF dan Split Data Uji dan Latih

Setelah data sudah diberikan label dengan lexicon dilanjutkan dengan Transformasi TF-IDF. Sebelum dilakukan proses transformasi TF-IDF dilakukan pembagian data uji dan data latih (Muhammadi et al., 2022). Proses pembagian data uji dan data latih dengan komposisi data latih 70% data uji 30% dan data latih 80% dan data uji 20%. Setelah dilakukan pembagian dilakukan transformasi dan hasil dari pembobotan TF-IDF akan digunakan untuk membuat model machine learning.(Baiq Nurul Azmi et al., 2023)

Klasifikasi SVM

SVM adalah sebuah fungsi untuk mencari hyperplane (bidang pemisah) yang optimal untuk mengklasifikasi kelas dalam sebuah data. SVM termasuk dalam supervised learning dan membutuhkan data berlabel sebagai data latih. Hyperplane tersebut akan memisahkan kelas label untuk diklasifikasi. Pada penggunaan kernel linear digunakan nilai hyperparameter C

dengan 3 skenario yaitu nilai C = 0.1, C = 1, C = 10. Setelah dilakukan pelatihan model dilakukan pengujian menggunakan data uji yang telah dilakukan. Berikut adalah skenario yang akan dilakukan seperti pada Tabel 1

Tabel 1

Labeling	Hyperparameter	Data Latih	Data Uji
InSet Lexicon	C = 0,1; 1; 10	70%	30%
	C = 0,1; 1; 10	80%	20%
VADER Lexicon	C = 0,1; 1; 10	70%	30%
	C = 0,1; 1; 10	80%	20%

Confusion Matrix

Setelah itu dilakukan pengujian dengan data uji dan hasilnya akan dihitungan dengan confusion matrix. Contoh perhitungan confusion matrix untuk mengetahui akurasi dari model yang dibuat sebagai berikut :

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + TN + FN}, \text{ Dimana :}$$

True Positif (TP) : Jumlah data uji positif yang sesuai dengan klasifikasi positif

False Positif (FP) : Jumlah data uji negatif yang diklasifikasikan menjadi positif

True Negatif (TN) : Jumlah data uji negatif yang diklasifikasikan sesuai negatif

False Negatif (FN) : Jumlah data uji positif yang di klasifikasikan menjadi negatif

HASIL DAN PEMBAHASAN

Pengumpulan Data

Pengumpulan Data dilakukan dengan cara scraping data di aplikasi X. Pengumpulan tersebut menggunakan tweet harvest dan disimpan dalam bentuk CSV. Data yang dikumpulkan selama 14 hari diambil dari tanggal 29 Desember 2023-11 Januari 2024 Hasil dari pengumpulan data didapatkan jumlah 9117 baris dan 22 kolom. Pada Gambar 4.1 adalah hasil

pengumpulan data. Isi kolom tersebut diantaranya adalah tanggal tweet dibuat, isi tweet, id tweet dll. Setelah data terkumpul dilanjutkan proses preprocessing data.

Preprocessing Data

1. Penggabungan Data

Data yang terkumpul selama 14 hari digabungkan menjadi satu file CSV.

2. Data Cleaning

Pada data cleaning dilakukan pengecekan data kosong, data duplikat, dan penghapusan isi conten yang tidak digunakan. Pada data cleaning ditemukan 1285 data duplikat sehingga dilakukan penghapusan terhadap data tersebut. Selain itu pada data cleaning isi data yang tidak digunakan seperti URL, mention, hastag, mengubah huruf besar menjadi kecil dan tanda baca dihapus dengan menggunakan library regex. Pada Tabel 2 adalah hasil sebelum dan sesudah dilakukan cleaning

Tabel 2

Sebelum Data Cleaning	Sesudah Data Cleaning
@YurisAbrori @rgantas Kalo dari awal segampang itu dapet dana buat benerin fasilitas ga bakal kejadian kek gini. Nyatanya bukan cuma fasilitas tapi kriminalitas yg tinggi juga ga segampang itu diselesein pake uang. Coba deh baca, banyak kok di situs berita indo besar bahas topik kondisi rohingya	kalo dari awal segampang itu dapet dana buat benerin fasilitas ga bakal kejadian kek gini nyatanya bukan cuma fasilitas tapi kriminalitas yg tinggi juga ga segampang itu diselesein pake uang coba deh baca banyak kok di situs berita indo besar bahas topik kondisi rohingya

3. Tokenizing

Tokenizing dilakukan setelah data clenaing. Proses tersebut dilakukan untuk memisahkan kalimat menjadi kata per kata. Pada Tabel 3 merupakan hasil tokenizing.

Tabel 3

Sebelum Data Cleaning	Sesudah Data Cleaning
kalo dari awal segampang itu dapet dana buat benerin fasilitas ga bakal kejadian kek gini nyatanya bukan cuma fasilitas tapi kriminalitas yg tinggi juga ga segampang itu diselesein pake uang coba deh baca banyak kok di situs berita indo besar bahas topik kondisi rohingya	['kalo', 'dari', 'awal', 'segampang', 'itu', 'dapet', 'dana', 'buat', 'benerin', 'fasilitas', 'ga', 'bakal', 'kejadian', 'kek', 'gini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'tapi', 'kriminalitas', 'yg', 'tinggi', 'juga', 'ga', 'segampang', 'itu', 'diselesein', 'pake', 'uang', 'coba', 'deh', 'baca', 'banyak', 'kok', 'di', 'situs', 'berita', 'indo', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']

4. Normalisasi Data

Normalisasi data mengubah data yang memiliki kata tidak baku dan akan diubah menjadi kata baku. Proses pengubahan tersebut digunakan kamus bahasa slang. Pada Tabel 4 merupakan hasil dari Normalisasi data.

Tabel 4

Sebelum Normalisasi	Sesudah Normalisasi
['kalo', 'dari', 'awal', 'segampang', 'itu', 'dapet', 'dana', 'buat', 'benerin', 'fasilitas', 'ga', 'bakal', 'kejadian', 'kek', 'gini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'tapi', 'kriminalitas', 'yg', 'tinggi', 'juga', 'ga', 'segampang', 'itu', 'diselesein', 'pake', 'uang', 'coba', 'deh', 'baca', 'banyak', 'kok', 'di', 'situs', 'berita', 'indo', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']	['kalo', 'dari', 'awal', 'segampang', 'itu', 'dapat', 'dana', 'buat', 'benerin', 'fasilitas', 'enggak', 'bakal', 'kejadian', 'kayak', 'begini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'tapi', 'kriminalitas', 'yang', 'tinggi', 'juga', 'enggak', 'segampang', 'itu', 'diselesein', 'pakai', 'uang', 'coba', 'deh', 'baca', 'banyak', 'kok', 'di', 'situs', 'berita', 'indonesia', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']

5. Translate Data

Setelah dilakukan normalisasi dilanjutkan proses translate. Proses translate digunakan untuk proses pelabelan lexicon VADER. Proses translate mendapatkan hasil translate seperti pada Tabel 5 dibawah.

Tabel 5

Sebelum Translate	Sesudah Translate
['kalo', 'dari', 'awal', 'segampang', 'itu', 'dapet', 'dana', 'buat', 'benerin', 'fasilitas', 'ga', 'bakal', 'kejadian', 'kek', 'gini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'tapi', 'kriminalitas', 'yg', 'tinggi', 'juga', 'ga', 'segampang', 'itu', 'diselesein', 'pake', 'uang', 'coba', 'deh', 'baca', 'banyak', 'kok', 'di', 'situs', 'berita', 'indo', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']	If from the start it was that easy to get funds to repair the facilities, things like this wouldn't have happened like this. In fact, it's not just the facilities but high crime that is also not that easy to solve using money. Try reading a lot on big Indonesian news sites discussing the topic of Rohingya conditions.

6. Filtering Data

Filtering adalah membuang sebuah kata yang kurang memiliki maksa dalam kalimat. Kata tersebut dibuang karena tidak memiliki maksa. Hasil dari filtering data pada Tabel 6 dibawah.

Tabel 6

Sebelum Translate	Sesudah Translate
['kalo', 'dari', 'awal', 'segampang', 'itu', 'dapet', 'dana', 'buat', 'benerin', 'fasilitas', 'ga', 'bakal', 'kejadian', 'kek', 'gini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'tapi', 'kriminalitas', 'yg', 'tinggi', 'juga', 'ga', 'segampang', 'itu', 'diselesein', 'pake', 'uang', 'coba', 'deh', 'baca', 'banyak', 'kok', 'di', 'situs', 'berita', 'indo', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']	['awal', 'segampang', 'dana', 'buat', 'benerin', 'fasilitas', 'enggak', 'bakal', 'kejadian', 'kayak', 'begini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'kriminalitas', 'tinggi', 'enggak', 'segampang', 'diselesein', 'pakai', 'uang', 'coba', 'baca', 'banyak', 'kok', 'situs', 'berita', 'indonesia', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']

7. Stemming Data

Proses preprocessing paling akhir adalah stemming data. Stemming data akan mengubah kata yang berimbuhan menjadi kata dasar. Stemming dapat dilihat hasilnya pada Tabel 7

Tabel 7

Sebelum Translate	Sesudah Translate
['awal', 'segampang', 'dana', 'buat', 'benerin', 'fasilitas', 'enggak', 'bakal', 'kejadian', 'kayak', 'begini', 'nyatanya', 'bukan', 'cuma', 'fasilitas', 'kriminalitas', 'tinggi', 'enggak', 'segampang', 'diselesein', 'pakai', 'uang', 'coba', 'baca', 'banyak', 'kok', 'situs', 'berita', 'indonesia', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']	['awal', 'gampang', 'dana', 'buat', 'benerin', 'fasilitas', 'enggak', 'bakal', 'jadi', 'kayak', 'begini', 'nyata', 'bukan', 'cuma', 'fasilitas', 'kriminalitas', 'tinggi', 'enggak', 'gampang', 'diselesein', 'pakai', 'uang', 'coba', 'baca', 'banyak', 'kok', 'situs', 'berita', 'indonesia', 'besar', 'bahas', 'topik', 'kondisi', 'rohingya']

Labeling Data

Labeling data adalah proses untuk mengklasifikasi sentimen tersebut termasuk ke dalam sentimen apa. Proses pelabelan menggunakan metode lexicon InSet dan lexicon VADER. Berikut adalah hasil dari pelabelan lexicon tersebut.

1. Lexicon InSet

Berikut adalah hasil dari pelabelan lexicon InSet pada Tabel 8

Tabel 8 Hasil Labeling Lexicon InSet

Sentimen	Total
Negatif	5338
Positif	1338

Netral	509

2. Lexicon VADER

Berikut adalah hasil dari pelabelan lexicon InSet pada Tabel 9

Tabel 9 Hasil Labeling Lexicon VADER

Sentimen	Total
Negatif	2523
Positif	2748
Netral	1881

TF-IDF

Setelah dilakukan pelabelan dengan masing masing lexicon dilakukan pembobotan TF-IDF. Sebelum dilakukan pembobotan dilakukan pembagian data uji dan data latih dan penghapusan sentimen netral karena sentimen netral. Pada Tabel 10 adalah jumlah data uji masing masing

Tabel 10 Hasil Split Data Uji Latih dan Uji

Sentimen	Data Latih	Data Uji	Jumlah Data Latih	Jumlah Data Uji
InSet	70%	30%	4627	1983
	80%	20%	5288	1332
VADER	70%	30%	3690	1582
	80%	20%	4217	1055

Setelah dilakukan pembagian data dilakukam transformasi dengan TF-IDF.

Support Vector Machine

Setelah itu dilanjutkan untuk melatih model dengan SVM. Pada proses pelatihan model digunakan beberapa skenari hyperparameter C.

Confusion Matrix

Setelah dilakukan pelatihan proses paling akhir adalah hasil akurasi pelatihan model yang telah dilakukan. Berikut adalah hasil dari akurasi tersebut ditunjukkan pada Tabel 11

Tabel 11 Hasil Akurasi

Sentimen	Data Latih	Data Uji	C	Akurasi
InSet	70%	30%	0.1	84%
			1	89%
			10	90%
	80%	20%	0.1	84%
			1	90%
			10	90%
VADER	70%	30%	0.1	86%
			1	88%
			10	89%
	80%	20%	0.1	86%
			1	88%
			10	89%

Berdasarkan Tabel 11 tersebut didapatkan akurasi masing masing skenario berbeda. Dan dilakukan perhitungan rata rata dari masing masing lexicon didapatkan hasil sebagai berikut :

- Rata Rata Akurasi Lexicon InSet $=\frac{Total\ Akurasi}{Jumlah\ Skenario} = \frac{527}{6} = 87{,}83\%$
- Rata Rata Akurasi Lexicon VADER$=\frac{Total\ Akurasi}{Jumlah\ Skenario} = \frac{526}{6} = 87{,}66\%$

KESIMPULAN DAN SARAN

Penelitian tersebut didapatkan kesimpulan bahwa analisa sentimen Rohingya menggunakan lexicon InSet menghasilkan 5338 sentimen negatif, 1338 sentimen positif, dan 509 netral. Sedangkan pada lexicon VADER menghasilkan 2523 sentimen negatif, 2749 sentimen positif, dan 1881 netral. Dari kedua hal tersebut memiliki perbedaan dalam jumlah sentimen, pada labeling dengan InSet menghasilkan sentimen negatif yang lebih banyak

dibandingkan dengan positif, sedangkan pada labeling VADER menghasilkan sentimen yang lebih imbang antara sentimen positif dan negatif. Hasil kombinasi lexicon InSet SVM menghasilkan rata rata akurasi 87,83% dan pada kombinasi lexicon VADER SVM menghasilkan rata rata akurasi sebesar 87,66%. Dengan hal tersebut, lexicon InSet lebih unggul 0.17% dibandingkan lexicon VADER. Dan komposisi data latih dan data uji tidak memiliki pengaruh yang signifikan sedangkan pada komposisi hyperparameter C memiliki pengaruh terhadap hasil akurasi.

Saran untuk penelitian selanjutnya dilakukan penelitian dengan menggunakan metode lexicon lainnya dan jenis machine learning lain, percobaan lain dapat digunakan data lain untuk mengetahui performa. Selain itu bisa dibandingkan dengan teknik labeling manual untuk mengetahui perbedaan labeling otomatis dan manual.

DAFTAR REFERENSI

Arya, V., Mishra, A. K., & González-Briones, A. (2022). Sentiments analysis of covid-19 vaccine tweets using machine learning and vader lexicon method. Advances in Distributed Computing and Artificial Intelligence Journal, 11(4), 507–518. https://doi.org/10.14201/adcaij.27349

Azmi, B. N., Hermawan, A., & Avianto, D. (2023). Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 4(4), 281–290. https://doi.org/10.35746/jtim.v4i4.298

Biswas, S., Young, K., & Griffith, J. (2023). A Comparison of Automatic Labelling Approaches for Sentiment Analysis. Retrieved from https://www.researchgate.net/publication/370580498

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. Expert Systems with Applications, 162. https://doi.org/10.1016/j.eswa.2020.113746

Chan, J. Y. Le, Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2023). State of the art: a review of sentiment analysis based on sequential transfer learning. Artificial Intelligence Review, 56(1), 749–780. https://doi.org/10.1007/s10462-022-10183-8

D'Aniello, G., Gaeta, M., & La Rocca, I. (2022). KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. Artificial Intelligence Review, 55(7), 5543–5574. https://doi.org/10.1007/s10462-021-10134-9

El, I., Li, N. R., & Murphy, M. J. (n.d.). Theory and Applications Machine Learning in Radiation Oncology.

Geofany, N., & Liza, R. (n.d.). Klasifikasi Sentimen Tweet Pada Twitter Terhadap Pembelajaran E-Learning Menggunakan Metode k-Nearest Neighbor.

Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). Analisis Sentimen Aplikasi Ruang Guru di Twitter Menggunakan Algoritma Klasifikasi. Jurnal Teknoinfo, 14(2), 115. https://doi.org/10.33365/jti.v14i2.679

Gupta, N., & Agrawal, R. (2020). Application and techniques of opinion mining. In Hybrid Computational Intelligence (pp. 1–23). Elsevier. https://doi.org/10.1016/B978-0-12-818699-2.00001-9

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Retrieved from http://sentic.net/

Isnan, M., Elwirehardja, G. N., & Pardamean, B. (2023). Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model. Procedia Computer Science, 227, 168–175. https://doi.org/10.1016/j.procs.2023.10.514

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and Research: A Systematic Literature Review Through Text Mining. IEEE Access, 8, 67698–67717. https://doi.org/10.1109/ACCESS.2020.2983656

Koto, F., & Rahmaningtyas, G. Y. (2017). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. 2017 International Conference on Asian Language Processing (IALP), 391–394. https://doi.org/10.1109/IALP.2017.8300625

Machová, K., Mikula, M., Gao, X., & Mach, M. (2020). Lexicon-based sentiment analysis using particle swarm optimization. Electronics (Switzerland), 9(8), 1–22. https://doi.org/10.3390/electronics9081317

Muhammadi, R. H., Laksana, T. G., & Arifa, A. B. (2022). Combination of Support Vector Machine and Lexicon-Based Algorithm in Twitter Sentiment Analysis. Retrieved from https://github.com/evanmartua34/

Musfiroh, D., Khaira, U., Eko, P., Utomo, P., Suratno, T., Studi, P., Informasi, S., Sains, F., & Teknologi, D. (2021). Sentiment Analysis of Online Lectures in Indonesia from Twitter Dataset Using InSet Lexicon. Jurnal Teknologi Informasi dan Multimedia, 1, 24–33.

Tabassum, A., & Patil, R. R. (2020). A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. International Research Journal of Engineering and Technology. Retrieved from www.irjet.net

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

E-ISSN : 3046-7209 , P-ISSN : 3046-7217, Hal 62-76

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Zulkarnaini. (2023). Ratusan Pengungsi Rohingya Kembali Masuk Aceh.

Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM

ORIGINALITY REPORT

23%	21%	15%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.upnjatim.ac.id Internet Source	2%
2	Submitted to University College London Student Paper	1%
3	etd.repository.ugm.ac.id Internet Source	1%
4	Submitted to Swinburne University of Technology Student Paper	1%
5	journal.aptii.or.id Internet Source	1%
6	ijcaonline.org Internet Source	1%
7	etd.cput.ac.za Internet Source	1%
8	journal.widyakarya.ac.id Internet Source	1%
	jurnal.teknokrat.ac.id	

9	Internet Source	1 %
10	Submitted to Universitas Sebelas Maret Student Paper	1 %
11	jurnal.itscience.org Internet Source	1 %
12	kc.umn.ac.id Internet Source	1 %
13	www.mendeley.com Internet Source	1 %
14	Submitted to University of Huddersfield Student Paper	1 %
15	conference.upnvj.ac.id Internet Source	1 %
16	opac.uad.ac.id Internet Source	1 %
17	repository.ub.ac.id Internet Source	<1 %
18	repo.itpln.ac.id Internet Source	<1 %
19	journal.lpkd.or.id Internet Source	<1 %
20	digilib.uinsa.ac.id Internet Source	<1 %

21	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	<1 %
22	etheses.uin-malang.ac.id Internet Source	<1 %
23	Submitted to Kaplan College Student Paper	<1 %
24	ojs.unud.ac.id Internet Source	<1 %
25	iranarze.ir Internet Source	<1 %
26	Submitted to Universitas Brawijaya Student Paper	<1 %
27	repository.uin-suska.ac.id Internet Source	<1 %
28	Submitted to Universitas Sanata Dharma Student Paper	<1 %
29	jurnal.ilmubersama.com Internet Source	<1 %
30	Eka Rini Yulia, Kusmayanti - Solecha. "Implementasi Particle Swarm Optimization (PSO) pada Analysis Sentiment Review Aplikasi Trafi menggunakan Algoritma Naive Bayes (NB)", Jurnal Teknik Komputer, 2021 Publication	<1 %

31	dergipark.org.tr Internet Source	<1 %
32	journal.sinov.id Internet Source	<1 %
33	link.springer.com Internet Source	<1 %
34	openlibrarypublications.telkomuniversity.ac.id Internet Source	<1 %
35	pdfs.semanticscholar.org Internet Source	<1 %
36	Admi Syarif, Arafia Isnayu Akaf, Rizky Prabowo, Kurnia Muludi. "Analisis Sentimen Opini Masyarakat Terhadap Pelayanan BPJS Kesehatan Provinsi Lampung Berbasis Twitter", Jurnal Pepadun, 2022 Publication	<1 %
37	Nur Fajriyani, Enda Esyudha Pratama, Rina Septiriana. "Optimasi Hyperparameter pada Neural Network (Studi Kasus: Identifikasi Komentar Cyberbullying Instagram)", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2023 Publication	<1 %
38	es.scribd.com Internet Source	<1 %
	medium.com	

39	Internet Source	<1%
40	ojs.stmikplk.ac.id Internet Source	<1%
41	seminar.ilkom.unsri.ac.id Internet Source	<1%
42	yudiayutz.wordpress.com Internet Source	<1%
43	Afrian Hanafi, Adiwijaya Adiwijaya, Widi Astuti. "Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor", Jurnal Sisfokom (Sistem Informasi dan Komputer), 2020 Publication	<1%
44	Afriani Afriani, Herry Sujaini, Niken Candraningrum. "Analisis Perbandingan Metode Pengklasifikasi Gambar Jenis Tulisan Kaligrafi Arab", Jurnal Edukasi dan Penelitian Informatika (JEPIN), 2024 Publication	<1%
45	jurnal.polsri.ac.id Internet Source	<1%
46	publishing-widyagama.ac.id Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Perbandingan Performa Labeling Lexicon InSet dan VADER pada Analisa Sentimen Rohingya di Aplikasi X dengan SVM

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15