Model Klasifikasi Teks Produk Terlarang Menggunakan Algoritma Campuran (Hybrid) SVM dan Leksikon

Abdul Hanif Al-atho'illah[1)], Eka Dyar Wahyuni[2)], Amalia Anjani Arifiyanti[3)]
Email : [1)]abdulhanifalathoillah@gmail.com [2)]ekawahyuni.si@upnjatim.ac.id
[3)]amalia_anjani.fik@upnjatim.ac.id

Prodi Sistem Informasi, Fakultas Ilmu Komputer, UPN "Veteran" Jawa Timur

ABSTRAK

Jual beli secara daring merupakan aktivitas yang sudah menjadi hal yang biasa bagi masyarakat modern saat ini. Mulai dari usia muda hingga tua, 96% pengguna internet di Indonesia pernah menjelajahi web jual beli daring. Hal tersebut tentu karena peran teknologi yang membantu mempermudah aktivitas jual beli secara konvensional. Namun kemudahan tersebut belum didukung dengan sistem keamanan yang optimal. Masih banyak web jual beli daring membiarkan produk tertentu yang proses jual beli nya perlu dibatasi bahkan dilarang, karena beresiko mendukung tindak kriminalitas, Dsb. Namun tetap dapat tampil begitu saja pada berbagai web tersebut. Sehingga sebuah sistem klasifikasi teks yang dapat mengklasifikasikan data teks dari produk terlarang akan dapat menjadi salah satu opsi solusi dari permasalahan tersebut. Sistem klasifikasi teks yang merupakan bagian dari bidang penambangan kata, disusun oleh suatu model klasifikasi. Model klasifikasi dengan performa yang baik akan menghasilkan sistem klasifikasi yang baik pula. Penelitian ini membangun model klasifikasi dengan menguji 3 jenis pendekatan yaitu; pendekatan pengetahuan dengan kamus leksikon, pendekatan machine learning dengan algoritma Support Vector Machine (SVM), dan pendekatan gabungan dari keduanya (Hybrid) sehingga dapat menghasilkan suatu model dengan performa akurasi terbaik dalam klasifikasi teks produk. Dengan menggunakan 300 dataset produk didapatkan hasil akurasi terbaik mencapai 76,64%, recall sebesar 77%, dan presisi sebesar 78%. Setelah model klasifiksi dibuat, penelitian ini juga akan merancang sistem klasifikasi berbasis web. Sehingga sistem klasifikasi safety product dapat dibangun hingga diluncurkan dan dapat memberikan prediksi terhadap masukan teks dari pengunjung secara dinamis.

Kata Kunci : *Sistem Klasifikasi Teks, Produk Terlarang, Text Mining, Leksikon, SVM*

1. PENDAHULUAN

Pada era digital saat ini aktivitas jual-beli menjadi lebih mudah dengan adanya teknologi. Kegiatan jual-beli secara digital atau daring telah menjadi hal yang biasa di era saat ini. Teknologi telah memberikan kemudahan akses kepada siapapun dalam melakukan aktivitas jual-beli. Namun dengan kemudahan akses tersebut, juga membawa resiko penyalahgunaan dalam penggunaannya. Menurut survei pada yang dilakukan oleh Paypal[3] pada tahun 2017, pengguna *e-commerce* di Indonesia yang masih berusia dibawah 20 tahun dan berstatus sebagai pelajar atau mahasiswa mencapai 9% dari 4000 responden. Sedangkan survei dari 'We Are Social'[4] pada 2019 menyatakan bahwa 96% pengguna internet di Indonesia penah menggunakan *e-commerce*. Dengan kondisi tersebut maka akan sangat mungkin seseorang melakukan jual-beli barang atau produk yang seharusnya tidak untuk diperjual-belikan. Hal tersebut dapat membahayakan konsumen, dan juga dapat beresiko meningkatkan kriminalitas di dalam masyarakat.

Sehingga diperlukan adanya suatu mekanisme atau sistem yang dapat menyaring dan mengklasifikasi produk sebelum ditampilkan. Sistem tersebut akan bertugas untuk mengenali suatu produk yang akan di unggah oleh penjual merupakan produk yang dilarang atau tidak. Data suatu produk yang akan di iklankan hampir semuanya merupakan data jenis teks. Dalam dunia teknologi, pengolahan data teks sudah banyak di terapkan antara lain yaitu penambangan teks atau *text mining*.

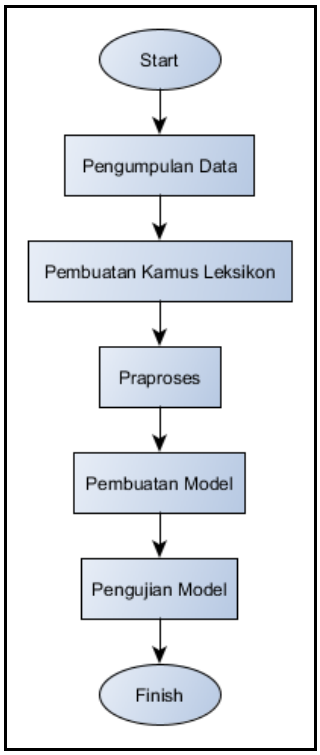
Text mining (Penambangan Teks) merupakan suatu proses untuk mengekstrak pola dalam mengeksplorasi pengetahuan dari sumber data yang berbentuk teks[7]. Penambangan teks telah banyak digunakan untuk memproses sebuah data berupa teks

untuk diekstrak menjadi sebuah informasi yang dapat dikenali oleh sistem salah satunya proses klasifikasi teks. Dalam klasifikasi teks terdapat 2 jenis pendekatan umum, yaitu pendekatan *Machine Learning* & pendekatan *Knowlede Based* (Berbasis Pengetahuan). Dalam proses pembuatan model klasifikasi teks, pada penelitian ini menggunakan bahasa pemrograman python. Python mendukung banyak *library packages* yang membantu dalam memproses data. Selain itu python juga termasuk bahasa yang cenderung relative mudah dipahami.

Pendekatan *machine learning* memerlukan dataset untuk digunakan sebagai data latih. Oleh karena itu, dibutuhkan usaha untuk mengumpulkan dan melakukan *class tag* pada sampel *data set* tersebut, selain itu proses training juga membutuhkan waktu[2] . Akurasi dari pendekatan klasifikasi *machine learning* sangat baik, akan tetapi performa klasifikasinya domain dependen terhadap *data set* yang digunakan pada saat pelatihan[5]. Sedangkan pendekatan berbasis pengetahuan bergantung pada *dictionary* atau kamus leksikon yang digunakan untuk melakukan penilaian terhadap fitur yang didapat. Sebagai luaran dalam penelitian ini akan dibandingkan hasil klasifikasi dari pendekatan SVM, Leksikon, dan pendekatan gabungan dari keduanya. Sehingga akan didapatkan performa terbaik sistem klasifikasi teks untuk produk terlarang.

2. METODOLOGI

Pada proses pembuatan model klasifikasi teks menggunakan text mining, setidaknya ada 5 tahap antara lain; pengumpulan data, pembuatan kamus leksikon, praproses, pembuatan model, hingga pengujian model. Gambar diagram alir penelitian dapat dilihat sebagai berikut



Gambar 1. Diagram Alir Penelitian

2.1 Pengumpulan Data

Proses ini adalah pengumpulan data yang akan digunakan untuk uji coba dalam perancangan model klasifikasi. Data yang digunakan merupakan data dari web portal jual beli daring yang populer[6] antara lain; Tokopedia, Shopee, dan Bukalapak. Data tersebut merupakan data produk yang ‘Boleh’ (Safety) dan ‘Dilarang’ (Non-Safety. Pada proses

ini akan dilakukan 2 sub-proses yaitu; pengambilan data, dan pelabelan data. Pengambilan data merupakan proses *scrapping web* yaitu mengumpulkan data dari web. Sedangkan proses selanjutnya adalah proses pelabelan data yaitu melabeli data secara manual terhadap dua kelas yaitu kelas 'Boleh'/ 'Positif' dan kelas 'Dilarang'/ 'Negatif'.

2.2 Pembuatan Kamus Leksikon

Salah satu pendekatan yang akan digunakan pada model klasifikasi teks, adalah pendekatan leksikon. Pendekatan leksikon adalah pendekatan klasifikasi teks berdasarkan daftar kosa kata yang terhimpun dalam kamus leksikon. Pada tahap membuat leksikon, peneliti mengumpulkan berbagai kosakata yang berhubungan dengan produk-produk yang dilarang sesuai yang telah ditentukan. Leksikon yang dibuat memuat antara lain; berbagai merk alkohol, daftar zat/ barang berbahaya, daftar nama liquid dan vape/ rokok elektrik, daftar obat-obatan terlarang, daftar nama senjata api, daftar nama hewan, daftar merk rokok, dan daftar barang terlarang lainnya. Kemudian daftar leksikon akan dibandingkan dengan sampel data yang telah dikumpulan untuk memastikan tidak ada kata yang bermakna ganda dan merujuk ke kata produk yang tidak dilarang. Leksikon yang telah dibuat kemudian disusun dalam format .csv.

2.3 Praproses

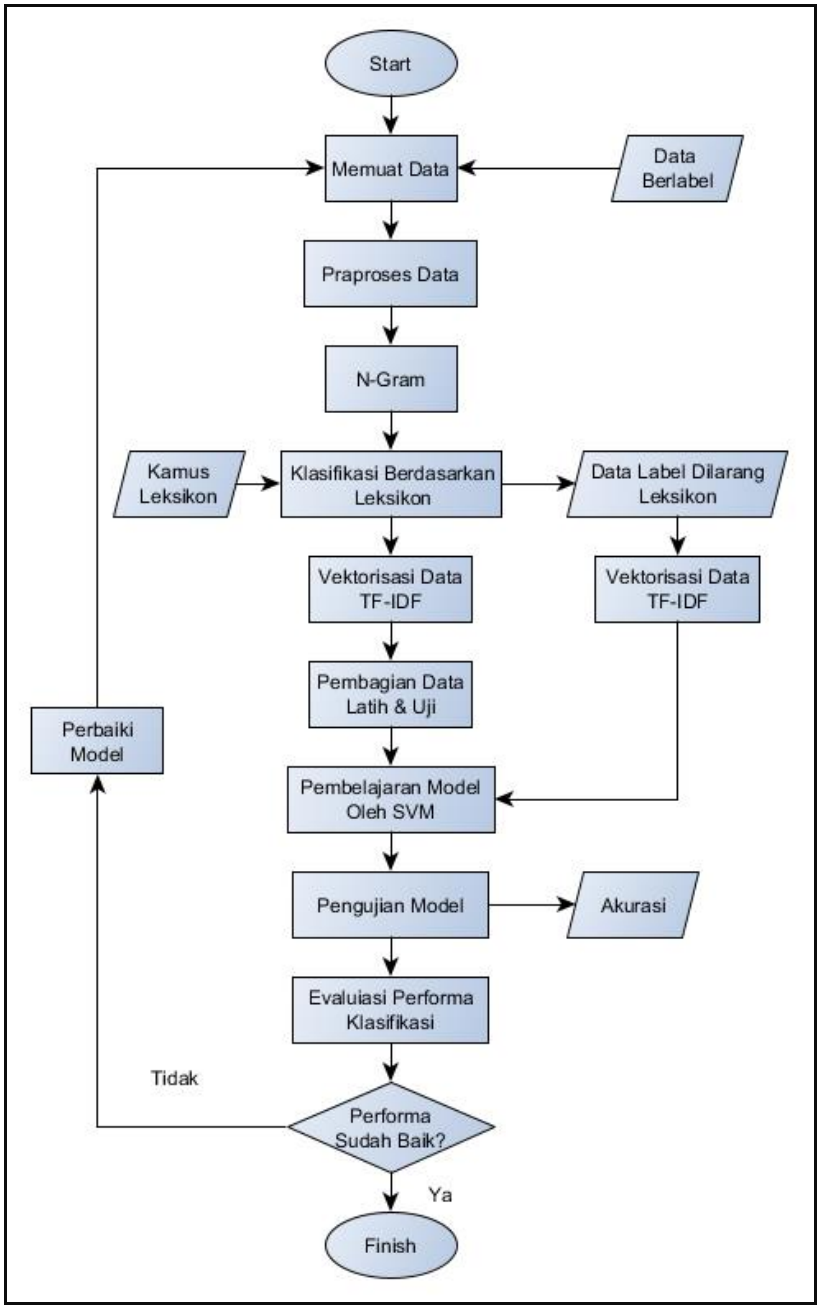
Pada tahap praproses, data mentah yang telah dilabeli akan dibersihkan agar model yang dijalankan dapat bekerja optimal dalam mengenali pola yang ada pada data. Dalam praproses ada dua tahap utama yaitu; seleksi fitur dan pembobotan fitur. Pada tahap seleksi fitur, pertama peneliti akan melakukan *cleansing* pada data untuk menghilangkan karakter yang tidak diperlukan. Proses selanjutnya melakukan tokenisasi (*tokenizing*) pada data yang berupa teks. Sehingga terpecah menjadi kumpulan token untuk tiap kata nya agar lebih mudah untuk di proses. Selanjutnya dilakukan proses *POS Tagging* atau pemberian tag pada tiap kata untuk mengenali kelas kata yang ada. Pemberian tag di penelitian ini akan mengambil kata kerja, kata sifat, kata benda, dan kata asing yang mana akan dikenali sebagai ciri atau entitas dari suatu produk. Proses selanjutnya dilakukan *stopword removal* untuk menghilangkan kata-kata yang tidak penting atau tidak memiliki arti seperti 'yang', 'dan', dsb. Terakhir data akan melalui proses *stemming* yang mana untuk menghilang kan imbuhan kata sehingga tiap kata akan kembali ke bentuk atau kata dasarnya.

Pada tahap pembobotan fitur, peneliti akan merubah bentuk data terlebih dahulu menggunakan metode n-gram. Pada metode n-gram data akan di pisahkan menjadi kumpulan fitur/ gram. Pada penelitian ini akan digunakan bentuk unigram dan bigram. Unigram adalah bentuk dimana satu fitur berisi satu kata. Sedangkan bigram adalah bentuk dimana satu fitur berisi dua kata. Setelah data berbentuk unigram dan bigram, selanjutnya data akan di bobot/ vektorisasi. Vektorisasi data adalah proses normalisasi data teks dengan pemberian nilai terhadap setiap fitur. Pada penelitian ini digunakan teknik TF-IDF untuk pemberian bobot fitur. Teknik ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap fitur di setiap dokumen dalam korpus. Secara sederhana, TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen. Pembobotan ini dilakukan agar sistem dapat menganalisa pola yang ada dari suatu teks terhadap suatu kelas. Selain itu karena sistem hanya dapat memproses suatu data vektor atau angka.

2.4 Pembuatan Model

Pada proses pembuatan model klasifikasi menggunakan penambangan teks, dalam penelitian ini menggunakan pendekatan berdasarkan Leksikon dan pendekatan *Machine Learning* algoritma SVM. Pendekatan campuran leksikon dan SVM ini akan dilakukan secara bertingkat/ bertahap[1][8]. Data akan di klasifikasi oleh leksikon

terlebih dahulu. Kemudian data yang tidak terklasifikasi oleh leksikon akan diklasifikasi menggunakan SVM dengan pembelajaran atas data latih. Berikut skema algoritma uang akan digunakan.



Gambar 2. Skema Model Klasifikasi Campuran Leksikon-SVM

Pada skema ini proses dimulai dari memuat data yang telah dilabeli pada proses sebelumnya. Kemudian dilanjutkan dengan praproses data hingga n-gram. Setelah data melalui proses n-gram, data akan diklasifikasi dengan mencocokkan tiap token dari bentuk n-gram dengan kamus leksikon yang sebelumnya sudah dibuat. Kemudian data akan diklasifikasi menggunakan kamus leksikon. Namun hasil data yang terlabel 'Positif' masih akan diklasifikasi dengan algoritma SVM. Setelah data melalui proses klasifikasi leksikon, data terlabel 'Negatif' dan 'Positif hasil klasifikasi leksikon akan divektorisasi dengan TF-IDF untuk tiap bentuk pada n-gram. Setelah bentuk menjadi vektor, data berlabel 'Positif' hasil klasifikasi leksikon akan dibagi menjadi dua; data

latih dan data uji. Pembagian data menggunakan teknik *hold-out* dengan perbandingan data latih sebesar 80%. Kemudian data terlabel 'Negatif' hasil klasifikasi Leksikon ditambahkan pada data latih. Namun label yang akan digunakan sebagai acuan untuk membantu proses pelatihan adalah label asli, bukan label 'Negatif' dari klasifikasi Leksikon. Proses selanjutnya adalah pelatihan algoritma SVM. Pada proses pelatihan SVM, akan ada beberapa jenis kernel yang akan diuji coba antara lain; Linear, Polinomial, dan RBF. Tahap selanjutnya dari skema ini adalah proses pengujian hasil dari proses klasifikasi untuk mendapatkan akurasi dari pemodelan ini. Namun dalam proses pengujian hasil pada skema ini akan membandingkan hasil prediksi dari klasifikasi leksikon dan hasil prediksi algoritma SVM dengan label asli atau label sebenarnya. Setelah proses pengujian, akan dilakukan evaluasi model. Jika model perlu diperbaiki, maka model akan diperbaiki dan mengulangi proses yang ada. Jika model cukup baik, maka proses selesai.

2.5 Pengujian Model

Pada tahap terakhir dalam proses model klasifikasi adalah pengujian model. Pada tahap ini data uji akan digunakan untuk menguji akurasi dan performa model dalam mengklasifikasikan data. Teknik yang digunakan dalam pengujian model adalah *Confusion Matrix* yang mana menghitung jumlah dari data positif yang ditebak dengan benar, data positif yang ditebak salah, data negatif yang ditebak benar dan data negatif yang ditebak salah. Hasil dari teknik ini akan didapatkan akurasi, *recall*, dan *precision* dari model yang telah dibuat terhadap data uji. Selain itu sebagai bahan perbandingan pada pengujian model juga akan ditampilkan hasil performa dari algoritma Leksikon dan algoritma SVM terhadap *data set* yang sama. Sehingga akan dapat dinilai pendekatan campuran leksikon-SVM dapat meningkatkan performa klasifikasi teks. Atau justru akan menurunkan performa dari algoritma tunggal yang menjadi penyusunnya. Sehingga terdapat tujuh skenario yang akan dibandingkan antara lain;

1) Skenario 1 : Algoritma Leksikon
2) Skenario 2 : Algoritma SVM (Kernel Linear)
3) Skenario 3 : Algoritma SVM (Kernel Polinomial)
4) Skenario 4 : Algoritma SVM (Kernel RBF)
5) Skenario 5 : Algoritma Leksikon-SVM (Kernel Linear)
6) Skenario 6 : Algoritma Leksikon-SVM (Kernel Polinomial)
7) Skenario 7 : Algoritma Leksikon-SVM (Kernel RBF)

Selain itu pada tahap ini juga akan dilakukan evaluasi model. Evaluasi model dilakukan untuk memperbaiki model jika ternyata hasil dari pengujian model ditemukan kekurangan atau kesalahan.

3. HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan tentang hasil dan pembahasan dari penelitian yang telah dilakukan. Pada penelitian ini terdapat lima tahapan utama dan beberapa sub tahapan didalamnya. Lima tahapan tersebut antara lain; Pengumpulan Data, Pembuatan Kamus Leksikon, Praproses, Pembuatan Model, dan Pengujian Sistem.

3.1 Pengumpulan Data

Pada tahap pengumpulan *data set* akan menggunakan aplikasi Gaboosh Media dan Dropshipkit. Data yang dikumpulkan berjumlah 150 data dengan kelas 'Negatif' dan 150 data dengan kelas 'Positif'. Sehingga total data berjumlah 300 data teks. Data yang akan digunakan hanya data teks dari judul dan deskripsi dari produk. Setelah data dihimpun, data akan diberi label 'Positif' dan 'Negatif' secara manual dan disimpan dalam format .csv.

3.2 Pembuatan Kamus Leksikon

Pada tahap pembuatan kamus leksikon, akan dikumpulkan kosa kata yang merujuk ke produk terlarang. Kosakata tersebut antara lain merk dari berbagai produk terlarang seperti; merk minuman keras, merk rokok, nama senjata api, nama obat atau bahan kimia yang berbahaya. Selain itu penyusunan kamus leksikon juga didasarkan dari pembelajaran atas 300 data sampel yang dikumpulkan. Sehingga kamus leksikon dapat lebih optimal karna berisi kosakata yang secara spesifik cenderung merupakan produk 'Negatif'. Sedangkan kosa kata yang dapat bermakna ganda dan cenderung bermakna 'Positif' akan dihilangkan. Jumlah kosa kata yang dikumpulkan berjumlah 1591 kosa kata.

3.3 Praproses

Pada tahap praproses data akan dibersihkan dari unsur – unsur yang tidak diperlukan dalam klasifikasi produk melalui proses seleksi fitur. Selain dibersihkan data akan di vektorisasi sehingga data yang awalnya berupa teks dapat diolah oleh sistem melalui proses pembobotan fitur.

3.3.1 Seleksi Fitur

Tahap seleksi fitur dimulai dengan *cleansing*, berfungsi untuk menghilangkan karakter yang tidak diperlukan dalam teks seperti; kapital, tanda baca, nomor/ angka, url. Dalam proses ini menggunakan fungsi yang tesedia dalam bahasa python antara lain; *lower* dan *replace*. Tahap tokenisasi akan merubah data yang berupa teks menjadi token. Token terdiri dari satu kosa kata data teks. Proses ini menggunakan *package library* nltk dan fungsi *tokenize*. Tahap *POS-Tagging* dilakukan dengan memberi label pada tiap token. Pada tahap ini digunakan package library CRFTagger bahasa Indonesia. Setelah diberi label, akan dilakukan seleksi untuk menyimpan token yang berlabel; NN (Kata Benda), JJ (Kata Sifat), dan FW (Kata Asing). Token dengan label selain itu akan dihilangkan untuk mengoptimalkan data. Selanjutnya melakukan *stopword removal*, akan digunakan *package library* Sastrawi. Jika token terdapat dalam daftar *stopword*, maka token kata tersebut akan dihilangkan. Daftar *stopword* berisi kata kata yang tidak memiliki arti atau makna yang akan lebih baik untuk dihilangkan agar proses klasifikasi dapat optimal. Tahapan *stemming* ditunjukan untuk merubah kata yang berimbuhan menjadi kembali ke kata dasarnya. *Library stemmer* yang digunakan pada penelitian ini adalah Sastrawi.

3.3.2 Pembobotan Fitur

a. N-gram

Tahap selanjutnya setelah data di preposes adalah dengan menerapkan n-gram yang merupakan teknik tokenisasi berdasarkan jumlah kata. Data akan di rubah ke dua jenis bentuk token yaitu unigram dan bigram. Unigram merupakan tokenisasi yang terdiri dari satu kata. Sedangkan bigram adalah tokenisasi yang terdiri dari dua kosakata. *Package library* yang digunakan dalam proses ini adalah nltk.

b. Vektorisasi TF-IDF

Proses ini adalah untuk merubah data yang berupa teks menjadi vektor yang dapat dibaca oleh sistem. Proses vektorisasi dilakukan menggunakan TF-IDF *Converter*. Algoitma TF-IDF *Converter* adalah merubah data teks menjadi numerik berdasarkan kemunculan atau frekuensi suatu kata dan keterkaitan kata tersebut terhadap keseluruhan data dan labelnya.

3.4 Pembuatan Model

Dalam pembuatan model klasifikasi teks agoritma leksikon-SVM data terlebih dahulu melalui praproses untuk dibersihkan. Kemudian data yang telah dalam bentuk

unigram dan bigram akan diklasifikasikan dengan algoritma leksikon. Hasil proses klasifikasi leksikon pada 300 produk dataset, sistem menemukan 90 data unigram dan 7 data bigram mengandung kosakata negatif. Sehingga total sistem melabeli 96 produk sebagai kelas 'Negatif' dan 204 sebagai produk yang akan di proses lebih lanjut dengan algoritma klasifikasi SVM. Hasil ini juga menjadi hasil dari klasifikasi leksikon yang akan diukur performanya sebagai pembanding. Data produk yang dilabeli sebagai negatif akan dipisahkan dan akan kemudian akan disatukan dengan data latih dalam proses klasifikasi SVM.

Sebelum masuk ke penerapan klasifikasi SVM, data terlebih dahulu masuk ke tahap pembagian data. Teknik yang digunakan dalam tahap pembagian data adalah *hold-out*, yang mana akan secara acak memilih data yang akan dijadikan data latih maupun data uji. Perbandingan yang dipilih adalah 80% data Latih dan 20% data uji. Data latih kemudian akan dipelajari oleh algoritma SVM. Sedangkan data uji akan digunakan untuk menguji hasil pembelajaran sistem. Pada tahap ini juga data dengan label 'Negatif' hasil klasifikasi leksikon akan ditambahkan sebagai data latih.

```
from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest = train_test_split(xdata,ydata,test_size=0.2, random_state=1)
```

```
ytrain.value_counts()
```

```
1    100
0     63
Name: label, dtype: int64
```

```
ytest.value_counts()
```

```
1    28
0    13
Name: label, dtype: int64
```

Gambar 3. Hasil Pembagian Data Latih dan Uji Algoritma Leksikon-SVM

Hasil pembagian dengan teknik *hold-out* didapat 163 data latih dan 41 data uji. Kemudian diketahui data berlabel 'Negatif' hasil klasifikasi leksikon tedapat 96 data yang mana 74 data berlabel 'Negatif' dan 22 data berlabel 'Positif' pada label sebenarnya. Sehingga keseluruhan data latih berjumlah 259 data yang mana terdiri dari 137 data berlabel 'Positif' dan 122 data berlabel 'Negatif'.

```
ydataN.value_counts()
```

```
0    74
1    22
Name: label, dtype: int64
```

```
xtrain = xtrain.append(xdataN)
ytrain = ytrain.append(ydataN)
```

```
ytrain.value_counts()
```

```
0    137
1    122
Name: label, dtype: int64
```

Gambar 4. Hasil Akhir Data Latih Algoritma Leksikon-SVM

Setelah dilakukan pembagian data latih dan uji, proses pelatihan algoritma SVM dapat dilakukan. Proses pelatihan menggunakan fungsi *fit* pada kelas SVC algoritma

SVM. Model ini akan dilakukan ujicoba pada 3 kernel yaitu; Linear, Polinomial, dan RBF dengan pengaturan parameter default.

3.5 Pengujian Model

Dalam tahap pengujian model ini akan dilakukan dua aktivitas utama, yaitu pengujian skenario dan evaluasi model. Namun setelah pengujian skenario, ditemukan bahwa pada model klasifikasi yang menggunakan algortima SVM tidak memberikan hasil yang bagus. Algortima SVM pada kernel linear tidak mampu mengklasifikasikan data pada kelas 'Dilarang' atau label 0. Sedangkan pada pendekatan gabungan Leksikon-SVM di kernel linear dan kernel Polynomial. Model tidak mampu mengklasifikasikan kelas data 'Boleh' atau label 1. Sehingga dicoba menggunakan proses penyeimbangan data latih, dan menghasilkan performa yang lebih baik. Proses penyeimbangan data latih menggunakan teknik *Synthetic Minority Over-Sampling Technique* (SMOTE) *imbalance data handling* dengan *library imblearn's SMOTE*. Selanjutnya hasil dari skenario dapat dilihat sebagai berikut.

1) Skenario 1 : Algoritma Leksikon

Pada skenario ini data positif yang benar di prediksi sebagai positif sebanyak 74 data. Data positif yang salah di prediksi sebagai negatif sebanyak 22 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 76 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 128 data. Sehingga pada skenario ini mampu menghasilkan nilai akurasi sebesar 67.3%, nilai *recall* sebesar 67%, dan nilai *precision* sebesar 70%.

2) Skenario 2 : Algoritma SVM (Kernel Linear)

Pada skenario ini setelah menggunakan metode SMOTE data positif yang benar di prediksi sebagai positif sebanyak 25 data. Data positif yang salah di prediksi sebagai negatif sebanyak 22 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 8 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 5 data. Sehingga menghasilkan nilai akurasi sebesar 50%, nilai *recall* sebesar 50%, dan nilai *precision* sebesar 47%.

3) Skenario 3 : Algoritma SVM (Kernel Polinomial)

Pada skenario ini setelah menggunakan metode SMOTE data positif yang benar di prediksi sebagai positif sebanyak 6 data. Data positif yang salah di prediksi sebagai negatif sebanyak 3 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 27 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 24 data. Sehingga menghasilkan nilai akurasi sebesar 50%, nilai *recall* sebesar 50%, dan nilai *precision* sebesar 58%.

4) Skenario 4 : Algoritma SVM (Kernel RBF)

Pada skenario ini setelah menggunakan metode SMOTE data positif yang benar di prediksi sebagai positif sebanyak 21 data. Data positif yang salah di prediksi sebagai negatif sebanyak 11 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 12 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 16 data. Sehingga menghasilkan nilai akurasi sebesar 61,67%, nilai recall sebesar 62%, dan nilai precision sebesar 62%.

5) Skenario 5 : Algoritma Leksikon-SVM (Kernel Linear)

Pada skenario ini setelah menggunakan metode SMOTE data positif yang benar di prediksi sebagai positif sebanyak 74 data. Data positif yang salah di prediksi sebagai negatif sebanyak 27 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 13 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 23 data. Sehingga menghasilkan nilai akurasi sebesar 70,80%, nilai *recall* sebesar 71%, dan nilai *precision* sebesar 70%.

6) Skenario 6 : Algoritma Leksikon-SVM (Kernel Polinomial)

Pada skenario ini setelah menggunakan metode SMOTE data positif yang benar di prediksi sebagai positif sebanyak 74 data. Data positif yang salah di prediksi sebagai negatif sebanyak 23 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 13 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 27 data. Sehingga menghasilkan nilai akurasi sebesar 73,72%, nilai *recall* sebesar 74%, dan nilai *precision* sebesar 73%.

7) Skenario 7 : Algoritma Leksikon-SVM (Kernel RBF)

Pada skenario ini setelah menggunakan metode SMOTE data positif yang benar di prediksi sebagai positif sebanyak 85 data. Data positif yang salah di prediksi sebagai negatif sebanyak 27 data. Data negatif yang benar di prediksi sebagai negatif sebanyak 5 data. Sedangkan data negatif yang salah di prediksi sebagai negatif sebanyak 23 data. Sehingga menghasilkan nilai akurasi sebesar 76,64%, nilai *recall* sebesar 77%, dan nilai *precision* sebesar 78%. Sehingga skenario ini menghasilkan performa terbaik disbanding skenario yang lain.

Tabel 1. Perbandingan Hasil Uji Skenario

	kelas prediksi	kelas sebenarnya 0	kelas sebenarnya 1	Akurasi	Precision	Recall
skenario 1	0	74	76	67,33%	70%	67%
	1	22	128			
skenario 2	0	25	8	50,00%	47%	50%
	1	22	5			
skenario 3	0	6	27	50,00%	58%	50%
	1	3	24			
skenario 4	0	21	12	61,67%	62%	62%
	1	11	16			
skenario 5	0	74	13	70,80%	70%	71%
	1	27	23			
skenario 6	0	74	13	73,72%	73%	74%
	1	23	27			
skenario 7	0	85	5	**76,64%**	**78%**	**77%**
	1	27	23			

4. KESIMPULAN DAN SARAN

Pada bab ini dijelaskan tentang kesimpulan dari penelitian ini. Selain itu juga akan dipaparkan saran untuk penelitian selanjutnya agar dapat memperbaiki kekurangan pada penelitian ini.

4.1 Kesimpulan

Berdasarkan hasil analisis dan implementasi, maka dapat disimpulkan beberapa hal sebagai berikut:

A. Algoritma SVM merupakan algoritma yang performa nya akan semakin baik seiring banyaknya *data set* yang digunakan untuk melatih sistem. Sehingga performa hasil dari model yang menggunakan algoritma SVM akan dapat menjadi lebih baik jika data latih yang digunakan terus di tingkatkan. Sedangkan pada algortima leksikon

merupakan algoritma yang sangat dipengaruhi oleh kamus leksikon yang ada. Sehingga pemilihan kata dalam penyusunan kamus leksikon sangat penting dan diperlukan pemahaman terhadap data yang akan di klasifikasi dengan algoritma leksikon ini. Namun ini juga menjadi sangat efektif untuk menyaring data produk yang mengandung kosakata / istilah spesifik merupakan produk terlarang, seperti; jenis/ istilah narkoba, bahan kimia terlarang, alat kontrasepsi, dsb. Bahkan tanpa data latih sekalipun.

B. Model klasifikasi teks yang digunakan pada sistem yang dibuat menggunakan model klasifikasi gabungan antara leksikon dengan SVM. Data diklasifikasi dengan leksikon terlebih dahulu diimplementasikan. Kemudian selanjutnya akan diklasifikasi oleh SVM dengan kernel RBF yang sebelumnya telah dilakukan penyeimbangan data menggunakan SMOTE. Pendekatan ini menghasilkan performa terbaik dengan nilai akurasi sebesar 76,64%, nilai *recall* sebesar 77%, dan nilai *precision* sebesar 78%.

4.2 Saran

Adapun saran yang bisa diberikan untuk pengembangan dari penelitian ini untuk penelitian selanjutnya antara lain :

A. Penambahan jumlah data yang digunakan untuk dapat mengoptimalkan algoritma *machine learning* yang cenderung akan semakin baik dengan semakin banyaknya data latih.
B. Pengembangan kamus leksikon produk terlarang yang lebih sesuai dengan data yang ada di lapangan untuk meningkatkan kemampuan klasifikasi produk
C. Penambahan teknik vektorisasi fitur lain bersama TF-IDF untuk meningkatkan keakuratan algoritma SVM yang memberikan klasifikasi berdasakan vektor sebagai dimensi data.
D. Penelitian selanjutnya diharapkan dapat mengantisipasi negation-handling untuk memberikan klasifikasi yang lebih baik.

5. DAFTAR RUJUKAN

[1] El-Halees, Alaa. 2011. Arabic Opinion Mining Using Combined Classification Approach.
[2] Cao, Jiaping et al. 2014. Web-based traffic sentiment analysis: Methods and applications. IEEE transactions on Intelligent Transportation systems, 15(2), 844-853.
[3] Databoks. 2019. Pelaku e-commerce Didominasi Usia Muda. [Internet]. Laman: https://databoks.katadata.co.id/datapublish/ 2019/04/01/pelaku-e-commerce-didominasi-usia-muda. [diakses 13 September 2019].
[4] Databoks. 2019. 96% Pengguna Internet di Indonesia Pernah Menggunakan ECommerce. [Internet]. Laman: https://databoks.katadata.co.id/datapublish/2019/12/03/96-penggunainternet-di-indonesia-pernah-gunakan-e-commerce. [diakses 13 September 2019].
[5] Kundi, F. M., Asghar, M. Z. 2014. Lexicon-based sentiment analysis in the social web. Journal of Basic and Applied Scientific Research, 4(6)
[6] Solutech. 2019. 5 Marketplace Terbaik di Indonesia pada 2019. [Internet]. Laman: https://solutech.id/2019/07/18/5marketplace-terbaik-di-indonesia-pada-2019/. [diakses 18 Mei 2020].
[7] Talib, Ramzan et al. 2016. Text Mining: Techniques, Applications and Issues. International Journal of Advanced Computer Science and Applications. 7. 10.14569/IJACSA.2016.071153.
[8] Zhang, Lei et al. 2011. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.