

FAISAL PARSAKH NURSYAMSYI -
KLASIFIKASI SENTIMEN
TERHADAP APLIKASI IDENTITAS
KEPENDUDUKAN DIGITAL
MENGUNAKAN ALGORITMA
NAÏVE BAYES DAN SVM

by Layanan Perpustakaan UHAMKA

Submission date: 22-Dec-2023 10:32AM (UTC+0700)

Submission ID: 2263938524

File name: URSYAMSYI_TEKNIK_INFORMATIKA_TURNITIN_KE-2_-_Ujang_Bengkels.docx (931.44K)

Word count: 4289

Character count: 27631

KLASIFIKASI SENTIMEN TERHADAP APLIKASI IDENTITAS KEPENDUDUKAN DIGITAL MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SVM

Faisal Parsakh Nursyamsyi¹, Firman Noor Hasan²

6

¹ Fakultas Teknologi Industri dan Informatika, Teknik Informatika, Universitas Muhammadiyah Prof. DR. Hamka, Jakarta, Indonesia

Email: ¹denfaisal.270@gmail.com, ²firman.noorhasan@uhamka.ac.id

Email Penulis Korespondensi: firman.noorhasan@uhamka.ac.id

Abstrak– Kementerian Dalam Negeri mengumumkan penerapan sebuah aplikasi untuk mengikuti perkembangan teknologi dan informasi sekaligus memanfaatkan digitalisasi dalam upaya meningkatkan efisiensi pelayanan publik kepada masyarakat dalam hal data kependudukan dengan nama *Identitas Kependudukan Digital (IKD)*. Aplikasi *Identitas Kependudukan Digital* akan merepresentasikan informasi data identitas dalam bentuk digital. Sudah ada 5 juta pengguna yang telah mengunduh aplikasi dan sekitar 33 ribu orang yang memberikan ulasan mereka terhadap kepuasan setelah menggunakan aplikasi. Namun, penerapan aplikasi *Identitas Kependudukan Digital* masih menjadi pro dan kontra. Beragam sekali sentimen pengguna yang diberikan berdasarkan ulasan terhadap kepuasan mereka setelah menggunakan aplikasi tersebut. Dari permasalahan tersebut, peneliti mencoba melakukan penelitian sentimen klasifikasi dengan algoritma *Naïve Bayes* dan *Support Vector Machine* menggunakan *Rapid Miner Studio* untuk mengetahui respon masyarakat terhadap kepuasan mereka pada aplikasi *Identitas Kependudukan Digital* dengan menarik data ulasan pada aplikasi *Identitas Kependudukan Digital*. Sentimen pada data ulasan akan terbagi menjadi sentimen positif dan sentimen negatif. Tahapan yang dilakukan pada proses penelitian yaitu pengumpulan data, pelabelan data, pembersihan data, pembobotan kata dengan *TF-IDF*, *SMOTE Upsampling*, dan *Cross Validation* untuk menaung kedua algoritma klasifikasi, *apply model*, dan *performance*. Hasil proses analisis yang telah dilakukan, algoritma *Support Vector Machine* memiliki performa cukup baik dengan nilai akurasi 80,46%, presisi 0,73, dan recall 0,96%. Sedangkan *Naïve Bayes* memiliki nilai akurasi 80,22%, presisi 0,73 dan recall 0,94. Kedua algoritma dapat melakukan proses pengklasifikasian dengan baik terhadap proses analisis pada aplikasi *Identitas Kependudukan Digital*.

Kata Kunci: Naïve Bayes; SVM; Identitas Kependudukan Digital; Klasifikasi Sentimen; Aplikasi

Abstract– The Ministry of Home Affairs announced the implementation of an application to keep up with technological and information developments while utilizing digitalization in an effort to increase the efficiency of public services to the community in terms of population data under the name *Digital Population Identity (IKD)*. The *Digital Population Identity Application* will represent identity data information in digital form. There have been 5 million users who have downloaded the application and around 33 thousand people have provided reviews regarding their satisfaction after using the application. However, implementing the *Digital Population Identity application* still has pros and cons. There are various user sentiments given based on reviews regarding their satisfaction after using the application. From this problem. The researcher tried to conduct sentiment classification research using the *Naïve Bayes* algorithm and *Support Vector Machine* using *Rapid Miner Studio* to determine the public's response to their satisfaction with the *Digital Population Identity application* by pulling review data on the *Digital Population Identity application*. Sentiment in review data will be divided into positive sentiment and negative sentiment. The stages carried out in the research process are data collection, data labeling, data cleaning, word weighting with *TF-IDF*, *SMOTE Upsampling*, and *Cross Validation* to accommodate the two classification algorithms, *apply model*, and *performance*. As a result of the analysis process that has been carried out, the *Support Vector Machine* algorithm has quite good performance with an accuracy value of 80.46%, precision of 0.73, and recall of 0.96%. Meanwhile, *Naïve Bayes* has an accuracy value of 80.22%, precision of 0.73 and recall of 0.94. Both algorithms can carry out the classification process well in the analysis process in the *Digital Population Identity application*.

Keywords: Naïve Bayes; SVM; Identitas Kependudukan Digital; Sentiment Classification; Application

31

1. PENDAHULUAN

Dengan kemajuan teknologi informasi saat ini telah merubah kebiasaan-kebiasaan yang dilakukan manusia di setiap harinya. Perkembangan dari teknologi informasi dan komunikasi menjadikan aktivitas yang dilakukan manusia menjadi semakin efektif dan efisien [1][2]. Teknologi informasi juga telah menjadi bagian dari berbagai macam aspek seperti aspek sosial budaya, pertahanan, ekonomi, maupun kemasyarakatan. Pada sektor pemerintahan, teknologi informasi tidak bisa terpisahkan. Hal ini menjadi pendorong pemerintah untuk membuat sebuah inovasi, termasuk dalam proses administrasi kependudukan secara digitalisasi [3].

Dalam upaya digitalisasi proses administrasi data kependudukan, Direktorat Jenderal Kependudukan dan Pencatatan Sipil membangun sebuah aplikasi *mobile* yang dapat membantu memudahkan masyarakat dalam mencari informasi kependudukan dengan nama *Identitas Kependudukan Digital (IKD)* [4]. Aplikasi *Identitas Kependudukan Digital* bertujuan untuk menerapkan perkembangan teknologi informasi dan komunikasi saat ini menuju era digitalisasi



terhadap data kependudukan masyarakat. Aplikasi Identitas Kependudukan Digital dibangun dengan pendekatan teknologi digital dan terintegrasi dengan sistem mekanisasi pengolahan data. Pemerintah berharap dapat memberikan pelayanan publik yang mudah, cepat, dan tepat sasaran pada proses administrasi kependudukan di masyarakat [5]. Pro dan kontra terhadap aplikasi Identitas Kependudukan Digital masih menjadi perbincangan di kalangan masyarakat. Adapun, faktor lain yang bisa mempengaruhi aplikasi Identitas Kependudukan Digital yaitu tingkat keamanan data pribadi masyarakat Indonesia masih tergolong minim karena di tahun 2022 terjadi isu kebocoran data pendaftar kartu identitas ponsel atau kartu sim dengan total kebocoran sekitar 1,3 miliar data. Hal tersebut membuat beberapa masyarakat Indonesia masih merasa takut untuk memulai mendigitalisasi identitas miliknya [6].

Berdasarkan informasi yang ada pada aplikasi *Google Playstore*, sudah ada 5 juta pengunduh aplikasi Identitas Kependudukan Digital dan hampir 33 ribu ulasan diberikan oleh pengguna terhadap kepuasan mereka dalam menggunakan aplikasi Identitas Kependudukan Digital ini. Banyak sekali ulasan yang diberikan oleh pengguna aplikasi IKD baik sentimen positif ataupun negatif setelah mereka mencoba menggunakan aplikasi Identitas Kependudukan Digital.

Analisis Sentimen adalah suatu proses analisis dengan melakukan ekstraksi, memahami dan memproses data terhadap kata pada sebuah teks data berdasarkan opini untuk memahami makna dan informasi yang terkandung pada teks data. Analisis sentimen dilakukan untuk mengetahui kecenderungan dari suatu opini atau pendapat terhadap sebuah objek permasalahan berdasarkan sudut pandang orang lain dengan mempertimbangkan apakah pandangan tersebut bersifat positif atau negatif dari suatu permasalahan [7].

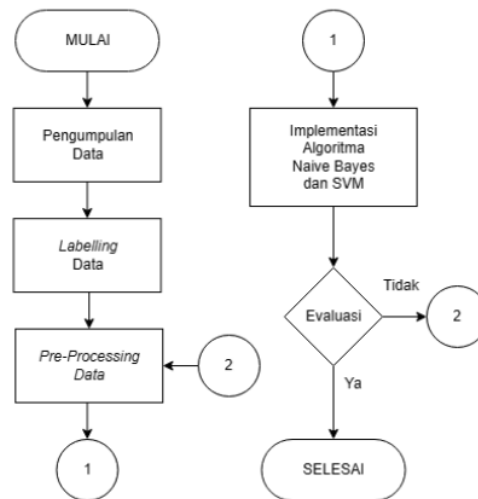
Naïve Bayes dan *Support Vector Machine* (SVM) adalah kedua metode *machine learning* yang sangat populer digunakan dalam proses analisis dengan keunggulan pada masing-masing metode [8][9]. Klasifikasi Naïve Bayes memiliki konsep model dengan menghitung probabilitas pada suatu parameter dan menentukan nilai yang paling baik menggunakan nilai dari parameter lain [10][11][8]. *Support Vector Machine* memiliki model mencari garis batas *hyperplane* (batas keputusan) terbaik yang berfungsi untuk memisahkan menjadi dua kelas yaitu negatif dan positif [12][13].

Berdasarkan penelitian terdahulu, Algoritma Naïve Bayes dan *Support Vector Machine* (SVM) telah digunakan sebagai metode klasifikasi pada analisis sentimen. Penelitian [14] yang dilakukan Yerik Afrianto Singgalen melakukan proses analisis sentimen dengan melakukan *scraping* data menggunakan *webharvy* dengan total data 3850 data dan diolah menggunakan *software RapidMiner* dengan melakukan *pre-processing* (*tokenize, transform cases, stopwords, dan Stem*), *processing* (Naïve Bayes), dan *performance evaluation* (*accuracy, recall, precision, area under curve*) dengan output akurasi sebesar 96,36%, *precision* sebesar 93,23%, dan *recall* 100% [14]. Adapun penelitian [15] yang dilakukan oleh peneliti Dianati Duei Putri, dkk. melakukan proses analisis sentimen dengan melakukan *crawling* data *tweet* menggunakan *python* dengan hasil total *crawling* sebesar 1546 data. Peneliti mulai melakukan *pre-processing* data, *labelling*, dan *processing* Naïve Bayes) didapat hasil tingkat akurasi sebesar 80% dengan nilai rata-rata *precision* positif 75%, netral 79%, dan negatif 82% [15]. Selain itu, dari penelitian [16] yang dilakukan oleh peneliti Herwinsyah dan Arita Witanti, peneliti melakukan *crawling* data sebesar 14.000 data dan dilakukan *pre-processing* (*cleansing, use folding, tokenizing, filtering stopword, stemming*), *labelling*, dan penerapan *Support Vector Machine* (SVM) dengan perbandingan 80% untuk data *training* dan 20% untuk data *testing* dengan hasil akurasi 89%, *f1-score* 93%, *precision score* 88%, dan *recall* 99% [16].

Dalam penelitian ini, metode klasifikasi Naïve Bayes dan *Support Vector Machine* (SVM) akan digunakan terhadap proses analisis sentimen pada aplikasi Identitas Kependudukan Digital. Seluruh penarikan data ulasan pada aplikasi Identitas Kependudukan Digital dilakukan dengan metode *web scraping* dengan menggunakan *Google Colab* yang kemudian disimpan kedalam dokumen *Microsoft Excel* dengan format *.csv*. Total data yang berhasil ditarik pada ulasan aplikasi Identitas Kependudukan Digital yaitu 2000 data dengan rentang waktu Bulan Juli 2023 hingga Oktober 2023. Selanjutnya data tersebut akan dilakukan proses *labelling, cleaning, pre-processing, bobot TF-IDF*, dan penerapan algoritma klasifikasi Naïve Bayes dan *Support Vector Machine* (SVM). Penelitian ini dilakukan dengan tujuan untuk mendapatkan informasi persepsi yang muncul dari pengguna aplikasi Identitas Kependudukan Digital berdasarkan kepuasan mereka setelah menggunakan aplikasi IKD. Pembagian terhadap sentimen dibagi menjadi sentimen positif dan negatif. Adapun tujuan lain yaitu sebagai pengujian terhadap tingkat akurasi, *precision*, dan *recall* dari algoritma Naïve Bayes dan *Support Vector Machine* (SVM) terhadap data ulasan aplikasi Identitas Kependudukan Digital.

2. METODOLOGI PENELITIAN

Proses yang dilakukan peneliti pada penelitian ini adalah menerapkan algoritma Naïve Bayes dan SVM (*Support Vector Machine*) untuk mengklasifikasi opini kepuasan pengguna aplikasi berdasarkan ulasan yang diberikan dari aplikasi Identitas Kependudukan Digital sehingga mendapatkan hasil tingkat akurasi dari sentimen berdasarkan kedua algoritma tersebut. Dalam proses pengumpulan data, data ulasan didapat dengan melakukan *Web Scraping* melalui *website Google Colab* dan proses klasifikasi data menggunakan *tools RapidMiner* Adapun alur penelitian yang dilakukan peneliti dapat dijelaskan melalui diagram alir pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

2.1 Pengumpulan Data

Proses yang dilakukan dalam pengumpulan data ulasan dikumpulkan dengan menggunakan proses *Web Scraping*. Proses *Web Scraping* dilakukan melalui *website Google Colab* dengan memasukkan *apps id* dari aplikasi Identitas Kependudukan Digital di *Google Playstore* dan menentukan jumlah data yang akan dikumpulkan sesuai yang diinginkan. Seluruh data yang telah dikumpulkan dalam proses *Web Scraping* kemudian disimpan dengan format dokumen *csv* untuk dilakukan ke tahap proses selanjutnya.



Gambar 2. Proses pengumpulan data ulasan

2.2 Labelling Data

Setelah proses pengumpulan dataset telah berhasil dilakukan, seluruh dataset dari proses sebelumnya harus dilakukan proses pelabelan dengan menganalisis setiap ulasan dan diproses secara manual pada file *.csv* menggunakan *Google Spreadsheet*. Proses *labelling* akan dibagi menjadi dua sentimen pada setiap ulasan yaitu sentimen positif dan negatif. Proses ini dilakukan untuk merepresentasi suatu objek data dan membantu mempelajari pola identifikasi data [17].

2.3 Pre-Processing Data

Tahap *pre-processing* adalah tahapan untuk menyiapkan suatu dataset dengan cara melakukan pembersihan dan mengatur suatu isi data dari dataset mentah menjadi dataset bersih dan teratur. *Pre-Processing* dilakukan dengan menggunakan dataset yang sebelumnya telah diberi label menggunakan *software RapidMiner*. Proses ini dilakukan dengan tujuan agar dataset dapat diproses dengan baik pada saat proses pengklasifikasian [18]. Pada proses *pre-processing* terdiri dari beberapa tahap diantaranya *cleansing*, *case folding*, *tokenize*, *spelling normalization*, *filter stopwords*, *filter tokens (by length)*, dan *stemming*. *Cleansing* merupakan tahap untuk membersihkan kata yang tidak dibutuhkan pada dataset seperti *emoji*, simbol tanda baca, angka [9]. *Case folding* adalah tahapan untuk merubah keseluruhan huruf pada dataset menjadi huruf kecil [17]. *Tokenize* yang merupakan tahap untuk memecah sebuah kalimat teks menjadi kata per-kata [19]. *Spelling Normalization* adalah tahapan untuk merubah kata-kata singkat, salah penulisan kata, maupun kata yang memiliki huruf berlebih pada suatu text data menjadi kata yang sebenarnya. Contoh "yg", "tdk", "penyebabnya", "bagusssssss" menjadi "yang", "tidak", "penyebabnya", "bagus" [6]. *Filter stopwords* adalah tahapan untuk menghapus kata-kata yang kurang memiliki arti berdasarkan kamus *list stopwords* bahasa Indonesia [20]. *Filter tokens (by length)* merupakan proses membatasi kata yang digunakan dengan ketentuan minimal dan maksimal dari jumlah kata [17]. *Stemming* adalah tahap untuk mengubah suatu kata imbuhan menjadi kata dasar [7].

2.4 Implementasi Algoritma Naïve Bayes

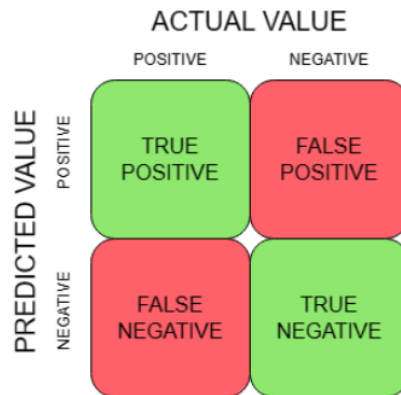
Klasifikasi Naïve Bayes sebuah proses pengklasifikasi dengan konsep model menghitung probabilitas pada suatu parameter dan menentukan nilai yang paling baik menggunakan nilai dari parameter lain. Pada metode klasifikasi Naïve Bayes, seluruh parameter nilai akan memberikan kontribusi dalam pengambilan keputusan karena Naïve Bayes memiliki asumsi independensi yang kuat pada masing-masing parameter [10][11][8]. Penerapan Naïve Bayes dilakukan untuk melakukan pembobotan pada setiap ulasan yang ada pada dataset setelah proses *pre-processing* [17].

2.5 Implementasi Algoritma Support Vector Machine (SVM)

Support Vector Machine adalah salah satu *supervised learning* terhadap teks dengan melakukan pengklasifikasian data dengan performa akurasi yang cukup baik [21]. SVM merupakan salah satu model pengklasifikasian dengan konsep *hyperplane* yang fungsi memisahkan antar kelas sehingga dapat mengoptimalkan margin antar kelas data [21], [22]. Algoritma SVM merupakan algoritma *machine learning* yang banyak digunakan pada suatu penelitian analisis seperti analisis sentimen dan menjadi algoritma yang populer oleh para peneliti [23].

2.6 Evaluasi

Tahap evaluasi adalah tahap terakhir untuk mengetahui apakah hasil proses analisis data terhadap setiap algoritma telah mencapai hasil yang baik. Pada tahap evaluasi, dilakukan pengukuran terhadap *accuracy*, *recall*, dan *precision* dengan melakukan perhitungan. Proses perhitungan pada *accuracy*, *recall*, dan *precision* dihitung berdasarkan hasil *confusion matrix*.



Gambar 3. Confusion Matrix

Accuracy merupakan proses menghitung dan mengukur tingkat keakuratan pada model klasifikasi terhadap prediksi yang benar. *Recall* adalah proses menghitung dan mengukur banyaknya kelas positif yang berhasil di klasifikasi oleh model. *Precision* yaitu proses menghitung dan mengukur banyaknya prediksi yang sebenarnya benar [17][24].

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{1}$$

$$\text{Recall} = \frac{TP}{(TP+FP)} \tag{2}$$

$$\text{Precision} = \frac{TP}{(TP+FN)} \tag{3}$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Sebelum melakukan analisis, perlu dilakukan pengumpulan dataset. Proses pengumpulan dataset dilakukan dengan metode *web scrapping* menggunakan *Google Collab*. Hasil dataset yang terkumpul pada proses *web scrapping* sejumlah 2000 data dengan rentang bulan Juli 2023 hingga Oktober 2023. Gambar. 4 merupakan proses dalam melakukan pengumpulan data menggunakan metode *web scrapping* pada *Google Collab*.

```

from google_play_scraper import Sort, reviews

result, continuation_token = reviews(
    'gov.dukcapil.mobile_id',
    lang='id',
    country='id',
    sort=Sort.NEWEST,
    count=2000,
    filter_score_with=None
)
    
```

Gambar 4. Proses web scrapping menggunakan Google Collab

3.2 Labelling Data

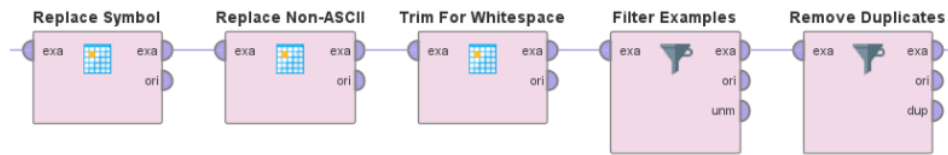
Dari hasil proses penarikan data yang dilakukan pada tahap sebelumnya, proses selanjutnya yaitu proses pelabelan **25** ada dataset. Proses labelling data dilakukan secara manual dan dilakukan oleh tiga orang. Proses pelabelan pada **dataset terbagi menjadi dua** sentimen yaitu **sentimen positif dan** sentiment **negatif** dengan penulisan label “Positive” untuk positif dan “Negative” untuk negatif. Proses pelabelan dilakukan menggunakan *Spreadsheet* dengan memahami makna dan maksud pada setiap ulasan untuk mengetahui apakah ulasan tersebut termasuk sentimen positif atau sentimen negatif. Hasil proses pelabelan yang telah dilakukan, sentimen negatif memiliki total 1049 dan sentimen positif 913. Gambar. 5 adalah hasil proses labelling yang dilakukan pada dataset

userName	content	sentiment
Djunaidi	Bermanfaat dan berguna	POSITIVE
suebta halimah	Buka dokumen loading terus tidak ada menu log out kalau sudah masuk hanya ada menu masuk dan menu daftar tidak ada saat suc	NEGATIVE
gunawan _c	RIBET RUWET ngapain juga harus ke dukcapil? TOLOL	NEGATIVE
SHON HAJI	Sangat oke	POSITIVE
Acimpathul Husli	Pelayanan di ikd kenapa terjadi kesalaha terus di saat ingin perubahan data	NEGATIVE
yabes samosir si mantap...	Indonesia maju... luar biasa	POSITIVE
The Saputra	Masih banyak kekurangan dan harus dibenahi lagi,akte kelahiran ga ada,gak muncul,gabisa dibuka,,harus belajar lagi bikin aplikasi r	NEGATIVE
suka mas	Scan barcode nya sangat susah,alasannya hp nya bermasalah,padahal sudah android 12,sungguh kurang bagus,dan saya kira kura	NEGATIVE
Eko Verizqo	Alhamdulillah mantap	POSITIVE
Fahmi Finishtriar	Aplikasi kntll, boomer sok sok an melek teknologi malah makin ribet.	NEGATIVE
muh Yusuf nabil	Ok	POSITIVE
Putra Adi	Gak ada pembaruan sama sekali Tah	NEGATIVE
Reo Zander	Susah loginnya muter doang,padahal sudah coba pake WiFi/data , lawak bnget	NEGATIVE
Mochammad Ma	Tidak bisa memuat koneksi padahal punya kuota	NEGATIVE
Taufiq qur Rohmi	Semoga bermanfaat dan aplikasi lebih baik	POSITIVE
Junaidi Upara	Mantap	POSITIVE
Parepare TV	Aplikasi apa ini	NEGATIVE
Waitina Sulu	Mantap	POSITIVE
Dandy AML	Sangat berguna untuk memenuhi ruang ponsel anda Kalo bisa unduh 3x sangat berguna banget parah	NEGATIVE
Soraya Attamimi	Lebih efisien dan terjaga rapi	POSITIVE
Engkos Koswara	Aplikasi kurang mendukung!	NEGATIVE
Anton Setiawan	Setahun ngurus E-KTP RUSAK sudah datang dari kelurahan&kecamatan sampai ke kantor walikota disuruh lewat aplikasi tapi Samp	NEGATIVE
Dwi Hartanto	saran saya kalau memang mempermudah alangkah lebih baiknya...untuk scan qr verifikasi bisa dilakukan secara online juga tanpa h	POSITIVE
Mas Wisdom	Nge bug, jaringan udah 5g, pake wifi udah stabil juga masih aja masalah di koneksi.	NEGATIVE
Alexi Achmad	IT kacangan....cpek brxx insta cba mulu	NEGATIVE
Miskin OFFICIAL	Ngabisin dana, gabisa didonlod,pindah daerah pada ngga tau apknya kintil la	NEGATIVE
Junior	Banyak bug nya	NEGATIVE
Usep Hendryana	Susah banget kebuka dokumen nya dari mulai daftar sampai sekarang gak kebuka	NEGATIVE

Gambar 5. Hasil proses labelling data

3.3 Pre-Processing Data

Tahapan *pre-processing* data dilakukan menggunakan *software* yaitu *RapidMiner Studio* untuk menyiapkan dataset dengan dilakukan beberapa tahapan. Tahap pertama diawali dengan melakukan *cleansing* pada dataset. Pada Gambar.6 proses *cleansing* dilakukan dengan tujuan untuk membersihkan setiap teks ulasan yang memiliki *emoji*, simbol tanda baca, angka, membersihkan *whitespace* berlebih, dan membersihkan teks ulasan berulang supaya dapat mengoptimalkan proses pengklasifikasian.

Gambar 6. Proses *cleansing* pada datasetTabel 1. Hasil *Cleansing*

Sebelum <i>Cleansing</i>	Setelah <i>Cleansing</i>
Susah amat buka nya Trus buat apa di buat... Kalau gak bisa di buka... Tak ada manfaatnya 🙄🙄🙄	Susah amat bukanya Trus buat apa di buat Kalau gak bisa di buka Tak ada manfaatnya
Proses nya cepat dan pelayanannya pun sangat baik sekali. Saya sangat terbantu sekali dengan Aplikasi IKD ini. Terima kasih... 🙏🙏🙏	Proses nya cepat dan pelayanannya pun sangat baik sekali Saya sangat terbantu sekali dengan Aplikasi IKD ini Terima kasih
Proses Login atau masuk ke menu Lama sekali, bahkan sudah pakai jaringan internet yg kencang saja tetap lama, notgod 😞	Proses Login atau masuk ke menu Lama sekali bahkan sudah pakai jaringan internet yg kencang saja tetap lama notgod

Adapun beberapa tahapan yang dilakukan pada pre-processing yaitu tahap *Case Folding*, *Tokenize*, *Spelling Normalization*, *Filter Stopwords*, *Filter Tokens (by Length)*, dan *Stemming*. Pada proses ini, seluruh tahapan dikumpulkan pada satu operator yaitu *process document from data*. Gambar 7. merupakan seluruh tahapan yang dilakukan untuk menyiapkan sebuah dataset.

Gambar 7. Operator pada *pre-processing*Tabel 2. Hasil *Case Folding*

Sebelum <i>Case Folding</i>	Setelah <i>Case Folding</i>
Susah amat bukanya Trus buat apa di buat Kalau gak bisa di buka Tak ada manfaatnya	susah amat bukanya trus buat apa di buat kalau gak bisa di buka tak ada manfaatnya
Proses nya cepat dan pelayanannya pun sangat baik sekali Saya sangat terbantu sekali dengan Aplikasi IKD ini Terima kasih	proses nya cepat dan pelayanannya pun sangat baik sekali saya sangat terbantu sekali dengan aplikasi ikd ini terima kasih
Proses Login atau masuk ke menu Lama sekali bahkan sudah pakai jaringan internet yg kencang saja tetap lama notgod	proses login atau masuk ke menu lama sekali bahkan sudah pakai jaringan internet yg kencang saja tetap lama notgod

Case folding merupakan tahap untuk merubah keseluruhan huruf pada dataset menjadi huruf kecil [17]. Proses ini dilakukan agar memudahkan proses pencarian pada setiap kata pada tahap *tokenize*. Tabel 2. Adalah contoh dari hasil proses *case folding* sebelum dan sesudah dari proses yang sudah dilakukan.

Tabel 3. Hasil *Tokenize*

Sebelum <i>Tokenize</i>	Setelah <i>Tokenize</i>
semoga bermanfaat dan aplikasi lebih baik	['semoga', 'bermanfaat', 'dan', 'aplikasi', 'lebih', 'baik']
gak bisa membuka dokumen muter terus tolong diperbaiki lagi	['gak', 'bisa', 'membuka', 'dokumen', 'muter', 'terus', 'tolong', 'diperbaiki', 'lagi']
aplikasi aneh mau di pake malah ga kebuka force close terus	['aplikasi', 'aneh', 'mau', 'di', 'pake', 'malah', 'ga', 'kebuka', 'force', 'close', 'terus']

Setelah melewati tahap *case folding*, dilanjutkan pada tahap *tokenize*. *Tokenize* yaitu tahap untuk memecah sebuah kalimat teks menjadi kata per-kata [19]. Pada *tokenize* pertama, proses ini dilakukan guna memecah kata yang sebelumnya telah di proses pada tahap *case folding* sehingga mempermudah proses yang akan dilakukan pada tahap selanjutnya. Berikut Tabel 3. merupakan contoh dari hasil proses *tokenize* sebelum dan sesudah pada penelitian ini.

Tabel 4. Hasil *Spelling Normalization*

Sebelum <i>Spelling Normalization</i>	Setelah <i>Spelling Normalization</i>
tolong apknya di fix lagi karna ada masalah setelah selesai daftar dan mengecek dokumen loading nya sangat lama sekalipun pake wifi hingga saat ini saya blm bisa membuka dokumen ktp saya tolong untuk developer nya	tolong aplikasinya di fix lagi karena ada masalah setelah selesai daftar dan mengecek dokumen loading nya sangat lama sekalipun pake wifi hingga saat ini saya belum bisa membuka dokumen ktp saya tolong untuk developer nya
mohon tingkatkan lagi scan barcode tdk bsa lihat dokumen tdk bisa gagal bsa d blg apk gak guna	mohon tingkatkan lagi scan barcode tidak bisa lihat dokumen tidak bisa gagal bisa di bilang aplikasi gak berguna
skrg sering terjadi error dokument seperti ktp dan kk tdk bisa di lihat sama tambahkan sedikit coba tambahkan fitur cetak pdf cetak dokumen	sekarang sering terjadi error dokumen seperti ktp dan kk tidak bisa di lihat sama tambahkan sedikit coba tambahkan fitur cetak pdf cetak dokumen

Tahap selanjutnya yaitu tahap *spelling normalization*. *Spelling Normalization* dilakukan dengan menggunakan operator *Replace Tokens* pada *RapidMiner* tujuan untuk merubah kata-kata singkat, salah penulisan kata, maupun kata yang memiliki huruf berlebih pada suatu text data menjadi kata yang sebenarnya [6]. Tabel 4. merupakan contoh dari hasil proses *spelling normalization* sebelum dan sesudah pada penelitian ini.

Tabel 5. Hasil *Filter Stopwords*

Sebelum <i>Filter Stopwords</i>	Setelah <i>Filter Stopwords</i>
alhamdulillah sangat ngebanu untuk aplikasinya terima kasih	alhamdulillah ngebanu aplikasinya terima kasih
seharusnya membuat aplikasi yang memudahkan masa untuk membuat akun harus scan kode qr ke kantor dukcapil	aplikasi memudahkan akun scan kode qr kantor dukcapil
katanya serba online tapi tetap aja ribet kalau harus mendatangi dukcapil setempat gimana dengan orang yang kerja diluar kota	serba online aja ribet dukcapil gimana orang kerja diluar kota

Selanjutnya, dilakukan *tokenize* kembali dengan tujuan untuk memperbaharui pemecahan kata per-kata setelah dilakukan *spelling normalization*. Setelah dilakukan tahap *tokenize* kedua, dilanjutkan tahap *filter stopwords* dengan tujuan untuk menghapus kata-kata yang kurang memiliki arti berdasarkan kamus *list stopwords* bahasa Indonesia [20]. Proses *filter stopwords* menggunakan sebuah file kamus bahasa indonesia yang didapat pada *website* www.kaggle.com. Tabel 5. adalah beberapa contoh hasil proses *filter stopwords* yang dilakukan pada penelitian ini.

Tabel 6. Hasil *Filter Tokens (by Length)*

Sebelum <i>Filter Tokens (by Length)</i>	Setelah <i>Filter Tokens (by Length)</i>
proses nya cepat pelayanannya terbantu aplikasi ikd terima kasih	Proses cepat pelayanannya terbantu aplikasi terima kasih
bagus cuman ditingkatkan kk ktp semoga kedepannya didownload	bagus cuman ditingkatkan semoga kedepannya didownload
daftar cek nik ktp repot suruh scan barcode dukcapil scan barcode nya bingung darimana dah	daftar repot suruh scan barcode dukcapil scan barcode bingung darimana

Setelah tahap *filter stopwords* telah dilakukan, selanjutnya akan dilakukan proses *filter tokens (by Length)*. Proses ini dilakukan untuk membatasi kata yang digunakan dengan ketentuan minimal dan maksimal dari jumlah kata [17]. Pada penelitian ini, jumlah kata yang akan digunakan yaitu 4 kata untuk ketentuan minimal kata dan 25 kata untuk ketentuan maksimal. Berikut Tabel 6. adalah beberapa contoh hasil dari proses *filter tokens (by length)* yang dilakukan di penelitian ini.

Tabel 7. Hasil *Filter Stemming*

Sebelum <i>Stemming</i>	Setelah <i>Stemming</i>
dokumen dibuka harap diperbaiki	dokumen buka harap perbaiki
aplikasi bagus mempermudah	aplikasi bagus mudah
mengapresiasi aplikasi bagus bermanfaat data kependudukan terima kasih dukcapil membantu mempermudah urusan data kependudukan	apresiasi aplikasi bagus manfaat data penduduk terima kasih dukcapil bantu mudah urus data penduduk

Data yang sudah di proses pada filter stopwords, akan dilakukan proses yaitu *stemming*. *Stemming* dilakukan dengan tujuan untuk mengubah kata-kata yang memiliki sebelumnya berbentuk imbuhan menjadi bentuk kata dasar [7]. Proses

ini memerlukan file dengan ekstensi .txt untuk membuat formula *Regex* agar dapat menjalankan proses *stemming*. Berikut Tabel 7. adalah beberapa contoh hasil pada proses *stemming* yang dilakukan pada penelitian ini.

3.4 Pembobotan TF-IDF

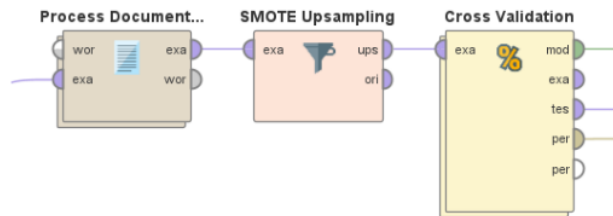
Setelah seluruh proses *pre-processing* dilakukan, seluruh kata yang ada pada dataset akan dilakukan pembobotan oleh *TF-IDF*. *TF-IDF* memiliki fungsi diantaranya *TF (Term Frequency)* untuk menghitung seberapa sering kemunculan pada suatu kata berdasarkan bobot kata tersebut dan *IDF (Inverse Document Frequency)* untuk mengukur bobot pada seluruh kata dimana jika kata tersebut jarang muncul maka bobot kata tersebut tinggi dan jika kata tersebut sering muncul maka bobot kata tersebut rendah [7][17].

text	abal	abis	absurd	aceh	ada	admin
mantap	0	0	0	0	0	0
dokumen buka harap perbaiki	0	0	0	0	0	0
buka dokumen kelurahan kasih fisik alasan bla	0	0.384	0	0	0	0
tooop	0	0	0	0	0	0
jelek	0	0	0	0	0	0
alangkah cepat	0	0	0	0	0	0
download	0	0	0	0	0	0
download aneh	0	0	0	0	0	0
aplikasi online daftar dukcapil bohong	0	0	0	0	0	0
sistem rusak product gagal	0	0	0	0	0	0
bagus salah menit	0	0	0	0	0	0
aplikasi mudah sulit	0	0	0	0	0	0
kacau dukcapil makassar salah kirim email kiri	0	0	0	0	0	0
loading sinyal stabil mohon perbaiki	0	0	0	0	0	0
dokumen buka	0	0	0	0	0	0

Gambar 8. Hasil proses *TF-IDF*

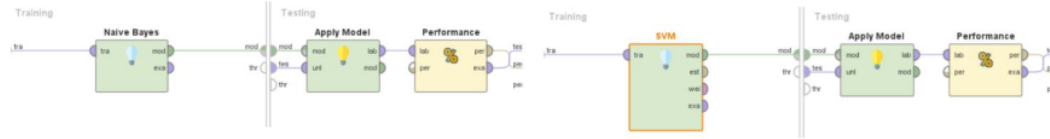
3.5 Implementasi Algoritma

Setelah dataset telah berhasil diproses pada tahap *pre-processing*, maka proses selanjutnya yaitu proses implementasi algoritma. Namun, sebelum memasuki proses implementasi algoritma, perlu menambahkan salah satu proses sebelum melakukan klasifikasi yaitu *Synthetic Minority Oversampling Technique (SMOTE Upsampling)*. *SMOTE Upsampling* adalah salah satu proses dengan tujuan untuk memodifikasi ketidakseimbangan pada dataset dengan melakukan pembuatan sintetik baru dari minoritas kelas supaya meningkatkan kinerja pada model klasifikasi [14]. Lalu pada penelitian ini, proses penerapan klasifikasi menggunakan operator *Cross Validation* dengan nilai *k-fold* yaitu 10.



Gambar 9. Penggunaan operator *SMOTE Upsampling* dan *Cross Validation*

Operator *Cross Validation* akan menampung beberapa operator lainnya yakni *Naïve Bayes* dan *SVM*, *Apply Model*, dan *Performance*. *Apply Model* bertujuan untuk membuat pengklasifikasian baru terhadap sebuah data yang belum dipelajari oleh metode klasifikasi yang digunakan. *Apply Model* akan melakukan *input* data, kemudian akan memperoleh sebuah *output* dari hasil prediksi terhadap metode klasifikasi. *Performance* digunakan untuk mengetahui seberapa baik proses yang dilakukan pada model klasifikasi.



Gambar 10. Cross Validation pada Naïve Bayes dan SVM

3.6 Evaluasi

Proses evaluasi adalah proses untuk menghitung hasil pada setiap proses pengklasifikasian yakni Naïve Bayes dan SVM. Pada proses klasifikasi Naïve Bayes, berdasarkan hasil *confusion matrix* didapat tingkat *accuracy* mencapai 80,22% dengan hasil 978 pada *True Positive*, 689 pada *True Negative*, 350 pada *False Positive*, dan 61 pada *False Negative*.

Tabel 8. Hasil klasifikasi Naïve Bayes

accuracy : 80,22% +/- 5,10%			
	true Positive	true Negative	class precision
pred. Positive	978	350	73,64%
pred. Negative	61	689	91,87%
class recall	94,13%	66,31%	

Adapun hasil perhitungan yang dilakukan untuk menghitung nilai *accuracy*, *recall*, dan *precision* dari nilai yang dihasilkan berdasarkan *confusion matrix* pada klasifikasi Naïve Bayes.

$$\begin{aligned} \text{Accuracy} &= \frac{(TP+TN)}{(TP+TN+FP+FN)} \\ &= \frac{(978+689)}{(978+689+350+61)} = \frac{1667}{2078} = 0,8 \end{aligned}$$

$$\begin{aligned} \text{Recall} &= \frac{TP}{(TP+FN)} \\ &= \frac{978}{(978+61)} = \frac{978}{1039} = 0,94 \end{aligned}$$

$$\begin{aligned} \text{Precision} &= \frac{TP}{(TP+FP)} \\ &= \frac{978}{(978+350)} = \frac{978}{1328} = 0,73 \end{aligned}$$

Pada proses klasifikasi *Support Vector Machine* (SVM), berdasarkan hasil *confusion matrix* didapat tingkat *accuracy* mencapai 80,46% dengan hasil 1001 pada *True Positive*, 671 pada *True Negative*, 368 pada *False Positive*, dan 38 pada *False Negative*.

Tabel 9. Hasil klasifikasi SVM

accuracy : 80,46% +/- 5,79%			
	true Positive	true Negative	class precision
pred. Positive	1001	368	73,12%
pred. Negative	38	671	94,64%
class recall	96,34%	64,58%	

Adapun hasil perhitungan untuk menghitung nilai *accuracy*, *recall*, dan *precision* dengan nilai yang dihasilkan berdasarkan *confusion matrix* pada klasifikasi *Support Vector Machine* (SVM).

$$\begin{aligned} \text{Accuracy} &= \frac{(TP+TN)}{(TP+TN+FP+FN)} \\ &= \frac{(1001+671)}{(1001+671+368+38)} = \frac{1672}{2078} = 0,8 \end{aligned}$$

$$\begin{aligned} \text{Recall} &= \frac{TP}{(TP+FN)} \\ &= \frac{1001}{(1001+38)} = \frac{1001}{1039} = 0,96 \end{aligned}$$

$$\begin{aligned} \text{Precision} &= \frac{TP}{(TP+FP)} \\ &= \frac{1001}{(1001+368)} = \frac{1001}{1369} = 0,73 \end{aligned}$$

Berdasarkan proses pengklasifikasian terhadap ulasan aplikasi Identitas Kependudukan Digital, didapat bahwa diantara kedua algoritma memiliki selisih persentase *accuracy* yang tidak jauh. *Support Vector Machine* mendapatkan nilai *accuracy* yaitu 81,01% dan algoritma Naïve Bayes mendapatkan nilai *accuracy* yaitu 80,81%. Adapun hasil nilai perhitungan untuk *accuracy*, *recall*, *precision* pada algoritma Naïve Bayes yaitu *accuracy* 0,8, *recall* 0,93, *precision* 0,74. Pada algoritma *Support Vector Machine* didapat yaitu *accuracy* 0,81, *recall* 0,93, *precision* 0,73.

3.7 Visualisasi WordCloud

Visualisasi *WordCloud* adalah sebuah hasil visual yang sudah dilakukan pada proses pengklasifikasian pada algoritma Naïve Bayes dan *Support Vector Machine* dengan tujuan untuk memperlihatkan kata yang sering muncul pada ulasan aplikasi Identitas Kependudukan Digital. Dari hasil visualisasi *wordcloud*, kata 'aplikasi' adalah kata yang paling sering digunakan pada ulasan aplikasi Identitas Kependudukan Digital karena memiliki total frekuensi yaitu 670. Adapun kata lainnya yang memiliki kaitan terhadap aplikasi Identitas Kependudukan Digital yaitu 'dukcapil', 'digital' dan 'dokumen'.



Gambar 12. Visualisasi *WordCloud*

4. KESIMPULAN

Dari proses analisis yang sudah dilakukan pada penelitian aplikasi Identitas Kependudukan Digital dapat disimpulkan bahwa aplikasi Identitas Kependudukan Digital masih tergolong belum cukup optimal dan masyarakat masih merasa kesulitan untuk menggunakan aplikasi tersebut. Ada sekitar 1049 sentimen negatif atau 53,46% dan 913 sentimen positif atau 46,54% terhadap ulasan aplikasi IKD. Dari hal tersebut, perlu adanya perbaikan untuk aplikasi IKD supaya bisa lebih optimal dan efisien dalam kinerja pelayanan data kependudukan masyarakat sehingga masyarakat mendapatkan kepuasan lebih terhadap penggunaan aplikasi IKD. Adapun hal lainnya yaitu dari total data 2000 data yang berhasil dikumpulkan, dilakukan proses pelabelan pada setiap ulasan, dan dilanjutkan pada proses *cleansing*, diperoleh 1601 dataset bersih yang akan digunakan pada tahap proses klasifikasi. Setelah dilakukan proses pembobotan *T20DF*, *SMOTE Upsampling*, dan *Cross Validation* untuk menjalankan proses klasifikasi terhadap kedua algoritma yaitu Naïve Bayes dan *Support Vector Machine*, didapat hasil akhir bahwa *Support Vector Machine* memiliki performa cukup baik dalam melakukan proses klasifikasi dibanding Naïve Bayes. *Support Vector Machine* mendapatkan nilai *accuracy* yaitu 80,46% dan Naïve Bayes mendapatkan nilai *accuracy* 80,22%. Adapun perhitungan *accuracy*, *recall*, dan *precision* diantara kedua algoritma klasifikasi. *Support Vector Machine* memiliki hasil *accuracy* yaitu 0,81, *recall* yaitu 0,93, dan *precision* yaitu 0,73. Sedangkan Naïve Bayes memiliki hasil *accuracy* yaitu 0,8, *recall* yaitu 0,93, *precision* yaitu 0,74. Namun dari hasil *accuracy* diantara kedua algoritma, performa antara kedua algoritma memiliki selisih yang tidak begitu jauh. Saran dari peneliti untuk mengembangkan penelitian yaitu menambahkan jumlah data dan melakukan proses analisis dengan menerapkan metode klasifikasi lainnya seperti *Deep Learning*, *Logistic Regression*, *K-NN*, dan lainnya dengan tujuan mengetahui perbandingan hasil terhadap masing-masing metode klasifikasi.

FAISAL PARSAKH NURSYAMSYI - KLASIFIKASI SENTIMEN TERHADAP APLIKASI IDENTITAS KEPENDUDUKAN DIGITAL MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SVM

ORIGINALITY REPORT

15%

SIMILARITY INDEX

13%

INTERNET SOURCES

9%

PUBLICATIONS

6%

STUDENT PAPERS

PRIMARY SOURCES

1 Submitted to Universitas Bina Darma 3%
Student Paper

2 repository.uhamka.ac.id 1%
Internet Source

3 repository.uin-suska.ac.id 1%
Internet Source

4 repository.mercubuana.ac.id 1%
Internet Source

5 vdocuments.net 1%
Internet Source

6 www.stmik-budidarma.ac.id 1%
Internet Source

7 www.researchgate.net 1%
Internet Source

8 ejurnal.teknokrat.ac.id 1%
Internet Source

ecampus.pelitabangsa.ac.id

9

Internet Source

<1 %

10

e-journal.undikma.ac.id

Internet Source

<1 %

11

Herlawati Herlawati, Rahmadya Trias Handayanto, Prima Dina Atika, Fata Nidaul Khasanah et al. "Analisis Sentimen Pada Situs Google Review dengan Naïve Bayes dan Support Vector Machine", Jurnal Komtika (Komputasi dan Informatika), 2021

Publication

<1 %

12

Rina Noviana, Isram Rasal. "PENERAPAN ALGORITMA NAIVE BAYES DAN SVM UNTUK ANALISIS SENTIMEN BOY BAND BTS PADA MEDIA SOSIAL TWITTER", Jurnal Teknik dan Science, 2023

Publication

<1 %

13

etheses.uin-malang.ac.id

Internet Source

<1 %

14

pontianak.tribunnews.com

Internet Source

<1 %

15

www.scilit.net

Internet Source

<1 %

16

Friska Aditia Indriyani, Ahmad Fauzi, Sutan Faisal. "Analisis sentimen aplikasi tiktok menggunakan algoritma naïve bayes dan

<1 %

support vector machine", TEKNOSAINS :
Jurnal Sains, Teknologi dan Informatika, 2023

Publication

17

Submitted to Universitas Komputer Indonesia

Student Paper

<1 %

18

Submitted to Universitas Negeri Semarang

Student Paper

<1 %

19

j-ilkominfo.org

Internet Source

<1 %

20

media.neliti.com

Internet Source

<1 %

21

Frizka Fitriana, Ema Utami, Hanif Al Fatta.
"Analisis Sentimen Opini Terhadap Vaksin
Covid - 19 pada Media Sosial Twitter
Menggunakan Support Vector Machine dan
Naive Bayes", Jurnal Komtika (Komputasi dan
Informatika), 2021

Publication

<1 %

22

Submitted to Universitas Brawijaya

Student Paper

<1 %

23

es.scribd.com

Internet Source

<1 %

24

www.dbpia.co.kr

Internet Source

<1 %

25

Fajar Romadoni, Yuyun Umaidah, Betha
Nurina Sari. "Text Mining Untuk Analisis

<1 %

Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma Support Vector Machine", Jurnal Sisfokom (Sistem Informasi dan Komputer), 2020

Publication

26

Imam Kurniawan, Ajib Susanto. "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019", Eksplora Informatika, 2019

Publication

<1 %

27

disdukcapil.dumaikota.go.id

Internet Source

<1 %

28

downloads.hindawi.com

Internet Source

<1 %

29

Hendrik Setiawan, Ema Utami, Sudarmawan Sudarmawan. "Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naive Bayes", Jurnal Komtika (Komputasi dan Informatika), 2021

Publication

<1 %

30

REYNALDA NABILA CIKANIA. "IMPLEMENTASI ALGORITMA NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE PADA KLASIFIKASI SENTIMEN REVIEW LAYANAN TELEMEDICINE HALODOC", Jambura Journal of Probability and Statistics, 2021

<1 %

31 **doku.pub** <1 %
Internet Source

32 **ejournal.nusamandiri.ac.id** <1 %
Internet Source

33 **ejurnal.seminar-id.com** <1 %
Internet Source

34 **jurnal.untan.ac.id** <1 %
Internet Source

35 **repository.its.ac.id** <1 %
Internet Source

36 **ojs.uajy.ac.id** <1 %
Internet Source

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off