 ISSN: 2684-8473

Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing

Hidayatul Ma'rifah[1], Aji Prasetya Wibawa[2], Muhammad Iqbal Akbar[3]

Jurusan Teknik Elektro, Universitas Negeri Malang, Malang, Indonesia
[1] marifahhidayatul@gmail.com; [2] aji.prasetya.ft@um.ac.id; [3] iqbal.elektro.um@gmail.com

INFORMASI ARTIKEL

Histori Artikel
Diterima : 8 Maret 2020
Direvisi : 22 Maret 2020
Diterbitkan : 4 April 2020

Kata Kunci:
Klasifikasi
Preprocessing
Text mining
Stemming
Stopwords removal
K-Nearest neighbour

ABSTRAK

Penelitian ini bertujuan untuk menemukan kombinasi dan urutan preprocessing dalam text mining yang paling maksimal untuk klasifikasi bidang jurnal berbahasa Indonesia berdasarkan judul dan abstraknya. Tahap-tahap preprocessing yang akan diterapkan terdiri dari case folding, stemming, stopwords removal, transformasi VSM (Vector Space Model), dan SMOTE. Namun, pengamatan tiap skenario berfokus pada stemming dan dua teknik stopwords removal, yaitu stopwords removal berbasis kamus, dan berbasis document frequency setelah melewati proses transformasi ke dalam bentuk VSM dengan pembobotan TF-IDF (Term Trequency–Inverse Document Frequency). Proses klasifikasi mengadopsi algoritma k-NN (K-Nearest Neighbour), yang menentukan kelas suatu data tes dengan melihat tetangga terdekatnya. Dalam penelitian ini, metrik untuk menemukan jarak tetangga terdekat adalah Cosine Similarity. Pengujian klasifikasi menggunakan 10-Fold Cross Validation untuk menghasilkan confusion matrix sebagai hasil akhir. Kinerja klasifikasi terbaik dicapai dengan persentase accuracy sebesar 72.91% dan precision mencapai 73,36%.

I. Pendahuluan

Saat ini permasalahan klasifikasi artikel berbahasa Indonesia dapat dipecahkan secara komputasi matematis menggunakan metode-metode dalam *text mining*, antara lain penerapan algoritma *K-Nearest Neighbor*, teknik *preprocessing* data, dan evaluasi untuk membentuk suatu sistem klasifikasi artikel yang sangat membantu penulis mentukan kategori dari artikel yang telah ditulisnya. Pemecahan masalah yang serupa telah dikembangkan sebelumnya dimana metode yang digunakan adalah pendekatan *cosine similarty*, dengan *preprocessing* yang terdiri dari *case folding*, *tokenizing* dan *stopword removal* berbasis *term frequency* (Nurfadila, 2019).

Penelitian-penelitian tersebut, tentu saja memberikan hasil yang mampu mengatasi problematika klasifikasi artikel. Akan tetapi, masih ada permasalahan lain yang menghambat, yaitu keberagaman teknik *preprocessing* untuk *text mining* berbahasa Indonesia yang belum semuanya diterapkan dan belum diketahui kombinasi seperti apa yang menghasilkan ketepatan klasifikasi secara maksimal. Oleh sebab itu, dalam penelitian ini diterapkan beberapa metode *preprocessing* tambahan serta dilakukan pengujian beberapa kombinasi dari metode-metode yang ada, kemudian dibandingkan satu sama lain agar tampak komposisi dan urutan *preprocessing* mana yang menghasilkan klasifikasi terbaik.

II. Material dan Metode

Pada *Dataset* dalam penelitian ini jurnal Fakultas Ekonomi yang diolah terlebih dahulu menggunakan beberapa tahap *preprocessing.* Penelitian terbagi menjadi 9 skenario yang masing-masing berisi urutan dan komposisi *preprocessing* yang berbeda.

Proses klasifikasi untuk kesembilan skenario ersebut adalah sama, yaitu mengadopsi algoritma *k-NN (K-Nearest Neighbour)*, yang menentukan kelas suatu data tes dengan melihat tetangga terdekatnya. Dan metrik pengukur jarak tetangga adalah *Cosine Similarity.* Pengujian klasifikasi menggunakan 10-Fold Cross Validation untuk menghasilkan *confusion matrix* sebagai hasil akhir.

A. Dataset Penelitian

Data dikumpulkan dari situs resmi eJournal Universitas Negeri Malang, bagian Jurnal Fakultas Ekonomi pada Maret 2019. Yang diambil hanya judul beserta abstrak jurnal, tanpa menyertakan nama penulis ataupun tanggal terbit.

Dokumen yang diambil berasal dari 4 kategori yaitu Ekonomi Bisnis, Pendidikan Akuntasi dan Bisnis, Pendidikan Bisnis dan Manajemen, serta Pendidikan Akuntasi. Atribut judul dan abstrak yang telah dikumpulkan akan menjadi data latih setelah memasuki preprocessing. Sementara atribut kategori menjadi label atau kelas data untuk supervised learning. Tabel 1 menampilkan daftar kelas dalam dataset, dengan contoh data data di masing-masing label pada Tabel 2.

Tabel 1. Daftar kelas dalam dalam dataset

Label	Jumlah data	Persentase
Ekonomi Bisnis	29	23.01%
Pendidikan Akuntansi dan Bisnis	31	24.60%
Pendidikan Bisnis dan Manajemen	30	23.81%
Pendidikan Akuntansi	36	28.57%
Total	126	100%

Tabel 2. Sampel data

Label	***Judul dan Abstrak***
Ekonomi Bisnis	PENGARUH CELEBRITY ENDORSEMENT TERHADAP BRAND CREDIBILITY DAN BRAND EQUITY PADA ONLINE SHOP (Study pada Online Shop Vanilla Hijab Indonesia) Tujuan dari penelitian ini adalah untuk mengetahui pengaruh penggunaan Celebrity Endorsement terhadap Brand Credibility dan Brand Equity pada online shop yang ada di sosial media Instagram. Jumlah sampel dalam penelitian adalah sebanyak 110 responden konsumen Vanilla Hijab yang ada di seluruh Indonesia.
Pendidikan Akuntansi & Bisnis	PENERAPAN PROBLEM BASED LEARNING DALAM KERANGKA LESSON STUDY UNTUK MENINGKATKAN KEMAMPUAN BERPIKIR KRITIS DAN HASIL BELAJAR AKUNTANSI SISWA Penelitian ini dilaksanakan dengan tujuan untuk mendeskripsikan proses pembelajaran dengan menggunakan metode PBL dalam kerangka lesson study dalam meningkatkan kemampuan berpikir kritis.
Pendidikan Bisnis & Manajemen	PENGEMBANGAN MEDIA MOVIE MAKER PADA MATA PELAJARAN KEARSIPAN KELAS X ADMINISTRASI PERKANTORAN Tujuan penelitian ini untuk menghasilkan produk media pembelajaran movie maker dalamï kompetensi dasar Mengidentifikasi Alat dan Bahan Kearsipan serta Menjelaskan Pengurusan Surat Masuk
Pendidikan Akuntansi	PENGARUH KECERDASAN INTERPERSONAL TERHADAP PEMAHAMAN AKUNTANSI DENGAN KEPERCAYAAN DIRI SEBAGAI VARIABEL INTERVENING PADA SISWA JURUSAN AKUNTANSI Penelitian ini menguji pengaruh kecerdasan interpersonal terhadap pemahaman akuntansi dengan kepercayaan diri sebagai variabel intervening pada siswa jurusan akuntansi SMK Muhammadiyah 1 Kota Pasuruan

B. Preprocessing

Preprocesing diperlukan untuk memaksimalkan kinerja algoritma klasifikasi [2]. Umumnya ada empat tahap preprocessing untuk dokumen teks, yaitu *case folding*, *tokenizing*, *stopwords removal* dan *stemming* [3]. Namun dalam penelitian ini ada dua jenis *stopwords removal* yang diterapkan yaitu *stopwords removal* berbasis kamus dan berbasis *document frequency.* Kemudian SMOTE diaplikasikan di akhir *preprocessing.*

Case folding dilakukan dengan meratakan seluruh teks pada atribut Judul dan Abstrak menjadi *lowercase* serta menghilangkan karakter *non-word* seperti simbol, tanda baca dan angka sehingga yang tersisa hanyalah teks alfabetis a sampai z. Tabel 3 menunjukkan perubahan data sebelum dan sesudah case folding.

Tabel 3. Perubahan data sebelum dan sesudah case folding

Kelas	Judul (Sebelum Case Folding)	Judul (Setelah Case Folding)
Ekonomi Bisnis	PENGARUH INTERAKSI PERSONAL, KEBIJAKAN, ASPEK FISIK, RELIABILITAS, DAN PEMECAHAN MASALAH TERHADAP LOYALITAS PELANGGAN RITEL : SEBUAH KONTEKS TOKO BUKU	pengaruh interaksi personal kebijakan aspek fisik reliabilitas dan pemecahan masalah terhadap loyalitas pelanggan ritel sebuah konteks toko buku

Tabel 4. Perubahan data sebelum dan sesudah tokenizing

Kelas	Sebelum *Tokenizing*	Setelah *Tokenizing*
Ekonomi Bisnis	pengaruh interaksi personal kebijakan aspek fisik reliabilitas dan pemecahan masalah terhadap loyalitas pelanggan ritel sebuah konteks toko buku	'pengaruh' 'interaksi' 'personal' 'kebijakan' 'aspek' 'fisik' 'reliabilitas' 'dan' 'pemecahan' 'masalah' 'terhadap' 'loyalitas' 'pelanggan' 'ritel' 'sebuah' 'konteks' 'toko' 'buku'

Langkah selanjutnya adalah Tokenizing, yaitu memenggal dokumen menjadi satuan kata. Tokenizing diperlukan untuk proses *stopwords removing* berbasis kamus yang berjalan dengan perulangan pada tiap-tiap kata dalam dokumen. Tabel 4 menunjukkan perubahan data sebelum dan sesudah tokenizing.

Kamus yang dijadikan acuan untuk *stopwords removal* dalam penelitian ini adalah *stoplist* yang diajukan Fadillah Z. Tala pada tahun 2013 [4]. *Stoplist* tersebut dirinci dari hasil analisis frekuensi kata dalam Bahasa Indonesia, isinya adalah kata-kata Bahasa Indonesia yang paling sering yang sering muncul namun tidak memiliki arti tertentu seperti kata ‘adalah’, ‘yaitu’, ‘sebaliknya’, ‘begitu’, ‘demikian’, dan semacamnya. Bila kata-kata dalam kamus ditemukan dalam dokumen, maka kata tersebut dihapus. Tabel 5 menunjukkan perubahan data sebelum dan sesudah *stopwords removal* berbasis kamus Tala.

Tabel 5. Perubahan data sebelum dan sesudah *stopwords removal* berbasis kamus Tala

Kelas	Judul (Sebelum *Stopwords removal*)	Judul (Setelah *Stopwords removal*)
Ekonomi Bisnis	'pengaruh' 'interaksi' 'personal' 'kebijakan' 'aspek' 'fisik' 'reliabilitas' 'dan' 'pemecahan' 'masalah' 'terhadap' 'loyalitas' 'pelanggan' 'ritel' 'sebuah' 'konteks' 'toko' 'buku'	pengaruh interaksi personal kebijakan aspek fisik reliabilitas pemecahan loyalitas pelanggan ritel konteks toko buku

Stemming bertujuan untuk menyaring kata dasar dari setiap kata yang yang ada dalam dokumen. Sehingga setiap kata yang berbeda tetapi serupa misal seperti ‘mengukur’ dan ‘pengukuruan’ dianggap satu kata yang sama yaitu ‘ukur’. Program stemming dokumen Bahasa Indonesia sesuai PUEBI paling populer dikembangkan dalam penelitian berjudul “Stemming Indonesian: A confix-stripping approach.” [5]. Stemming menggunakan algoritma Nazief menghasilkan precision yang lebih tinggi dibanding algoritma stemming lainnya [6]. Dari algoritma Nazief, dikembangkan sebuah library stemming Bahasa Indonesia yang bernama Sastrawi [7]. Tabel 6 menunjukkan perubahan data sebelum dan sesudah *Stemming*.

Tabel 6. Perubahan data sebelum dan sesudah *Stemming*

Kelas	Judul (Sebelum *Stemming*)	Judul (Setelah *Stemming*)
Ekonomi Bisnis	pengaruh interaksi personal kebijakan aspek fisik reliabilitas pemecahan loyalitas pelanggan ritel konteks toko buku	pengaruh interaksi personal bijak aspek fisik reliabilitas pecah loyalitas langgan ritel konteks toko buku

Perbedaan stopwords removal yang berikutnya terletak pada pemilihan kata-kata yang hendak dihapus. Stopwords removal berbasis document frequency berarti ada pembatasan minimal dan maksimal frekuensi dokumen yang menampikan sebuah kata agar kata itu dipertahankan.

Frekuensi dokumen maksimal adalah 60 , artinya, bila ada kata yang muncul di lebih dari 60 dokumen dalam data latih, maka kata tersebut dihilangkan. 60 dipilih sebab dalam penelitia sebelumnya dimana stopwords removal berbasis term frequency diterapkan, 60 memberikan hasil maksimal [1]. Frekuensi dokumen minimal yaitu 2. Ini berarti, semua kata yang muncul dalam kurang dari 2 dokumen maka kata tersebut diabaikan. Pembatasan document frequency minimal ini mencegah gangguan yang disebabkan oleh kata-kata salah ketik, atau kata terlalu asing yang kemungkinan besar tidak akan muncul lagi pada data tes. Document frequency merupakan salah satu parameter TF-IDF, sehingga urutan stopwords removal berbasis document frequency ini harus diletakan di belakang transformasi VSM menggunakan TF-IDF.

TF-IDF sendiri adalah gabungan dari Term Frequency dan Inverse Document Frequency. TF digunakan untuk mengukur bahwa berapa kali suatu kata atau frasa hadir dalam dokumen, sedangkan IDF memberikan bobot lebih rendah untuk kata-kata yang sering muncul dan memberikan bobot lebih besar untuk kata-kata yang jarang muncul [8][9]. Metode ini paling umum digunakan pada text mining sebab efisien, mudah dan memiliki hasil yang akurat.[10].

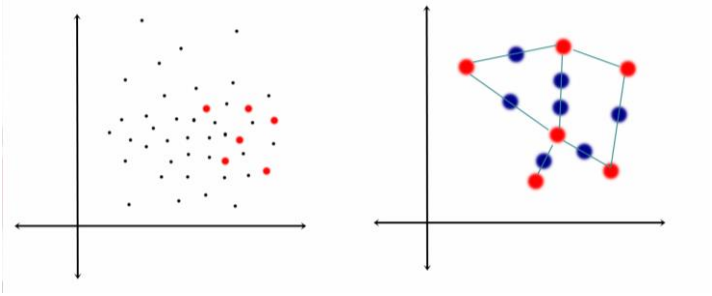
Tabel 7 menampilkan perhiungan term frequency dan document frequency untuk query “pengaruh interaksi personal bijak aspek fisik reliabilitas pecah loyalitas langgan ritel konteks toko buku” pada data latih. Kata ‘pengaruh’ dihapus sebab melebihi document frequency maksimal. Begitu pula kata ‘personal’, ‘fisik’, ‘pecah’, ‘loyalitas’, dan ‘toko’ dihapus sebab tidak memenuhi document frequency minimal.

Tahap preprocessing yang paling belakang adalah SMOTE (Synthetic Minority Over-sampling Technique), untuk mengatasi ketidakseimbangan data yang terjadi jika jumlah objek suatu kelas data lebih banyak dibandingkan dengan kelas lain. Kelas data yang objeknya lebih banyak disebut kelas mayor sedangkan lainnya disebut kelas minor[11].

Tabel 7. *Term* dan *document frequency* untuk *query* "pengaruh interaksi personal bijak aspek fisik reliabilitas pecah loyalitas langgan ritel konteks toko buku"

Q	tf				df
	d1	*d2*	*d3*	*d4*	
pengaruh	91	40	79	73	65
interaksi	0	5	5	2	3
personal	0	0	1	4	2
bijak	24	0	0	0	5
aspek	1	2	2	3	6
fisik	0	0	0	0	0
reliabilitas	0	0	1	3	4
pecah	0	0	3	0	1
loyalitas	6	0	0	0	1
langgan	10	0	2	0	2
ritel	0	0	2	0	2
konteks	0	2	10	0	4
toko	0	0	0	0	0
buku	0	9	9	1	4

Cara kerja SMOTE yaitu menciptakan data sintetis di tengah dua data terdekat. Dalam menentukan dua data terdekat ini, pengukurannya menggunakan euclidean distance. [12]. Pada Gambar 1, titik-titik hitam mewakili kelas mayor, titik-titik merah mewakili kelas minor. Titik-titik biru yang terbentuk di sekitar titik merah adalah data sintetis dari hasil SMOTE.



Gambar. 1. Ilustrasi SMOTE

Tabel 8. Perubahan data sebelum dan sesudah SMOTE

Label	Sebelum SMOTE		Seteah SMOTE	
	Jumlah data	*Persentase*	*Jumlah data*	*Persentase*
Ekonomi Bisnis	29	23.01%	36	25 %
Pendidikan Akuntansi dan Bisnis	31	24.60%	36	25%
Pendidikan Bisnis dan Manajemen	30	23.81%	36	25%
Pendidikan Akuntansi	36	28.57%	36	25%
Total	126	100%	288	100%

C. Skenario Preprocessing

Penelitian akan dilakukan dalam sembilan skenario yang berbeda. Masing-masing skenario memiliki perbedaan pada bagian preprocessing. Setiap skenario diuji dengan komposisi dan urutan teknik preprocessing yang berbeda-beda pula. Sembilan skenario tersebut ditampilkan pada Tabel 9.

Seperti semua skenario yang akan diuji, preprocessing pada Skenario 1 diawali case folding dan tokenizing, yaitu pembersihan dokumen dari karakter non-alfabetis dan memenggalnya menjadi satuan kata. Kemudian, dokumen ditransformasi menjadi Vector Space Model (VSM) menggunakan pembobotan TF-IDF. ssSetelah diketahui bobot document frequency yang dijadikan acuan dalam memotong kata-kata tidak penting pada dokumen (stopwords removal). Stopwords removal berbasis frekuensi dokumen ini menghapus kata-kata yang tidak memenuhi documenf frequency minimal atau maksimal. Terakhir SMOTE meratakan kelas yang tidak seimbang dengan membuat data sintetis bagi kelas-kelas yang memiliki sedikit instance.

Teknik stemming Bahasa Indonesia ditambahkan pada Skenario 2 setelah proses tokenizing. Urutan preprocessing dengan stemming dan stopwords removal berbasis document frequency hanya dapat seperti ini sebab setelah vektorisasi TF-IDF, dokumen tidak dapat melakukan proses stemming. Begitu pula stopwords berbasis document frequency yang tidak dapat diletakkan sebelum TF-IDF.

Tabel 9. Skenario Peprocessing

Skenario	*Preprocessing*	Klasifikasi
1	*Case folding, tokenizing,* TF-IDF, *stopwords removal (document frequency),* SMOTE	K-Nearest Neighbour dengan metrik Cosine Similarity, 10-fold Cross vaidation.
2	*Case folding, tokenizing, stemming,* TF-IDF, *stopwords removal (document frequency),* SMOTE	
3	*Case folding, tokenizing, stopwords removal (Tala),* TF-IDF, SMOTE	
4	*Case folding, tokenizing, stemming, stopwords removal (Tala),* TF-IDF, SMOTE	
5	*Case folding, tokenizing, stopwords removal (Tala), stemming,* TF-IDF, SMOTE	
6	*Case folding, tokenizing, stemming,* TF-IDF, SMOTE	
7	*Case folding, tokenizing, stopwords removal (Tala),* TF-IDF, *stopwords removal (document frequency),* SMOTE	
8	*Case folding, tokenizing, stemming, stopwords removal (Tala),* TF-IDF, *stopwords removal (document frequency),* SMOTE	
9	*Case folding, tokenizing, stopwords removal (Tala), stemming,* TF-IDF, *stopwords removal (document frequency),* SMOTE	

Skenario 3 dan 4 memiliki komposisi yang sama dengan Skenario dan 2, hanya saja stopwords removal yang digunakan berbasis kamus stopwords Bahasa Indonesia milik [4]. Perbedaan lainnya adalah urutan stopwords removal berada sebelum transformasi TF-IDF, sebab stopwords removal yang ini tidak memerlukan nilai document frequency dan tidak dapat dilakukan pada dokumen yang sudah berbentuk VSM.

Lain dengan stopwords removal berbasis document frequency, penggunaan stopwords removal dengan kamus dapat dilakukan sebelum stemming. Skenario 5 membalik posisi stemming dan stopwords removal yang ada pada Skenario 4. Bila Skenario 1 dan 3 memberikan komposisi preporcessing dengan stopwords removal tanpa stemming, maka pada Skenario 6 diberikan komposisi preprocessing dengan stemming tanpa satu pun teknik stopwords removal. Skenario ini menguji sejauh apa performa stemming tanpa kolaborasi stopwords removal bersamanya.

Dua teknik stopwords removal yang memiliki cara kerja berbeda perlu diujikan performanya pada preprocessing yang menyertakan keduanya. Skenario 7 sampai 9 pada dasarnya adalah Skenario 3 sampai 5 yang ditambahi stopwords removal berbasis document frequency setelah transformasi TF-IDF dan sebelum SMOTE. Sisanya, urutan dan komposisi lain seperti stemming sama persis dengan Skenari 3 sampai 5.

D. Klasifikasi

K-Nearest Neighbor (KNN) merupakan metode yang digunakan untuk melakukan klasifikasi berdasarkan data pembelajaran yang paling dekat dengan objek [13]. Algoritma KNN termasuk ke dalam Instance Based Learning. KNN mencari pola k yang terdekat dengan pola masukan, kemudian menentukan kelas keputusan berdasarkan jumlah pola terbanyak di antara pola k [14].

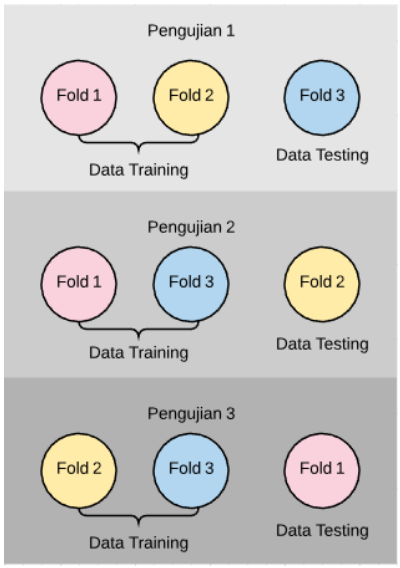
Pada penelitian ini, parameter k yang digunakan adalah k=1, itu artinya kelas data ditentukan berdasarkan kelas yang dimiliki oleh 1 tetangga terdekatnya. k=1 dipilih sebab untuk data yang telah melalui proses SMOTE, k=1 lah yang terbaik, mengingat pembuatan data sintetis pada SMOTE terletak di antara dua data yang kedekatannya dihitung menggunakan algoritma 1-nearest neigbour (Chawla, 2002).

Dalam mengukur jarak tetangga, metrik yang digunakan yaitu cosine similarity Metrik ini memodelkan dokumen teks sebagai vektor kata. Dengan model ini, kesamaan antara dua dokumen dapat diturunkan oleh menghitung nilai kosinus antara vektor kata dua dokumen[15]. Metrik Cosine Similarity ditunjukkan pada persamaan (1) [16].

$$\cos\alpha = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_1}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

Pada persamaan (1), A adalah Vektor A, yang akan dibandingkan kemiripannya, B adalah Vektor B, yang akan dibandingkan kemiripannya, $A \cdot B$ adalah dot product antara vektor A dan vektor B, $|A|$ adalah panjang vektor A, $|B|$ adalah panjang vektor B, dan $|A||B|$ adalah cross product antara |A| dan |B|.

Cross validation adalah metode statistik yang mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua yaitu data training data testing. Bentuk dari cross validation adalah k-fold cross validation.[17] Jumlah data yang menjadi data training pada setiap pengujian adalah sebanyak k-1 folds, sedangkan satu fold yang lain menjadi data testing. Gambar 2 menunjukkan ilustrasi bagaimana K-Folds Cross Validation melakukan pengujian.



Gambar. 2. Ilustrasi pengujian data dengan 3-folds Cross Validation.

Jumlah folds yang umum digunakan adalah 10 [17], begitu pula yang diterapkan dalam penelitia ini. Untuk 10-fold cross validation, data di bagi menjadi sepuluh kelompok. Pada masing-masing kelompok terjadi pembagian data latih dan data uji dengan perbandingan 9:1. Pembagian ini bertujuan agar seluruh bagian dataset pernah digunakan sebagai data latih maupun data uji.

Confusion matrix diterapkan sebagai pengukur keakuratan hasil klasifikasi.[18] Hasil dari tabel confusion matrix digunakan untuk menghitung accuracy, precision, dan recall. Accuracy merupakan tingkat kedekatan antara nilai prediksi dan nilai aktual. Precision didefinisikan sebagai tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. True positive rate (recall) adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

III. Hasil dan Pembahasan

Bagian Penelitian yang dilakukan sesuai dengan prosedur yang telah dijelaskan pada bab sebelumnya memberikan hasil yang beragam pada tiap-tiap skenarionya. Tabel 10 menunjukkan hasil penelitian yang telah dilakukan.

Nilai accuracy terendah dimiliki oleh Skenario 6 yang menggunakan komposisi preprocessing; case folding, tokenizing, stemming, vektorisasi denan TF-IDF dan SMOTE tanpa satu pun proses stopwords removal. Nilai ini bahkan lebih rendah dari Skenario 1 dan 3 yang hanya menggunakan salah satu teknik stopwords removal tanpa stemming.

Tabel 10. Hasil penelitian

	Hasil			
	Accuracy	***Precision***	***Recall***	***Jumlah Kata***
Skenario 1	69,44%	70,03%	69,44%	829
Skenario 2	71,53%	72,33%	71,53%	712
Skenario 3	68,75%	69,99%	68,75%	1840
Skenario 4	69,44%	70,20%	69,44%	1435
Skenario 5	70,83%	71,68%	70,83%	1490
Skenario 6	68,05%	69,98%	68,05%	1624
Skenario 7	71,53%	72,37%	71,53%	661
Skenario 8	72,92%	73,09%	72,92%	573
Skenario 9	72,92%	73,36%	72,92%	605

Nilai accuracy tertinggi dicapai oleh dua skenario terakhir, yang komposisi preprocessing-nya terdiri dari case folding, tokenizing, stopwords removal berbasis kamus, stemming, vektorisasi dengan TF-IDF, stopwords removal berbasis document frequency, dan SMOTE, dengan urutan stemming yang dibolak-balik

Fluktuasi yang sedikit berbeda terjadi pada perbandingan precision. Bila pada accuracy, nilai tertingginya dicapai oleh dua skenario, 8 dan 9, pada precision tampah bahwa Skenario 9 memiliki nilai yang lebih tinggi dari semua Skenario yang telah diujikan. Nilai precision terendah masih dimiliki oleh Skenario 6 dengan selisih 0.004% dibanding Skenario 3.

Proses SMOTE untuk mengatasi ketidakseimbangan data menjadikan nilai recall dan accuracy menjadi sangat dekat. Tidak ada perbedaan naik dan turun grafik Skenario 1 hingga 9, termasuk hasil terendah yang ada pada Skenario 6 dan hasil tertingi pada Skenario 8 dan 9.

Teknik preprocessing yang paling banyak memotong dokumen adalah stopwords removal berbasis document frequency, sebagaimana yang diterapkan pada Skenario 1, 2, 7, 8, dan 9. Kata unik yang tersisa setelah pengaplikasian stopwords removal berbasis document frequency ada di bawah 1.000. Pada Skenario 3~6, jumlah kata unik yang tersisa masih 1.400 ke atas, sebab stopwords removal berbasis kamus dan stemming tidak menyusutkan terlalu banyak jumlah kata unik.

Pada Skenario 1, data yang sudah melewati proses case folding dan tokenizing, data kemudian ditransformasi menjadi VSM menggunakan TF-IDF. Setelahnya, dokumen memasuki proses stopwords removal berbasis document frequency. Skenario 2 melewati proses yang sama. Bedanya, sebelum ditransformasi menjadi VSM, dilakukan stemming terlebih dahulu. Data dipangkas setiap katanya dari imbuhan dan hanya diambil kata dasarnya saja.

Stemming berpengaruh pada stopwords removal berbasis document frequency, sebab banyak kata yang sebelumnya dianggap berbeda seperti “pasar” dan “pemasaran” setelah stemming menjadi sama-sama “pasar”. Misal sebelumnya kata “pasar” yang hanya muncul satu kali dihapus sebab tidak memenuhi syarat minimal document frequency, di sini kata “pasar” tidak dihapus karena muncul lebih dari satu kali. Begitu pula misal sebelumnya “pasar” tidak dihapus karena muncul 60 kali, di sini kata tersebut dihapus karena ketambahan “pasar” lain hasil pemangkasan imbuhan.

Apabila dibandingkan, peningkatan yang terjadi pada Skenario 2 lebih tinggi dari Skenario 4 atau 5, ini artinya stopwords removal berbasis document frequency dapat memberikan hasil yang lebih maksimal. Ini karena stopwords removal berbasis kaus tidak dapat menyingkirkan kata-kata yang frekuensinya meningkat setelah stemming.

Secara keseluruhan, proses stemming Bahasa Indonesia pada dokumen memberikan kolaborasi yang baik dengan penghapusan stopwords. Terbukti pada Skenario 2, 4, serta 8 terjadi peningkatan di masing-masing accuracy, precision, dan recall dibanding skenario sebelumnya yang tidak menggunakan stemming. Jumlah kata unik menurun sebab semua kata yang terbentuk dari kata dasar yang sama telah diserupakan.

Dari skenario yang disebut sebelumnya tampak bahwa stemming meningkatkan performa klasifikasi, akan tetapi, pada Skenario 6 yang hanya menerapkan stemming tanpa satu pun proses stopwords removal yang, ketepatan prediksi data tes rendah, bahkan lebih rendah dari skenario hanya dengan stopwords removal tanpa stemming. Hal ini dikarenakan stemming menyamakan semua kata dengan kata dasar yang serupa. Sedangkan dalam proses klasifikasi dokumen teks, semakin banyak kesamaan kata dalam tiap-tiap kelas, semakin sulit data tes diprediksi. Ternyata, stemming Bahasa Indonesia menunjukkan performa yang buruk apabila tidak dikombinasikan dengan stopwords removal sama sekali.

Penempatan stemming dalam urutan preprocessing pun memberi pengaruh terhadap hasil klasifikasi. Skenario 5 dan 9 menunjukkan bahwa stemming lebih baik diletakkan setelah stopwords removal berbasis kamus. Sebab, daftar stopwords di dalam kamus Tala mengandung banyak sekali kata-kata berimbuhan. Apabila kata-kata dalam dokumen melalui proses stemming dahulu, maka kata yang seharusnya cocok dengan kata dalam kamus jadi lolos dari eliminasi.

Tanpa stemming, Skenario 1 dan 3 yang menerapkan teknik stopwords removal berbeda menampilkan hasil yang berbeda pula. Stopwords removal berbasis document frequency memberikan hasil yang lebih baik dibanding stopwords removal yang berpatok pada stoplist ajuan Tala.

Dua teknik stopwords removal yang ada dapat dikombinasikan karena memiliki cara kerja yang berbeda. Stopwords removal berbasis kamus menghapus kata-kata yang tidak penting dalam keseluruhan dokumen, kemudian, stopwords removal berbasis document frequency menghapus kata-kata yang lolos namun terlalu sering muncul atau terlalu asing dalam data latih. Kerja sama yang baik ini menjadikan nilai accuracy, precision, dan recall Skenario 7 lebih tinggi dari skenario-skenario dengan teknik stopwords removal berbasis kamus dan berbasis frekuensi tanpa satu sama lain.

Dua teknik stopwords removal yang digunakan berbagi tugas dalam membuang kata-kata yang mengganggu, dan di skenario-skenario sebelumnya menunjukkan bahwa stemming memaksimalkan hasil

kerja stopwords removal berbasis frekuensi. Oleh karena itu, memadukan ketiga teknik ini dapat memberikan nilai akurasi yang tinggi, seperti pada Skenario 8 dan 9.

Hasilnya, sebagaimana Skenario 5, urutan stemming lebih baik diletakkan di belakang stopwords removal berbasis kamus, menjadikan Skenario 9 lebih baik dari Skenario 8 dan semua skenario yang telah diujikan.

Kemudian, apabila menengok jumlah kata yang tersisa di masing-masing Skenario, tampak banyak atau tidaknya kata yang dihapus tidak mempengaruhi ketepatan klasifikasi, misalnya Skenario 1 dan 5 bila dibadingkan, Skenario 5 menghasilkan hasil klasifikasi yang lebih baik walau kata yang dihapus tidak sebanyak Skenario 1. Adi Namun, pemotongan kata dapat memberi pengaruh besar terhadap running time program, terutama bila kasusnya data latih merupakan data yang amat besar.

IV. Kesimpulan

Dari penelitian yang telah dilakukan, ditarik kesimpulan bahwa klasifikasi artikel berdasarkan judul dan abstrak menggunakan algoritma k-Nearest Neighbor dengan metrik Cosine Similarity dapat dimaksimalkan dengan 9 jenis Skenario yang terdiri dari teknik-teknik dan urutan yang berbeda. Kombinasi teknik-teknik preprocessing dan urutannya mempengaruhi performa algoritma klasifikasi, ada yang menjadikan lebih baik dan ada yang memperburuk.

Skenario yang dapat memberikan hasil klasifikasi terbaik adalah Skenario 9, dengan nilai accuracy, precision, dan recall masing-masing 72.92%, 73.36%, dan 72.92%. Komposisi prerocessing Skenario 9 terdiri dari; case folding, tokenizing, stopwords removal berbasis kamus, stemming, transformasi ke bentuk VSM, stopwords removal berbasis frekuensi dokumen, dan yang terakhir, sebagaimana skenario lainnya; SMOTE. Komposisi ini harus berurutan atau hasilnya akan berubah. Sedangkan Skenario yang menghasilkan hasil klasifikasi paling rendah dalam penelitian ini adalah Skenario 6, dengan nilai accuracy, precision, dan recall masing-masing 68.05%, 69.98%, dan 68.05%. Komposisi preprocessing pada Skenario 6 terdiri dari; case folding, tokenizing, stemming, transformasi ke bentuk VSM, dan SMOTE.

Daftar Pustaka

[1] P. D. Nurfadila, “Klasifikasi Jurnal Menggunakan Metode Cosine Similarity dengan Pengurangan Konten pada Judul dan Abstrak Berbasis Frequency.”

[2] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, “Data Processing and Text Mining Technologies on Electronic Medical Records: A Review,” J. Healthc. Eng., vol. 2018, pp. 1–9, 2018.

[3] S. F. Crone, S. Lessmann, and R. Stahlbock, “The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing,” Eur. J. Oper. Res., vol. 173, no. 3, pp. 781–800, 2006.

[4] F. Z. Tala, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia,” M.Sc. Thesis, Append. D, vol. pp, pp. 39–46, 2003.

[5] M. Adriani, B. Nazief, J. Asian, and H. E. Williams, “Stemming Indonesian: A confix-stripping approach,” ACM Trans. Asian Lang. Inf. Process., vol. 6, no. 4, 2007.

[6] L. Agusta, “Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Babahasa Indonesia,” pp. 196–201, 2009.

[7] A. Librian, “High quality stemmer library for Indonesian Language (Bahasa),” 2017. .

[8] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” Int. J. Comput. Appl., vol. 181, no. 1, pp. 25–29, 2018.

[9] V. Amrizal, “Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim),” J. Tek. Inform., vol. 11, no. 2, pp. 149–164, 2019.

[10] A. A. Maarif, “Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah,” Univ. Dian Nuswantoro Semarang, no. 5, p. 4, 2015.

[11] R. A. Barro, I. D. Sulvianti, and F. M. Afendi, “Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu,” Xplore J. Stat., vol. 1, no. 1, 2013.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. January, pp. 321–357, 2002.

[13] F. Agus, H. R. Hatta, and Mahyudin, “Pengklasifikasian Dokumen Berbahasa Arab Menggunakan K-Nearest Neighbor,” JSM STMIK Mikroskil, vol. 18, no. 1, pp. 43–56, 2017.

[14] Suyanto, Machine Learning Tingkat Dasar dan Lanjut, 1st ed. Bandung: Informatika Bandung, 2018.

[15] F. Rahutomo, T. Kitasuka, and M. Aritsugi, “Semantic Cosine Similarity,” Semant. Sch., vol. 2, no. 4, pp. 4–5, 2012.

[16] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Jurnal Teknik Elektro," J. Tek. Elektro, vol. 9, no. 1, pp. 18–23, 2017.

[17] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in Encyclopedia of database systems, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer, 2009.

[18] P.-N. Tan, M. Steinbach, and K. Vipin, Introduction to data mining. Pearson Education India, 2006.