Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)

By Samsuryadi Samsuryadi

WORD COUNT 4948 TIME SUBMITTED 19-JUN-2024 04:41PM

PAPER ID 109726762

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)

Willy*, Dian Palupi Rini, Samsuryadi

Fakultas Ilmu Komputer, Magister Ilmu Komputer, Universitas Sriwijaya, Palembang, Indonesia
Email: [1,*]willywilly279@gmail.com, [2]dian.palupi.rini@gmail.com,[3]samsuryadi@gmail.com
Email Penulis Korespondensi: willywilly279@gmail.com

Abstrak−*Fake news* merupakan informasi palsu yang menyerupai seakan-akan itu adalah benar. Berita dapat juga dikatakan sebagai senjata politik yang kebenarannya tidak bisa dipertanggungjawabkan yang disebarkan secara sengaja untuk mencapai suatu tujuan tertentu. Klasifikasi teks berita membutuhkan kalkulasi suatu metode terhadap setiap kata pada dokumen. Setiap kata yang diproses per dokumen mengartikan bahwa jumlah dimensi data sama dengan jumlah kata. Semakin banyak jumlah kata pada suatu dokumen, semakin banyak juga jumlah dimensi pada setiap data (*high dimension*). Jumlah dimensi yang besar (*high dimension*), menyebabkan proses pembuatan model (*training*) yang lama dan juga terlihat jelas kekurangannya dalam melihat kemiripan dokumen (*document similarity*). Dataset yang digunakan pada penelitian ini berjumlah 20000 dan 17 atribut. Metode yang digunakan dalam penelitian dengan menggunakan *Random Forest Classifier(RFC), Support Vector Machine(SVM)* dan *Logistic Regression(LR)* dengan *high dimension* dan hasil penelitian ini untuk mendapatkan perbandingan nilai akurasi pada masing-masing metode yang digunakan.

Kata Kunci: Fake News; High Dimension; Dataset; RFC; SVM; LR

Abstract−Fake news is false information that looks like it is true. News can also be said as a political weapon whose truth cannot be accounted which is spread intentionally to achieve a certain goal. Classification of texts requires calculating a method for each word in the document. Each word processed per document means that the number of data dimensions is equal to the number of words. The more the number of words in a document, the more the number of dimensions in each data (high dimension). The large number of dimensions (high dimension), causes the model-making process (training) to be long and the shortcomings are also clearly visible in seeing the similarity of documents (document similarity). The dataset used in this study amounted to 20000 and 17 attributes. The method used in this study uses Random Forest Classifier (RFC), Support Vector Machine (SVM) and Logistic Regression (LR) with high dimensions and the results of this study are to obtain a comparison of the accuracy values for each method used.

Keywords: Fake News; High Dimension; Dataset; RFC; SVM; LR

1. PENDAHULUAN

Fake news atau berita palsu merupakan suatu informasi palsu serta berbahaya karena dapat mempengarui persepsi tentang informasi tersebut sebagai suatu kebenaran serta dapat mempengarui banyak orang dapat merusak suatu citra dan kredibilitas [1]. *Fake news* merupakan informasi palsu yang menyerupai seakan-akan itu adalah benar [2]. Berita dapat juga dikatakan sebagai senjata politik yang kebenarannya tidak bisa dipertanggungjawabkan yang disebarkan secara sengaja untuk mencapai suatu tujuan tertentu. *Fake news* juga dapat berdampak buruk bagi seseorang ataupun sekelompok orang atau golongan, *fake news* juga dapat membuat masyarakat menjadi panik sehingga dapat berpengaruh buruk di dalam kehidupan sehari-hari [3]. *Fake news* bertujuan untuk mempengaruhi para pencari informasi, berita di media sosial ataupun berita di *website*, sehingga pencari berita dapat melakukan tindakan yang tidak baik, berita palsu juga dapat membuat ketakutan dimasyarakat, maka berita palsu harus dijelaskan, diidentifikasi serta diklarifikasi [4]. Melakukan validasi fakta pada setiap berita yang ada di media digital ataupun media fisik membutuhkan sumber daya (*resource*) yang banyak, oleh karena itu dibutuhkannya teknik perbandingan klasifikasi yang bertujuan membantu otomatisasi validasi fakta berita tersebut [5].

Klasifikasi teks berita membutuhkan kalkulasi suatu metode terhadap setiap kata pada dokumen. Setiap kata yang diproses per dokumen mengartikan bahwa jumlah dimensi data sama dengan jumlah kata. Semakin banyak jumlah kata pada suatu dokumen, semakin banyak juga jumlah dimensi pada setiap data (*high dimension*). Jumlah dimensi yang besar (*high dimension*), menyebabkan proses pembuatan model (*training*) yang lama dan juga terlihat jelas kekurangannya dalam melihat kemiripan dokumen (*document similarity*) [6]. Pemodelan data dimensi tinggi menjadi pusat perhatian pada dua dekade terakhir. Munculnya data dimensi tinggi telah menghadirkan tantangan bagi data mining. Data dimensi tinggi dicirikan dengan peubah prediktor lebih banyak dari pengamatan. Data dimensi tinggi sering kali dijumpai pada data-data digital seperti *microarray*, teks maupun data hasil pengolahan sinyal. Data semacam ini memiliki prediktor yang berjumlah ribuan, namun hanya memiliki objek pengamatan yang jauh lebih sedikit.

Penelitian *high dimension* untuk *text classification* sudah banyak dilakukan sebelumnya, misalnya klasifikasi teks berita BBC dengan LR, RF dan KNN [7], klasifikasi lagu berdasarkan lirik lagu dengan SVM dan NB [8]. Hasil setiap penelitian tersebut memiliki akurasi diatas 75% dengan beberapa nilai *metric* lainnya yang juga diatas 75% (*precision*, *recall*, *F1-score*). Hasil penelitian tersebut menandakan bahwa metode yang

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

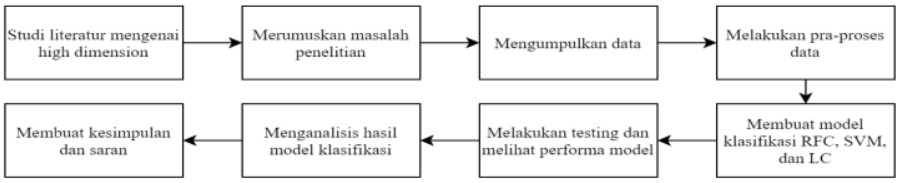
dipakai (LR, RF, KNN, SVM dan NB) cukup baik dalam melakukan klasifikasi *high dimension* data pada kasus seperti berita maupun lirik lagu. Penelitian terkait berita palsu pernah dilakukan olehF. A. Ozbay & B. Alatas pada tahun 2019 lalu dengan menggunakan metode *Decision Stump*,*Logistic Model Tree*dan J48 dengan mengklasifikasikan antara berita palsu dan berita asli, penelitian tersebut mendapatkan persentase hasil berturut-turut sebesar 56,4% untuk *Decision Stump*, 60,7% untuk *Logistic Model Tree* dan 55,8% untuk J48 [9]. Penelitian selanjutnya dilakukan oleh Eka Listiana dan Much Aziz Muslim dengan menggunakan metode SVM dengan hasil 62.5 % [10]. Penelitian selanjutnya pernah dilakukan Vinodhini dan Chandrasekaran menggunakan PCA sebagai reduksi dimensi serta teknik hibrida untuk klasifikasi sentimen dengan menghasilkan akurasi tertinggi 83,3% [11].

Pada penelitian ini akan dibangun sebuah perbandingan algoritma *Random Forest Classifier, Support Vector Machine dan Logistic Regression* pada masalah *high dimension* untuk melihat algoritma yang paling sesuai pada masalah klasifikasi berita fake news. *Dataset fake news* diambil pada situs *https://www.kaggle.com/* sebanyak 20000 data dan *Hyperparameter* yang digunakan untuk *Random Forest, Support Vector Machine, Logistic regression*. Tujuan yang ingin dicapai dari penelitian ini adalah mengetahui solusi dari permasalahan klasifikasi pada *high dimension* dan mengetahui hasil perbandingan metode *random forest* (RF), *support vector machine* (SVM), dan *logistic regression* (LR) pada masalah *high dimension*. Diharapkan penelitian ini bertujuan untuk mendapatkan hasil nilai akurasi masing-masing metode serta perbandingannya, mendapatkan performa dari masing masing algoritma, serta menjadi referensi bagi penelitian selanjutnya yang bertujuan meningkatkan performa high dimension.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian merupakan detil kegiatan yang dilakukan selama penelitian berlangsung. Tahapan penelitian pada penelitian ini dapat dilihat pada gambar 1.



Gambar 1. Tahapan Metodologi Penelitian

a. Tahap pertama diawali menentukan masalah yang akan diteliti pada penelitian ini, yaitu masalah klasifikasi pada data *high dimension*, dilanjutkan studi literatur beberapa model klasifikasi yang dapat melakukan klasifikasi pada data *high dimension*.
b. Pada tahap rumusan masalah, didapatkan 3 model klasifikasi yang cocok untuk klasifikasi data *high dimension* yaitu *random forest classifier* (RFC), *support vector machine* (SVM), dan *logistic regression* (LR). Tiga model klasifikasi ini nantinya akan dilakukan perbandingan *metric sensitivity*, *specificity*, *precision*, *f1-score*, *accuracy*, dan *error*.
c. Pada tahap pengumpulan data, dilakukan pengambilan data dari situs web *kaggle* yang terdiri dari 33.342 data gabungan antara *fake news* dan *real news* beserta dengan labelnya masing-masing.
d. Pada tahap pra-proses data, akan dilakukan proses *case folding*, *remove punctuation*, *remove number*, *stop words removal*, dan *stemming*. Kemudian akan dilihat apakah data tersebut *balance* atau *imbalance*. Jika data *imbalance* maka akan dilakukan penghapusan data agar data label tiap *class* menjadi seimbang. Data yang telah seimbang kemudian akan dilakukan proses *feature extraction* sebanyak 300 fitur dan semua data disimpan dalam format *pickle*.
e. Setelah data selesai di pra-proses, dilanjutkan dengan membuat model klasifikasi RFC, SVM, dan LR. Data dalam format *pickle* akan dibaca kemudian dilakukan *splitting data* antara *data testing* dan *data training*. Kemudian setiap data akan dilakukan proses *training* dengan *split data*.
f. Setelah fase *training model* selesai, dilanjutkan dengan fase *testing*. Setiap model memiliki *best parameter*-nya masing-masing. Hasil tiap model akan dibuat *confusion matrix*-nya kemudian diukur menggunakan *metrics sensitivity*, *specificity*, *precision*, *f1-score*, *accuracy* dan *error*.

2.2 Pre-processing

Tahap *case folding* pada tahap *case folding* akan dijalankan fungsi yang mengubah semua karakter/huruf pada kalimat di data *fake news* menjadi karakter/huruf kecil. Kemudian *Remove Punctuation* dimana input kalimat dari data *fake news* mempunyai kemungkinan untuk mengandung karakter non-alphabet seperti tanda seru (!),

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

tanda titik koma (;), tanda dollar ($), tanda persentase (%) dan tanda *punctuation* lainnya. Oleh karena ini pada tahap ini akan dihapus karakter *punctuation* tersebut dari data *fake news*. Tahapan ke tiga yaitu *Remove Number* dilakukan untuk menghapus angka yang terdapat pada *dataset fake news*. Angka akan dihapus pada awal kalimat, pertengahan kalimat, dan akhir kalimat. Ke empat adalah *Stop Words Removal* merupakan kumpulan kata yang sering digunakan (*commonly used words*) tetapi kurang memiliki makna pada kalimat tersebut. Tahap ini digunakan untuk menghapus kalimat-kalimat yang kurang bermakna pada *dataset fake news*, seperti *the*, *is*, *i*, *me*, *myself*, *it*, *them*, *does*, dan kata *stop words* lainnya. Tahap ke lima *Stemming* merupakan proses mengubah kata menjadi kata dasarnya tanpa imbuhan seperti awalan (*prefix*), akhiran (*suffix*), sisipan (*infix*), dan gabungan awal-akhir (*confix*). Tahap ini dilakukan untuk mengubah kata pada setiap data *fake news* menjadi kata dasarnya.

2.3 Metode

Random Forest Classifier (RFC) merupakan salah satu jenis metode dari *ensemble learning algorithms*. Pada umumnya, pembelajaran (*learning*) seperti *perceptron*, ADALINE, *support vector machine*, mencari 1 hipotesa sebagai solusi, **hipotesa** merupakan **hasil klasifikasi** dari data yang dimasukkan (*input data*) [11]. Sedangkan pada metode *ensemble learning*, pembelajaran dilakukan dengan membuat kumpulan hipotesa kemudian dilakukan *voting*. Hipotesa (umumnya berupa *class*) dengan *vote* tertinggi dijadikan solusi (*output class*) [11]. Konsep RF berawal dari Tin Kam Ho yang menggunakan *random subspace method* sebagai pemilihan acak fitur untuk pembuatan *tree* [12], dilanjutkan dengan pengembangan dari Leo Breiman yang menggunakan *bagging* (*bootstrap aggregating*) sebagai teknik pembentukan *data training* dengan *resampling* [13]. Pada tahun 2001, Leo Breiman [14] menggabungkan kedua teknik ini yang sekarang disebut sebagai *random forest classifier*. Metode RF terdiri dari 2 tahapan utama, yaitu pembentukan *forest* dan *voting* hasil klasifikasi/hipotesa [15]

$$forest = \{h(\mathrm{x}, \Theta_k), k = 1, \ldots\} \quad (1)$$

Keterangan:

h : *Hypothesis* atau *classifier* (*decision tree*)
x : Input *vector*
Θ_k : *Independent and identically distributed* (IID) *random vectors*

Persamaan 1 menyatakan bahwa *forest* terdiri dari kumpulan *decision tree* sebagai *classifier* yang berjumlah *k*. Masukkan (*input*) tiap *decision tree* berupa data *x* yang di-*resampling* dengan *random vectors* dari data *x* itu sendiri (Θ_k).

$$\hat{\mathrm{C}}_{rf} = majority\ vote\ \{\ \hat{\mathrm{C}}_n(x)\ \}_{n=1}^{N} \quad (2)$$

Keterangan:

$\hat{\mathrm{C}}_{rf}$: *Class* hasil *random forest*. Operator *hat* pada $\hat{\mathrm{C}}$ menadakan bahwa *class* tersebut merupakan *class* hasil estimasi
x : Input *vector*
$\hat{\mathrm{C}}_n$: *Class* prediksi dari *tree* ke-n pada *random forest*

Setelah selesai pembuatan *random forest*, kemudian dilanjutkan dengan *voting* untuk mengukur performa *random forest* seperti pada persamaan 2.

Support Vector Machine (SVM) awalnya dikembangkan oleh Boser dan Vapnik berdasarkan teori Vapnik-Chervonenkis (VC) dan *structural risk minimization* (SRM), yaitu minimalisasi *error* ketika *training* dan memaksimalkan *margin*, model ini sering dikenal sebagai *Maximal Margin Classifier* [16][17][18].

Pada kasus nyata, model *maximal margin classifier* selalu sensitif terhadap *noise* dan *outliers* (*low bias* & *high variance*), yang menyebabkan susahnya klasifikasi kelas secara akurat. Oleh karena itu, dilakukan pengembangan lebih lanjut oleh Cortes & Vapnik yang memperbolehkan kesalahan dalam melakukan klasifikasi data *noise* atau data *outliers* (*misclassification*). Model ini memakai istilah *soft margin*, yaitu jarak antara data yang dilihat (observasi) dengan *threshold*. Model yang memakai *soft margin* disebut sebagai *Support Vector Classifier* (SVC) [18][19].

Model *maximal margin classifier* (MMC) ataupun *support vector classifier* (SVC) umumnya dipakai untuk memisahkan data secara linear (*linearly separable*). Tetapi nyatanya, banyak kasus yang tidak dapat dipisahkan secara linear, oleh karena itu dilakukan transformasi dimensi data menjadi dimensi tingkat tinggi (*higher dimension*). Kemudian dilakukan SVC untuk mencari pemisah data (*hyperplane*) pada *higher dimension* [18].

Kalkulasi SVC menggunakan yang namanya *kernel function* untuk mencari pemisah data (*hyperplane*). *Kernel function* didefinisikan dengan perkalian titik (*dot product*) antar 2 vektor pada *higher dimension*. *Kernel function* yang dipakai pada umumnya adalah *polynomial* dan *Gaussian* (*radial-basis function* (RBF)), yaitu [20][21][22]

$$K(x_i, x_j) = (1 + x_i x_j)^p \quad (3)$$

Keterangan:

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

$K(x_i, x_j)$: *Kernel function* untuk vektor x_i dan x_j
p : order *polynomial*

$$K(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2 * length^2}\right) \quad (4)$$

Keterangan:
$K(x_i, x_j)$: *Kernel function* untuk vektor x_i dan x_j
exp() : Fungsi eksponensial
$d(x_i, x_j)$: Jarak vektor x_i dan x_j menggunakan e*uclidean distance*
length : Skala yang digunakan untuk kontrol panjang *kernel*, dimana *length* > *0*

Pada umumnya, persamaan 2.4 sering ditulis sebagai:

$$K(x_i, x_j) = \exp\left(-\gamma \, ||x_i, x_j||^2\right) \quad (5)$$

Dimana $\gamma = \frac{1}{2 * length^2}$

Dalam kalkulasi menggunakan *kernel* ada istilah yang disebut sebagai *kernel trick*, yaitu menghitung relasi antar vektor pada *higher dimension* tanpa melakukan transformasi data. Hal ini dilakukan untuk mengurangi komputasi pada perhitungan SVC (pencarian *hyperplane*), karena *kernel trick* menghindari kalkulasi matematika pada transformasi data dari *low dimension* ke *high dimension*.

Metode *Logistic Regression* (LR) berawal dari pengembangan fungsi logistic (*logistic function*) untuk pemodelan pertumbuhan populasi oleh Pierre François Verhulst pada tahun 1838 sampai tahun 1847. Kemudian dilanjutkan pengembangan oleh Pearl-Reed pada tahun 1920 dan dilanjutkan dengan pengembangan statistika lainnya hingga saat ini [23]. Metode LR merupakan bentuk khusus dari *generalized linear model* (GLM), artinya model LR memiliki karakteristik yang mirip dengan model *linear regression*, yang membedakan adalah *linear regression* memiliki nilai minus tak hingga sampai tak hingga [$-\infty, \infty$], sedangkan LR memiliki nilai hasil berupa probabilitas dari nol hingga satu [0,1] [24]. Metode LR melakukan kalkulasi *conditional probability* pada kelas *y* dengan input vektor *x* sesuai dengan persamaan (6) [25]:

$$Proba(y_i \mid x_i) = \frac{1}{1 + \exp\left(-y_i(x_i^T w + c)\right)} \quad (6)$$

Keterangan:
$Proba(y_i \mid x_i)$: *Conditional probability* kelas y_i dengan vektor x_i
exp() : Fungsi eksponensial
w *dan* c : Parameter untuk kontrol hasil *logistic regression*
$x_i^T w$: Perkalian antar vektor x_i dan w

Untuk mendapatkan parameter *w* dan *c* terbaik, dilakukan minimalisasi yang disebut sebagai *logistic regression problem*. Ketika jumlah data (*n*) lebih besar dibandingkan jumlah fitur, maka LR cenderung menghasilkan *overfitting*. Ketika *overfitting* terjadi, banya fitur memiliki bobot yang tinggi tetapi tidak memiliki pengaruh yang signifikan terhadap hasil keluaran (*output*) [25]. Salah satu cara untuk mengurangi *overfitting* pada LR adalah menambahkan *penalty* terhadap *loss function* yaitu ℓ_2 *penalty* :

$$\min_{w,c} \ell_{avg}(w, c) + C||w||^2 \quad (7)$$
$$= \min_{w,c} \frac{1}{n} \sum_{i=1}^{n} \log\left(\exp\left(-y_i(x_i^T w + c)\right) + 1\right) + C \sum_{j=1}^{p} w_j^2$$

Keterangan:
$\min_{w,c}(f(w,c))$: Mencari nilai minimum untuk nilai *w* dan *c* pada fungsi *f(w,c)* pada kasus ini fungsinya adalah ℓ_{avg} ditambah *penalty* ℓ_2
$\ell_{avg}(w,c) + C||w||^2$: *Average logistic loss function* ditambah *penalty* ℓ_2
$\log(\exp(-x) + 1)$: *Logistic loss function* dimana *x* adalah $y_i(x_i^T w + c)$
C : Tingkat regularisasi untuk *penalty* ℓ_2 *regularization* bernilai *real* positif, umumnya bernilai logaritma, yaitu [0.001, 00.1, 0.1, 1, 10, 100] dan seterusnya

Setelah mendapatkan nilai terbaik untuk nilai *w* dan nilai *c*, dilanjutkan dengan menentukan ambang batas (*threshold*) untuk menentukan apakah vektor *x* masuk ke *class true* atau *class false*. Misal:
Class true = 1, *Class false* = -1, *threshold* = 0.6
$Proba(1 \mid x_i) = 0.7$ dan $Proba(-1 \mid x_i) = 0.3$

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

Maka x_i memiliki hasil berupa *class* ***true*** karena probabilitas x_i pada *class* 1 (*true*) lebih besar dibandingkan nilai *threshold* yang ditentukan.

3. HASIL DAN PEMBAHASAN

Pada bagian akan menjelaskan data yang akan dipakai untuk penelitian, kemudian dilanjutkan tahapan penelitian

3.1 Pengumpulan Data

Pendeteksian fake news yang dibuat dengan menggunakan metode klasifikasi ini menggunakan dataset didapat dari *kaggle* terdiri 17 atribut dan 33284 dataset berisi berita palsu dan asli didalam 1 tabel . program akan membaca dataset *fake news* yang diambil dari situs web *kaggle* dengan ekstensi file *csv*

3.2 *Case Folding*

Pada tahap ini, program akan mengubah setiap huruf pada kalimat menjadi huruf kecil, proses ini digunakan untuk membuat seluruh kata menjadi sama (tidak ada beda antara huruf kapital dan huruf kecil), berikut gambar hasil dari case folding

	Unnamed: 0	label	newstext
0	0	0	anti-zuma mp quits south africa's 'corrupt' an...
1	1	0	exclusive: u.s. agencies split over fingerprin...
2	2	0	clinton says trump is most divisive candidate ...
3	3	0	most republicans believe russia is meddling in...
4	4	1	this is how trump could win in a landslide – p...

Gambar 2. Hasil *Case Folding*

3.3 Remove Number & Punctuantion

Pada tahap ini akan dihapus angka dan tanda baca, contoh hasil *remove number* dan *remove punctuation* dapat dilihat pada gambar 3.

	Unnamed: 0	label	newstext
0	0	0	anti zuma mp quits south africa s corrupt an...
1	1	0	exclusive u s agencies split over fingerprin...
2	2	0	clinton says trump is most divisive candidate ...
3	3	0	most republicans believe russia is meddling in...
4	4	1	this is how trump could win in a landslide p...

Gambar 3. Hasil *Remove Number & Punctuantion*

3.4 Stop Word Removal

Pada tahap ini akan dihapus kalimat yang sering muncul sehingga kurang memiliki makna yang penting *Stop Word* merupakan kata umum yang sering digunakan dan tidak membawa informasi. Contoh hasil *stop word removal* dapat dilihat pada gambar 4.

```
'anti zuma quits south africa corrupt anc johannesburg reuters african national congress anc makhosi khoza str
ident critic scandal plagued president jacob zuma quit south africa ruling party thursday labeling nelson mand
ela year old liberation movement alien corrupt year old zulu linguistics expert anc supporter since age denoun
ced zuma july dishonorable disgraceful leader due litany scandals attracted eight years power comments earned
death threats provincial party disciplinary hearing khoza said prepared sit around wait verdict party said wil
lfully blind failings leader charged zuma charging makhosi khoza making mockery rule law making mockery anc co
nstitution said interview sabc state broadcaster charge zuma fire zuma anc know serious self correcting khoza
believed one around anc members parliament voted zuma ultimately unsuccessful aug parliamentary confidence vot
e conducted secret ballot anc spokesman zizi kodwa answer calls mobile phone serious allegations zuma relate f
riendship guptas family indian born businessmen accused using political influence secure lucrative contracts s
tate run companies remain law zuma guptas employ zuma son duduzane director least one companies denied wrongdo
ing say victims politically motivated witch hunt zuma time helm anc comes end december party chooses new leade
r although remain head state unless anc removes early president thabo mbeki khoza fulminated air zuma former w
ife nkosazana dlamini zuma sworn behind closed doors member parliament cementing belief preferred successor ch
allengers led deputy president cyril ramaphosa parliamentary seat gives dlamini zuma chairwoman african union
```

Gambar 4. Hasil Stop Word Removal

3.5 Stemming

Pada tahap ini akan dilakukan perubahan setiap kata pada kalimat, menjadi kata dasarnya, misal “quits” menjadi “quit”. Pada gambar 5 dapat dilihat hasil *stemming* yang didapatkan dari gambar 4.

```
'anti zuma quit south africa corrupt anc johannesburg reuter african nation congress anc makhosi khoza striden
t critic scandal plagu presid jacob zuma quit south africa rule parti thursday label nelson mandela year old l
iber movement alien corrupt year old zulu linguist expert anc support sinc age denounc zuma juli dishonor disg
rac leader due litani scandal attract eight year power comment earn death threat provinci parti disciplinari h
ear khoza said prepar sit around wait verdict parti said will blind fail leader charg zuma charg makhosi khoza
make mockeri rule law make mockeri anc constitut said interview sabc state broadcast charg zuma fire zuma anc
know seriou self correct khoza believ one around anc member parliament vote zuma ultim unsuccess aug parliamen
tari confid vote conduct secret ballot anc spokesman zizi kodwa answer call mobil phone seriou alleg zuma rela
t friendship gupta famili indian born businessmen accus use polit influenc secur lucr contract state run compa
ni remain law zuma gupta employ zuma son duduzan director least one compani deni wrongdo say victim polit moti
v witch hunt zuma time helm anc come end decemb parti choos new leader although remain head state unless anc r
emov earli presid thabo mbeki khoza fulmin air zuma former wife nkosazana dlamini zuma sworn behind close door
member parliament cement belief prefer successor challeng led deputi presid cyril ramaphosa parliamentari seat
give dlamini zuma chairwoman african union commiss addi ababa platform rais profil ahead decemb parti leadersh
```

Gambar 5. Hasil Stemming Gambar

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

3.6 Ubah Dataset Menjadi Balanced dataset

Pada tahap ini, akan dilakukan pengecekkan apakah setiap *class* (*fake news* dan *real news*) seimbang. Jika tidak, pada penelitian ini akan di *down sampling*, yaitu mengurangi sampel pada *class* mayoritas agar seimbang dengan *class* minoritasnya. Pada Gambar III-8 dapat dilihat jumlah *class* untuk *fake news* sebanyak **21,417** dan jumlah *class real news* sebanyak **11,868**. Oleh karena kita akan di *down sampling* menjadi sama-sama 10,000 data *fake news* dan 10,000 data *real news*.

3.7 Feature Extraction

Pada tahap ini, setiap kata akan diubah kedalam bentuk TF-IDF. Tahap ini dilakukan untuk mengubah seluruh data yang awalnya berupa teks, menjadi nilai numerik, dan nantinya akan dibuat dalam bentuk matriks. Pada gambar 6 dapat dilihat hasil *feature extraction*.

```
array([[0.        , 0.10114055, 0.09644771, ..., 0.        , 0.14314976,
        0.        ],
       [0.05874922, 0.        , 0.        , ..., 0.11227699, 0.07290578,
        0.        ],
       [0.        , 0.09146218, 0.        , ..., 0.        , 0.05424789,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.06719487,
        0.        ],
       [0.10065411, 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.0759319 , 0.        , 0.        , ..., 0.08340425, 0.        ,
        0.        ]])
```

Gambar 6. Hasil Feature Extraction

3.8 Save as Pickle file dan Read Pickle File

Pada tahap ini, hasil fitur, label, dan objek TF-IDF akan disimpan dalam ekstensi file *pickle*. Ekstensi *pickle* digunakan untuk menyimpan objek *python* menjadi format *binary* yang nantinya dapat diubah Kembali menjadi objek *python*. Tahap ini berguna untuk menyimpan hasil pra-proses kita kedalam *file* eksternal untuk dipakai pada tahap *training* model dan *testing* model.

3.9 Split Data Train dan Data Testing

Pada tahap ini akan dilakukan split dataset, menjadi *data training* dan *data testing* dengan rasio 75% : 25%. Pada tahap 6, *balancing dataset*, telah diambil data sebanyak 20,000. Dari 20,000 data tersebut maka didapatkan data *training* sebanyak 16,000 data dan *data testing* sebanyak 4,000.

3.10 Split Data Train dan Data Testing

Pada tahap ini akan ditentukan terlebih dahulu *hyperparameter* setiap model klasifikasinya. *Hyperparameter* merupakan parameter yang nilainya digunakan untuk mengontrol proses *training* suatu model. Pada tabel dapat dilihat list *hyperparameter* yang digunakan pada model *random forest*, *support vector machine*, dan *logistic regression*.

Tabel 1. *List parameter* Random Forest

No	Nama Parameter			
	n_estimators	max_depth	min_samples_split	min_samples_leaf
1	200	10	1	0.1
2	600	50	1	1

Tabel 2. *List parameter* Support Vector Machine

No	Nama Parameter		
	C	Gamma	Kernel
1	0.1	1	linear
2	1	10	rbf

Tabel 3. *List parameter* Logistic Regression

No	Nama Parameter	
	C	Penalty
1	0.01	L2
2	0.1	L2
3	1	L2

3.11 GridSearchCV dan Best Parameters

Pada tahap ini akan dilakukan *cross validation* untuk mencari nilai *hyperparameter* terbaik dari setiap model (RF, SVM, dan LR). Hasil dari *cross validation* tersebut akan disimpan kedalam variabel *best parameter* tiap

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

model klasifikasi. Nilai *best parameter* ini yang akan dipakai sebagai *parameter training* setiap model klasifikasi.

3.12 *GridSearchCV* dan *Best Param*

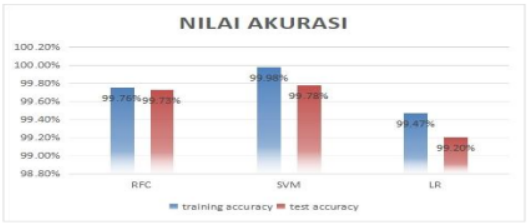
Pada tahap ini akan dilakukan *training* model menggunakan nilai *hyperparameter* terbaik dari hasil *cross validation*. Kemudian dilakukan *testing* menggunakan *data testing* yang telah di *split* sebelumnya.

3.13 Hasil *Persiapan Data*

Meskipun dataset yang baik telah digunakan, tetapi sebelum memasuki tahapan pra-pemrosesan data, perlu untuk mengecek kembali dataset yang digunakan, dengan tujuan untuk memastikan bahwa dataset tersebut sudah siap diolah kedalam tahapan pra-pemrosesan data. Hasil dari persiapan data dalam penelitian ini adalah didapatkan data dengan label untuk masing masing kategori berita tersebut, berita palsu diberi label kategori "*Fake*" dengan bobot nilai nol (1), Sedangkan berita asli diberi label "*Real*" dengan bobot nilai satu (0). Kemudian juga dalam tahapan ini menghasilkan dataframe baru yang di beri label '*Newstext*" dengan isi penggabungan antara dataframe "*title*" sebagai judul berita dan dataframe "*text*" sebagai isi dari berita tersebut. Untuk dataset, akan diambil sebanyak 33285 data berita asli dan data berita palsu.Kemudian kedua data berita tersebut digabungkan dengan catatan menghilangkan data yang muncul lebih dari satu kali, maka didapat data sebanyak 20000 data. Terdiri dari 10000 data set *fake* dan 10000 data set *real*

3.14 Hasil Perbandingan Metode *RFC, SVM, dan LRC*

Hasil dari perbandingan RFC, SVM, LRC dapat dilihat gambar 7 di bawah ini.



Gambar 7. Hasil Nilai Akurasi Klasifikasi RFC, SVM, LR Untuk Dataset

Dari gambar 7 percobaan nilai *Training Accuracy* tertinggi dengan nilai persentase sebesar 99,98% terdapat di SVM sedangkan *Training Accuracy* terendah dengan nilai persentase sebesar 99,47% terdapat di LR. Untuk *Testing Accuracy*, nilai tertinggi dimiliki oleh SVM dengan nilai persentase sebesar 99,78% sedangkan *Training Accuracy* terendah dengan nilai persentase sebesar 99,20% terdapat di LR. Untuk melihat perbandingan hasil setiap skenario percobaan dan setiap dataset, maka digunakanlah *Performance Measurements* untuk mengetahui nilai hasil rata-rata perkelasnya.

Gambar 8. Hasil Perbandingan *Performance Measurement* RFC, SVM, LR.

Gambar 8 merupakan hasil *Performance Measurements* dari skenario percobaan yang memiliki hasil dari klasifikasi LRC untuk dataset, yaitu untuk *Sensitivity* dengan nilai persentase tertinggi sebesar 99,96% terdapat pada LR sedangkan hasil *Sensitivity* dengan nilai persentase terendah sebesar 99,75% terdapat pada SVM. Untuk *Specificity* dengan nilai persentase tertinggi sebesar 99,80% terdapat pada SVM nilai persentase terendah sebesar 99,44% terdapat pada LR. Untuk *Precision* dengan nilai persentase tertinggi sebesar 99,80% terdapat pada SVM sedangkan nilai persentase terendah sebesar 99,45% terdapat pada LR. Untuk *F1-Score* dengan nilai persentase tertinggi sebesar 99,78% sedangkan dengan nilai persentase terendah sebesar 99,21% terdapat pada LR. Untuk *Accuracy* dengan nilai persentase tertinggi sebesar 99,78% terdapat pada SVM sedangkan nilai persentase terendah sebesar 99,20% terdapat pada LR. Untuk *Error* dengan nilai persentase tertinggi sebesar 0.80% terdapat pada LR sedangkan nilai persentase terendah sebesar 0.28 % terdapat pada SVM. Dari perbandingan di atas maka nilai paling baik terdapat pada SVM.

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

4. KESIMPULAN

Berdasarkan hasil dari keseluruhan proses yang dilakukan pada penelitian ini tentang klasifikasi berita palsu yang terdapat di web dengan menggunakan pendekatan dan metode yang diajukan, maka didapatkan kesimpulan *Dataset fake news high dimension* diambil pada situs *https://www.kaggle.com/* sebanyak 32000 data dan 17 atribut. Sistem identifikasi berita palsu dapat dibuat menggunakan metode *Machine Learning* dengan *high dimension* melalui tahapan *Text Preprocessing* yang meliputi, *Case Folding, Punctuation Removal, Number Removal, Removing Word <N Character, Stemming* dan *Lemmatization*, kemudian masuk ke tahapan *Extraction and Selection Feature* dan terakhir masuk ke tahapan klasifikasi dengan menggunakan metode *Random Forest Classifier, Support Vector Machine, Logistic Regression Classifier*. Nilai Training dan Test pada *high dimension* RFC menjelaskan detail hasil *Training Accuracy* 99,76%. Untuk *Testing Accuracy* 99,73%. Dari hasil yang di dapat,*Training Accuracy* hasil akurasi tertinggi. Hasil dari percobaan pada *high dimension* klasifkiasi SVM untuk dataset. Dari percobaan, nilai *Training Accuracy* dengan nilai persentase sebesar 99,78%. Untuk *Testing Accuracy*, dengan nilai persentase sebesar 99,98%. Hasil percobaan, nilai LR pada *high dimension Training Accuracy* dengan nilai persentase sebesar 99,47%. Untuk *Testing Accuracy*, nilai tertinggi dimiliki oleh skenario dengan nilai persentase sebesar 99,20%. Nilai performa terbaik dengan data *high dimension* dari semua percobaan terdapat pada SVM untuk nilai akurasi dataset dengan *Training Set Accuracy Test* persentase 99,97%. *Test Set Accuracy* persentasenya 99,77 %. Untuk hasil *Performance Measurements* dari percobaan yang memiliki hasil tertinggi dari klasifikasi SVM untuk dataset, untuk *Sensitivity* dengan nilai persentase sebesar 99,75%. Untuk *Specificity* nilai persentase sebesar 99,80%. Untuk *Precision* nilai persentase sebesar 99,80%. Untuk *F1-Score* nilai persentase sebesar 99,80%. Untuk *Accuracy* nilai persentase sebesar 99,78%. Untuk *Error* nilai persentase sebesar 0,23%. Nilai performa *Sensitivity* terbaik dari semua skenario percobaan LRC dengan persentase 99,96 % dibandingkan dengan yang lain.

REFERENCES

[1] Y. Y. Chen, S.-P. Yong, and A. Ishak, "Email Hoax Detection System Using Levenshtein Distance Method.," *JCP*, vol. 9, no. 2, pp. 441–446, 2014.
[2] C. D. MacDougall, *Hoaxes*, vol. 465. Dover Publications, 1958.
[3] H. Berghel, "Alt-News and Post-Truths in the" Fake News" Era," *Computer (Long. Beach. Calif).*, vol. 50, no. 4, pp. 110–114, 2017.
[4] T. Petkovic, Z. Kostanjcar, and P. Pale, "E-mail system for automatic hoax recognition," in *27th MIPRO International Conference*, 2005, pp. 117–121.
[5] P. Faustini and T. Covões, "Fake news detection using one-class classification," in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, 2019, pp. 592–597.
[6] S. Zhang, Y. Wang, and C. Tan, "Research on text classification for identifying fake news," in *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 2018, pp. 178–181.
[7] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augment. Hum. Res.*, vol. 5, no. 1, pp. 1–16, 2020.
[8] H. T. Sueno, B. D. Gerardo, and R. P. Medina, "Multi-class document classification using support vector machine (SVM) based on improved Na{\"\i}ve bayes vectorization technique," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, 2020.
[9] D. M. J. Lazer *et al.*, "The science of fake news," *Science (80-.).*, vol. 359, no. 6380, pp. 1094–1096, 2018, doi: 10.1126/science.aao2998.
[10] A. Choudhary and A. Arora, "Linguistic feature based learning model for fake news detection and classification," *Expert Syst. Appl.*, vol. 169, p. 114171, 2021, doi: 10.1016/j.eswa.2020.114171.
[11] T. G. Dietterich and Oregon, "Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Models," *Oncogene*, vol. 12, no. 2, p. pp 1-15(265-275), 1996.
[12] Tin Kam Ho, "Random Decision Forests Tin Kam Ho Perceptron training," *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, pp. 278–282, 1995.
[13] R. Richman and M. V. Wüthrich, "Nagging predictors," *Risks*, vol. 8, no. 3, pp. 1–26, 2020, doi: 10.3390/risks8030083.
[14] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12343 LNCS, pp. 503–515, 2020, doi: 10.1007/978-3-030-62008-0_35.
[15] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.
[16] V. N. Vapnik, *Statistics for Engineering and Information Science Springer Science+Business Media, LLC*. 2000.
[17] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999, doi: 10.1109/72.788640.
[18] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 9014–9022, 2011, doi: 10.1016/j.eswa.2011.01.120.
[19] N. H. Farhat, "Photonit neural networks and learning mathines the role of electron-trapping materials," *IEEE Expert. Syst. their Appl.*, vol. 7, no. 5, pp. 63–72, 1992, doi: 10.1109/64.163674.
[20] M. Seeger, "Gaussian processes for machine learning.," *Int. J. Neural Syst.*, vol. 14, no. 2, pp. 69–106, 2004, doi:

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 4, Oktober 2021, Page 1720-1728
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i4.3177

10.1142/S0129065704001899.

[21] D. Matić, F. Kulić, M. Pineda-Sánchez, and I. Kamenko, "Support vector machine classifier for diagnosis in electrical machines: Application to broken bar," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8681–8689, 2012, doi: 10.1016/j.eswa.2012.01.214.

[22] U. Ravale, N. Marathe, and P. Padiya, "Feature selection based hybrid anomaly intrusion detection system using K Means and RBF kernel function," *Procedia Comput. Sci.*, vol. 45, no. C, pp. 428–435, 2015, doi: 10.1016/j.procs.2015.03.174.

[23] J. S. Cramer, "The Origins of Logistic Regression," *SSRN Electron. J.*, 2005, doi: 10.2139/ssrn.360300.

[24] P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size," *Catena*, vol. 145, pp. 164–179, 2016, doi: 10.1016/j.catena.2016.06.004.

[25] R. Zakharov and P. Dupont, "for Feature Selection," no. May, 2014, doi: 10.1007/978-3-642-24855-9.

Submitted: **29/07/2021**; Accepted: **26/10/2021**; Published: **26/10/2021**

Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)

ORIGINALITY REPORT

10%
SIMILARITY INDEX

PRIMARY SOURCES

1	Dessy Santi, Meri Kristina Tongkuru. "Sistem Informasi Pengarsipan Surat-Surat Pada PT Sinergi Perkebunan Nusantara", Jurnal Ilmiah Intech : Information Technology Journal of UMUS, 2020 Crossref	96 words — 2%
2	repository.ub.ac.id Internet	76 words — 2%
3	www.ukm.edu.my Internet	27 words — 1%
4	ojs.trigunadharma.ac.id Internet	25 words — 1%
5	ejournal.unesa.ac.id Internet	20 words — < 1%
6	vdocuments.site Internet	18 words — < 1%
7	pt.scribd.com Internet	17 words — < 1%

8	www.coursehero.com Internet	16 words — < 1%
9	medscimonit.com Internet	11 words — < 1%
10	doctorpenguin.com Internet	10 words — < 1%
11	jurnal.stts.edu Internet	10 words — < 1%
12	text-id.123dok.com Internet	10 words — < 1%
13	www.biorxiv.org Internet	10 words — < 1%
14	J. Kim. "Geometric-Based Error Concealment for Concealing Transmission Errors and Improving Visual Quality", IEEE Transactions on Circuits and Systems for Video Technology, 8/2006 Crossref	9 words — < 1%
15	Zulfan, Lestari AKA, Dewi Maya Sari. "EFEKTIVITAS PENERAPAN UNDANG-UNDANG ITE TERHADAP PELAKU PENYEBARAN HOAKS COVID-19 DI MEDIA SOSIAL", Jurnal Transformasi Administrasi, 2021 Crossref	9 words — < 1%
16	mikrodata.bps.go.id Internet	9 words — < 1%
17	repositori.usu.ac.id Internet	9 words — < 1%
18	www.atlasconference.org	

	Internet	9 words — < 1%
19	1library.net Internet	8 words — < 1%
20	core.ac.uk Internet	8 words — < 1%
21	repository.unri.ac.id Internet	8 words — < 1%
22	"Web Information Systems Engineering – WISE 2020", Springer Science and Business Media LLC, 2020 Crossref	6 words — < 1%
23	journal.upgris.ac.id Internet	6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF