Deteksi Konten *Hoax* Berbahasa Indonesia di Twitter Menggunakan Fitur Ekspansi dengan *Word2Vec*

Friskadini Ismayanti[1], Erwin Budi Setiawan[2]

[1,2] Universitas Telkom, Bandung
[1]friskadiniism@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstrak

Penggunaan Internet di Indonesia semakin terus meningkat, dikarenakan semakin banyak orang yang dapat mengakses Internet maka semakin banyak juga penyalahgunaan dalam Internet itu sendiri khususnya media sosial. Perlu untuk dapat mendeteksi adanya *hoax* di media sosial seperti *Twitter*. *Twitter* sebagai salah satu media sosial memiliki peran yang signifikan dalam penyebaran informasi, semua orang dapat memberikan respons terhadap informasi yang didapat. Dalam isi sebuah *tweet* memungkinkan terjadinya ketidakcocokan kosakata. Oleh karena itu pada penelitian ini dilakukan penerapan metode *Feature Expansion Word2Vec* untuk mengatasi terjadinya ketidakcocokan kosakata. Penelitian ini melakukan pengembangan dan perbandingan sistem klasifikasi *hoax* Twitter menggunakan metode *Feature Expansion Word2Vec* dengan algoritma klasifikasi *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest* dan sistem tanpa metode *Feature Expansion Word2Vec*. Hasil dari penelitian ini, metode *Feature Expansion Word2Vec* pada algoritma klasifikasi Random Forest berhasil meningkat akurasi sistem sebesar 1,46% dengan nilai akurasi sebesar 89,53%.

Kata kunci : ***Word2Vec, Random Forest, Feature Expansion, Twitter, Klasifikasi Hoax***

Abstract

Internet use in Indonesia continues to increase, because the more people who can access the internet, the more abuse in the internet itself, especially social media. It is necessary to be able to detect hoaxes on social media such as Twitter. Twitter as one of the social media has a significant role in the dissemination of information, everyone can respond to the information obtained. In the content of a tweet allows for vocabulary incompatibility. Therefore, in this study, the application of the Word2Vec Feature Expansion method to overcome vocabulary incompatibility. This research conducted the development and comparison of Twitter's hoax classification system using the Word2Vec Feature Expansion method with Logistic Regression, Support Vector Machine (SVM), Random Forest and system without the Word2Vec feature expansion method. As a result of this study, Word2Vec's Feature Expansion method on the Random Forest classification algorithm managed to increase system accuracy by 1.46% with an accuracy value of 89.53%.

Keywords: ***Word2Vec, Random Forest, Feature Expansion, Twitter, Hoax Classification***

1. Pendahuluan

Media sosial adalah tempat orang membuat konten, berbagi konten, menandai konten, dan membangun jaringan dengan kecepatan luar biasa. Karena kemudahan penggunaan, kecepatan, dan jangkauannya, media sosial dengan cepat mengubah wacana publik di masyarakat dan memimpin tren dengan topik mulai dari lingkungan dan politik hingga teknologi dan hiburan [1]. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menulis bahwa jumlah pengguna aktif Internet di Indonesia sudah mencapai 196,7 juta [2]. Sehingga dengan penggunaan Internet di Indonesia semakin meningkat dapat mengakibatkan juga penyalahgunaan dalam Internet itu sendiri, khususnya media sosial. Contoh di *Twitter* banyak sekali cuitan atau *tweet* yang mengandung *hoax* baik secara individu maupun secara kelompok yang juga dapat mengakibatkan dampak negatif baik secara individu maupun kelompok.

Hoax adalah informasi yang tidak dapat dipercaya karena yang disampaikan adalah informasi palsu tetapi dianggap sebagai kebenaran. *Hoax* mampu mempengaruhi suatu citra dan kredibilitas pada banyak orang [3]. Berita *hoax* menyebar lebih cepat daripada berita sebenarnya. Meskipun *hoax* tidak dibuat untuk merusak program pada komputer dan sistem operasi, namun *hoax* dapat menghancurkan reputasi seseorang dan perusahaan yang bersangkutan, rekan kerja, teman, atau bahkan menghasilkan kerugian finansial. Penyebaran *hoax* pada perusahaan menghasilkan kerugian rata-rata mendekati $500.000 per perusahaan [4].

Twitter adalah mikroblog yang berbasis sebuah website sosial media yang diluncurkan pada tanggal 13 Juli 2006 [5]. Menurut *Twitter*, ada rata-rata 340 juta *tweet* yang telah dihasilkan masing-masing pada bulan Maret 2012 [6]. Keunikan *Twitter* adalah sebuah *tweet* memiliki limit sebanyak 140 karakter agar bisa secara maksimal pertukaran informasi menggunakan karakter sesedikit mungkin [7].

Peneliti menggunakan informasi penyebaran data dari *Twitter*, menganalisis, dan kemudian menentukan informasi tersebut bohong atau tidak. Untuk menentukan informasi tersebut, penelitian tentang sistem klasifikasi *hoax* telah dilakukan oleh beberapa peneliti. Salah satu studi penelitian membandingkan tiga algoritma, yaitu algoritma *Naïve Bayes*, algoritma *Support Vector Machine* (SVM), dan algoritma *Decision Tree* dalam klasifikasi *hoax*. Dan hasilnya adalah bahwa algoritma *Naive Bayes* menunjukkan akurasi terbaik, yaitu sebesar 91,36% [8]. Penelitian metode klasifikasi *hoax* lainnya adalah menggunakan SVM dengan bobot TF-IDF yang menunjukkan akurasi 95,83% [9]. Pada penelitian [10] yang menerapkan algoritma *Decision Tree* tingkat akurasinya mencapai 73,91%. Pada penelitian [11] yang menggunakan *Random Forest*, hasil akurasi tertinggi hingga 95%. Terdapat juga penelitian sebelumnya yang pernah dilakukan menggunakan metode *Word2Vec*, TF-IDF. Pendekatan standar mempelajari representasi kata adalah untuk melatih model log-biliner berdasarkan skip-gram yang diimplementasikan dalam *Word2Vec* [12] menghasilkan sebuah model dari representasi kata dan frasa terdistribusi serta komposisionalitasnya. Dalam sistem pendeteksi *hoax* hal yang dapat dilakukan yaitu dilakukannya beberapa tahap seperti memisahkan kata dan kemudian membandingkan kata tersebut dengan kata-kata yang sebelumnya sudah ada.

Oleh karena itu, pada penelitian ini penulis membahas apa pengaruh dari peformansi sistem yang diterapkan fitur ekspansi menggunakan *Word2Vec* pada algoritma *Support Vector Machine* (SVM), *Logistic Regression*, dan *Random Forest*.

Terdapat beberapa batasan penelitian dalam penulisan ini, yaitu data yang digunakan adalah data *hoax* dalam Bahasa Indonesia sebanyak 26984 *tweet* mengenai beberapa topik di Indonesia, proses pelabelan data dilakukan secara manual menjadi dua kategori, yaitu *hoax* dan *non hoax*, nilai matriks performansi yang digunakan adalah akurasi dan F1- Score, serta penggunaan *word embedding Word2Vec*.

Tujuan dari penelitian ini adalah bagaimana mengimplementasikan, mengukur nilai performansi khususnya akurasi dan F1 – Score, serta menganalisis hasil sistem klasifikasi data *hoax* yang dibangun menggunakan teknik *feature expansion* dan *word embedding Word2Vec* pada data *hoax* Bahasa Indonesia dalam *tweet* yang memiliki beberapa topik.

Tugas Akhir ini disusun dengan struktur berikut, pada Bab 2 akan dibahas teori atau studi literatur yang mendukung atau berkaitan dengan penelitian ini, Bab 3 membahas teori terkait penelitian dan pemodelan sistem yang dibangun. Pada Bab 4 membahas hasil eksperimen, analisis dan evaluasi model penelitian. Terakhir pada Bab 5 dijelaskan kesimpulan dan saran untuk penelitian lanjutan.

2. Studi Terkait

Terdapat penelitian menggunakan *feature expansion* untuk *sentiment analysis* pada Twitter. Pada penelitian ini [13], terdapat 3 algoritma klasifikasi yang digunakan yaitu *Support Vector Machine* (SVM), *Logistic Regression* dan *Naïve Bayes* penelitian ini menggunakan 2 *feature expansion* yaitu TF – IDF dan ekspansi berbasis tweet agar dapat meningkatkan akurasi dari sistem analisis tersebut. Peningkatan akurasi yang signifikan terjadi pada ekspansi berbasis tweet dengan akurasi 98,81% menggunakan *Logistic Classifier*.

Pada penelitian ini [6], fitur ekspansi yang digunakan ialah *word embedding* berdasarkan dari *Word2Vec* untuk mengurangi ketidakcocokan kosakata untuk topik klasifikasi *tweet*, dengan menggunakan 3 algoritma sebagai klasifikasi yaitu *Support Vector Machine* (SVM), *Logistic Regression* dan *Naïve Bayes*. Pada penelitian tersebut juga digunakan *corpus* tambahan untuk ekspansi fitur yaitu *IndoNews* dan *GoogleNews*, dan penggunaan *corpus GoogleNews* meningkatkan performansi secara konsisten pada klasifikasi *Logistic Regression* dan performansi menggunakan *Corpus GoogleNews* lebih baik dibandingkan dengan *IndoNews*.

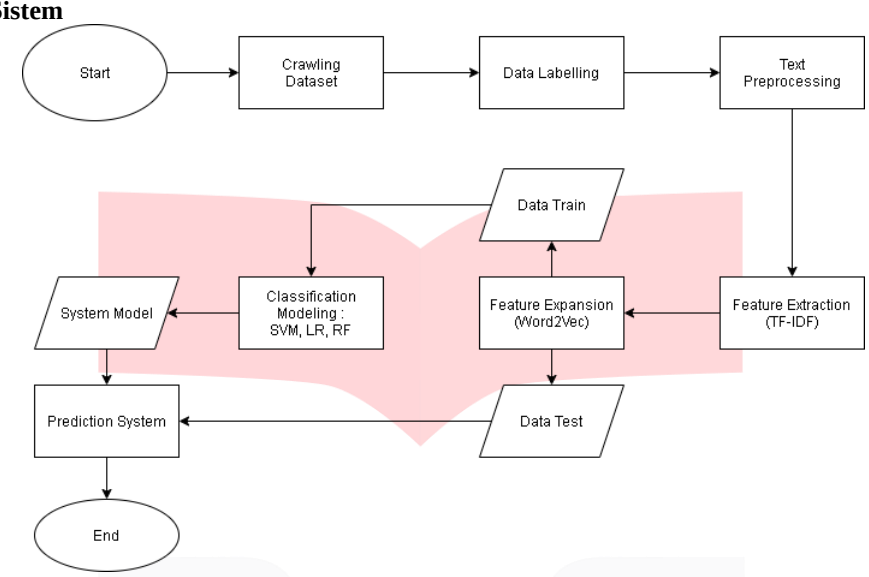
Pada penelitian ini [14], dilakukan sebuah penelitian mengenai komparasi *Random Forest, Naïve Bayes, Decision Tree, Support Vector Machines* (SVM), dan *Logistic Regression* untuk *multi class* teks klasifikasi, dan didapatkan bahwa klasifikasi *Logistic Regression* dengan akurasi (min 32,43%, max 58,50%).

Setelah mempelajari beberapa penelitian, belum ada penggunaan ekspansi fitur berbasis *word embedding Word2Vec* dalam klasifikasi *hoax* dengan algoritma SVM, *Logistic Regression*, *Random Forest*. Maka daripada itu pada penelitian ini menggabungkan fitur ekspansi *Word2Vec* dengan ketiga algoritma tersebut.

3. Sistem yang Dibangun

Sistem klasifikasi *hoax* yang akan dibangun pada penelitian ini dapat dilihat pada Gambar 1.

3.1 Gambaran Sistem



Gambar 1 Sistem Klasifikasi *Hoax* Menggunakan *Feature Expansion Word2Vec*

3.2 Crawling dan Pelabelan Data

Pada penelitian ini dataset yang dibangun dengan cara *crawling* data di *Twitter* dengan menggunakan *Application Program Interface* (API) yang telah disediakan oleh *Twitter*. Mengakses data *Twitter* dengan mendaftarkan akun sebagai pengembang dan menjelaskan alasan ingin mendapatkan akses ke API, dalam hal ini untuk penelitian akademis. Menggunakan data yang terkumpul kurang lebih 26984 *tweet* dengan kata kunci dan hashtag. Data tersebut akan dimasukkan dalam format CSV, dengan label 0 merepresentasikan non *hoax* dan 1 merepresentasikan *hoax*.

Tabel 1 Contoh *Tweet* Dalam Dataset

Tweet	Label
Diduga 53 Miliar Raib dari APBD Provinsi Gorontalo 2019	*Hoax*
Mari kita memberikan kritikan dan masukan yang baik dan benar bukan dengan cara fitnah atau menyebarkan berita bohong.	Non *Hoax*

3.3 Preprocessing

Preprocessing teks berarti pembersihan kebisingan seperti pembersihan kata-kata berhenti, tanda baca, istilah yang tidak membawa banyak bobot dalam konteks teks [15]. Dalam mempersiapkan *dataset* dibutuhkan tahap *preprocessing* data. Tujuan melakukan tahap *preprocessing* ini dilakukan agar meningkatkan kualitas data saat digunakan untuk pelatihan model data *hoax*. Pada proses ini digunakan *library String*, *NLTK* dan *Sastrawi*. Berikut beberapa tahap dalam *preprocessing*:

1. *Case Folding*
   Teknik ini digunakan untuk mengubah kalimat menjadi huruf kecil yang dibantu oleh *library* String
2. *Cleaning*

Teknik ini digunakan untuk menghilangkan semua angka, tanda baca, url ("http://", "www…com"), hastag (#), dan username (@username)

3. Normalisasi kata
   Teknik ini digunakan untuk mengubah kata singkatan, salah penulisan (typo), kata gaul, dan kata alay (informal) menjadi kata formal dengan bantuan kamus kata yang dibuat secara manual.
4. *Stopwords Removal*
   Teknik ini digunakan untuk menghapus kata-kata umum yang digunakan dan kata yang tidak memiliki arti khusus seperti kata ganti, preposisi, dan konjungsi. Proses penghapusan *stopwords* dalam Bahasa Indonesia digunakan *library* Phyton Natural Language Toolkit (NLTK).
5. *Stemming*
   Teknik ini digunakan untuk mengubah kata kembali dalam bentuk kata dasarnya dengan membuang imbuhan awal maupun akhir. Proses *stemming* menggunakan sebuah *library* khusus untuk pemprosesan Bahasa Indonesia, yaitu *library* Python Sastrawi.
6. *Tokenization*
   Teknik ini digunakan untuk memisah sebuah kalimat mejadi kata per kata yang biasa disebut *token*.

3.4 Word2Vec

Metode *Word2Vec* memiliki tujuan menemukan interaksi tersembunyi dalam sebuah kata. Setiap kata mewakili distribusi bobot dalam elemennya. Metode ini mempunyai dua model arsitektur, yaitu *Continous Bag of Words* (CBOW) dan *Skip Gram* [16]. *Word2Vec* adalah metode *embedding word* yang bermanfaat untuk merepresentasikan kata sebagai sebuah vektor dengan panjang N. Pada penelitian ini *Word2Vec* contoh model yang dipakai adalah *Continous Bag of Words* (CBOW) ditambah dengan ekspansi fitur.

3.5 *Term Frequency – Inverse Document Frequency* (TF-IDF)

TF – IDF merupakan proses dimana dilakukannnya transformasi data menurut data tekstual kepada data numerik agar dilakukan pembobotan dalam setiap kata atau fitur. TF- IDF ini merupakan ukuran statistik yang digunakan untuk mengevaluasi seberapa krusial sebuah kata pada sebuah dokumen [17]. Bobot kata akan semakin kecil apabila jarang muncul dalam suatu dokumen dan semakin besar apabila sering muncul dalam dokumen [18].

3.6 Logistic Regression

Logistic Regression merupakan metode analisis statistika agar dapat menggambarkan interaksi antara peubah respons (*dependent variable*) yang mempunyai dua kategori atau lebih menggunakan satu atau lebih peubah penjelas (*independent variable*) berskala interval atau kategori. *Logistic Regression* adalah regresi non linear, digunakan agar dapat menjelaskan hubungan antara X dan Y yang bersifat tidak linear, ketidaknormalan Y, keragaman respons non konstan yang tidak dapat dijelaskan menggunakan contoh regresi linear biasa [19].

3.7 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan teknik pembelajaran *supervised* untuk mengklasifikasikan aneka macam kategori data [20]. Secara sederhana konsep SVM merupakan sebuah usaha mencari *hyperplane* terbaik yang mempunyai peran krusial sebagai garis batas dua butir *class*. SVM mencari *hyperplane* menurut *support vectors* dan margin. *Support vectors* merupakan semua vektor data yang berjarak paling mendekati *hyperplane*, sedangkan margin menyatakan lebar menurut *separating hyperplane*. Tujuan SVM merupakan membuat sebuah contoh klasifikasi berupa fungsi sign(x), f(x)=y, supaya dapat mengklasifikasikan data dalam proses *testing* [21].

3.8 Random Forest

Random Forest merupakan metode *ensemble* yang dipakai dalam membangun model prediktif untuk masalah regresi dan klasifikasi [22], karena dalam menciptakan simpul anak pada setiap node dilakukan secara acak. Metode *random forest* adalah kumpulan metode pembelajaran memakai *decision tree* sebagai pengklasifikasi yang buat dan juga digabungkan. Ada tiga aspek dari metode hutan acak, *bootstrap sampling* yang digunakan membangun pohon prediktif. Setiap *decision tree* memprediksi dengan prediktor acak kemudian *Random Forest* membuat prediksi dengan cara menggabungkan hasil dari setiap pohon keputusan melalui sebagian besar suara untuk klasifikasi atau regresi.

3.9 Confusion Matrix

Confusion Matrix merupakan alat visualisasi yang digunakan pada pembelajaran *supervised*. Tiap kolom yang terdapat matriks adalah contoh yang terdapat pada kelas prediksi, sedangkan tiap baris merupakan kejadian di kelas yang sebenarnya. Hasil pada *confusion matrix* terdapat 4 keluaran, yaitu *recall*, *precision*, *accuracy*, dan *error rate* (Tabel 2).

Tabel 2 Tabel *Confusion Matrix*

Prediksi			
Aktual	*Hoax*	TN (True Negative)	FP (False Positive)
	Non *Hoax*	FN (False Negative)	TP (True Positive)

Keterangan:
TP = Jumlah data positif dan diprediksi benar.
TN = Jumlah data negatif yang diprediksi benar
FP = Jumlah data negatif namun diprediksi data positif.
FN = Jumlah data positif namun diprediksi data negatif

Precision merupakan nilai perbandingan antara data yang diminta dengan hasil prediksi model tersebut.

$$Precision = \frac{TP}{(TP + FN)} \times 100\% \quad (1)$$

Recall merupakan nilai yang menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi

$$Recall = \frac{TP}{(TP + FN)} \times 100\% \quad (2)$$

Akurasi merupakan seberapa akurat dapat mengklasifikasikan model dengan benar.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \quad (3)$$

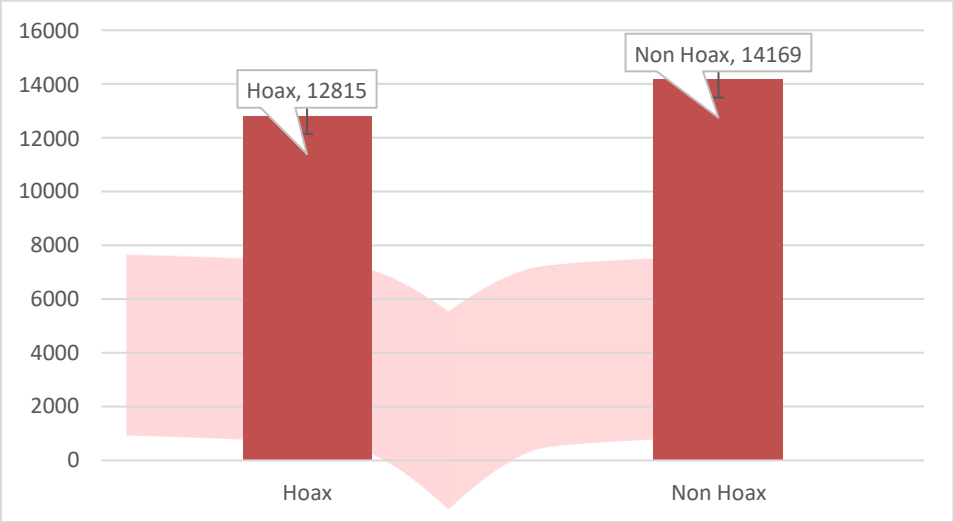
F1 – Score merupakan perbandingan rata-rata dari nilai *precision* dan nilai *recall* yang dibobotkan.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. Evaluasi

4.1 Data

Data *tweet* yang telah didapatkan berjumlah 26984 *tweet* berbahasa Indonesia terkait dengan bahasan *hoax* di Indonesia yang akan menjadi data *train* dan data *test*. Data ini terdiri dari beberapa *keyword* berbeda yang berkaitan dengan hastag, #vaksin, #presiden, #nkri, #mudik, #paspampres, #wakilbupati sangihe, #gajijakarta. Dengan jumlah penyebarannya data yang telah diberi label *hoax* dan *non hoax* dapat dilihat pada Gambar 2.



Gambar 2 Persebaran Data

Kemudian, terdapat data pelengkap pembuatan kamus kata menggunakan data yang diambil dari beberapa media berita seperti CNN Indonesia, Sindonews, Kompas, Tempo, Detik.com, Liputan6, dan Republika sebanyak 142.544 data. Komposisi data (*IndoNews*) yang digunakan untuk pembuatan kamus *similarity* dengan *word embedding Word2Vec* yang dapat dilihat pada Tabel 3.

Tabel 3 Data *IndoNews*

Nama Redaksi	Jumlah
CNN Indonesia	29349
Republika	53812
Kompas	15055
Tempo	15055
SindoNews	13702
Detik.com	7974
Liputan6	251
Total	**142544**

4.2 Preprocessing Data

Preprocessing dilakukan dengan tujuan memastikan dataset siap diproses. Beberapa tahapan yang harus dilakukan dalam preprocessing, dapat dilihat melalui Tabel 4 sebagai berikut.

Tabel 4 Preprocessing Data

Preprocessing	*Input*	*Output*
Case folding	@MrsRachelIn Mantap bapak bapak dewan yang terhormat maju Terus seret orang2 yang menyebar berita bohong dan fitnah Dana Haji ke gedung DPR minta pertanggung jawab mereka	@mrsrachelin mantap bapak bapak dewan yang terhormat maju terus seret orang 2 yang menyebar berita bohong dan fitnah dana haji ke gedung dpr minta pertanggung jawab mereka
Cleaning	@mrsrachelin mantap bapak bapak dewan yang terhormat maju terus seret orang 2 yang menyebar berita bohong dan fitnah dana haji ke gedung dpr minta pertanggung jawab mereka	mantap bapak bapak dewan yang terhormat maju terus seret orang yang menyebar berita bohong dan fitnah dana haji ke gedung dpr minta pertanggung jawab mereka

Normalisasi	mantap bapak bapak dewan yang terhormat maju terus seret orang yang menyebar berita bohong dan fitnah dana haji ke gedung dpr minta pertanggung jawab mereka	mantap bapak bapak dewan yang terhormat maju terus seret orang yang menyebar berita bohong dan fitnah dana haji ke gedung dpr minta pertanggung jawab mereka
Stopword Removal	mantap bapak bapak dewan yang terhormat maju terus seret orang yang menyebar berita bohong dan fitnah dana haji ke gedung dpr minta pertanggung jawab mereka	mantap bapak bapak dewan terhormat maju terus seret orang menyebar berita bohong fitnah dana haji gedung dpr minta pertanggung jawab
Steamming	mantap bapak bapak dewan terhormat maju terus seret orang menyebar berita bohong fitnah dana haji gedung dpr minta pertanggung jawab	mantap bapak bapak dewan hormat maju terus seret orang sebar berita bohong fitnah dana haji gedung dpr minta tanggung jawab
Tokenization	mantap bapak bapak dewan hormat maju terus seret orang sebar berita bohong fitnah dana haji gedung dp minta tanggung jawab	‘mantap’, ‘bapak’, ‘bapak’, ‘dewan’, ‘hormat’, ‘maju’, ‘terus’, ‘seret’, ‘orang’, ‘sebar’, ‘berita’, ‘bohong’, ‘fitnah’, ‘dana’, ‘haji’, ‘gedung’, ‘dpr’, ‘minta’, ‘tanggung’, ‘jawab’

4.3 Pembuatan Kamus Kata (*Corpus)*

Pembuatan *corpus* menggunakan teknik *word embedding Word2Vec* model *Continuous Bag of Words*. *Corpus* berupa kumpulan kata yang diurutkan nilai similaritasnya dari tertinggi hingga terendah sehingga didapatkan hasil yang terdapat pada Tabel 5-7.

1. *Corpus* Data *Tweet*

   Corpus data *tweet* mendapatkan hasil kosakata sebanyak 21.629 kata dan contoh dari nilai similaritasnya dapat dilihat pada Tabel 5.

Tabel 5 *Corpus Tweet*

Kata	**Ranking 1**	**Ranking 2**	**Ranking 3**	**Ranking 4**	**Ranking 5**
Provinsi	kabupaten	bandung	timur	riau	psbb
	Ranking 6	**Ranking 7**	**Ranking 8**	**Ranking 9**	**Ranking 10**
	kotawaringin	cikampek	trans	bal	wilayah

2. *Corpus* Data Berita (*IndoNews*)

   Corpus data berita (*IndoNews*) mendapatkan hasil kosakata sebanyak 225.623 kata dan contoh dari similaritasnya dapat dilihat pada Tabel 6.

Tabel 6 *Corpus IndoNews*

Kata	**Ranking 1**	**Ranking 2**	**Ranking 3**	**Ranking 4**	**Ranking 5**
Perintah	pemda	mesti	bijak	wenang	pemprov
	Ranking 6	**Ranking 7**	**Ranking 8**	**Ranking 9**	**Ranking 10**
	desentralisasi	realisasi	regulasi	fiskal	urgensi

3. *Corpus* Data *IndoNews* dan *Tweet*

   Corpus data *IndoNews* dan *tweet* mendapatkan hasil kosakata sebanyak 232.733 kata dan contoh dari similaritasnya dapat dilihat pada Tabel 7.

Tabel 7 *Corpus IndoNews + Tweet*

Kata	**Ranking 1**	**Ranking 2**	**Ranking 3**	**Ranking 4**	**Ranking 5**
usaha	bisnis	entitas	industri	bumn	operator
	Ranking 6	**Ranking 7**	**Ranking 8**	**Ranking 9**	**Ranking 10**
	umkm	emiten	persero	swasta	karyawan

4.4 Feature Expansion

Sebagai contoh untuk eksperimen *feature expansion* dengan ukuran fitur *top 1 similarity* menggunakan *Corpus IndoNews*, *Corpus Tweet* dan *Corpus IndoNews + Tweet*, pada representasi vektor fitur TF-IDF kata “kota”

memiliki bobot nol berbeda pada dokumen, *Corpus tweet* tersebut dan juga *Corpus IndoNews + Tweet*, dapat dilihat pada Tabel 8.

Tabel 8 *Feature Expansion* Contoh Kata pada *Corpus*

Contoh Kata	Top Similarity	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Kota	1	('bandung', 0.959925651550293)	('kotamadya', 0.7348856925964355)	('ibukota', 0.6510027050971985)
Provinsi	1	('kabupaten', 0.9924415349960327)	('propinsi', 0.779967188835144)	('propinsi', 0.8044645190238953)
Dana	1	('ribu', 0.9889853596687317)	('duit', 0.6612828373908997)	('duit', 0.6636890769004822)

4.5 Klasifikasi

Seletah melalui tahap *preprocessing*, pembobotan kata, dan proses *feature expansion*, kemudian proses dilanjutkan ke tahap klasifikasi menggunakan *Support Vector Machine* (SVM), *Logistic Regression*, dan *Random Forest*. Sebelum itu dilakukan penggantian rasio data latih dan data uji terlebih dahulu untuk masing-masing algoritma agar hasil yang diberikan lebih optimal. Pada tiap sistem klasifikasi dilakukan pengulangan eksekusi program sebanyak 5 kali yang diambil nilai rata-rata akurasinya dan menggunakan perbandingan data latih dan data uji 80:20. Lalu juga diambil paling tinggi akurasinya dari percobaan penggantian rasio pada data latih dan data uji.

4.6 Hasil Pengujian dan Skenario

Pengujian ini dilakukan agar dapat menganalisis tingkat keberhasilan dari performasi sistem yang dibuat untuk melihat kinerja identifikasi berita *hoax* pada media sosial *Twitter*, dan mengetahui pengaruh ekstraksi fitur dan fitur seleksi dalam pengujian ini.

Pengujian terhadap performa kinerja sistem menggunakan *Bag of Words* (BOW) dalam *unigram* tanpa ekspansi fitur dan ekstraksi fitur. Hasil pengujian dapat dilihat pada Tabel 9. Pengujian dilakukan sebanyak 5 kali dengan data uji 10% dan 20%, untuk menemukan rasio data uji untuk pengujian selanjutnya nanti.

Tabel 9 Hasil Peformansi BOW pada Data Uji

Rasio	*Logistic Regression*		*Support Vector Machine* (SVM)		*Random Forest*	
	Akurasi (%)	F1 - Score	Akurasi (%)	F1 - Score	Akurasi (%)	F1 - Score
90 : 10	82,40	0.8238	83,51	0,8350	**84,21**	**0,8417**
80 : 20	**82,56**	**0,8253**	**83,95**	**0,8392**	83.94	0.8391

Dari Tabel 9 dapat disimpulkan bahwa dengan rasio 20% data uji nilai akurasi rata-rata dari masing-masing klasifikasi memiliki nilai yang lebih baik dibandingkan dengan 10% data uji. Pada pengujian berikutnya, pengujian performa kinerja sistem menggunakan *Baseline Word2Vec* dan juga menggunakan pembobotan dengan TF–IDF tanpa menggunakan *upsampling minority class* dan ekspansi fitur dengan masing-masing menggunakan 1000 fitur baik di *Word2Vec* untuk *vector size* dan *max features* untuk TF–IDF. Hasil peformansi pada *Wor2Vec* dan *Word2Vec Upsampling* dapat dilihat pada Tabel 10. Pengujian dilakukan sebanyak 5 kali dengan data uji 20%.

Tabel 10 Perbandingan Peformansi pada *Word2Vec* dan *Word2Vec Upsampling*

Metode	*Logistic Regression*		*Support Vector Machine* (SVM)		*Random Forest*	
	Akurasi (%)	F1 - Score	Akurasi (%)	F1 - Score	Akurasi (%)	F1 - Score
Word2Vec	80,44	0,8041	82,19	0,8216	82,36	0,8233

***Word2Vec Upsampling* (Baseline)**	**80,73**	**0.8073**	**83,90**	**0.8384**	**88.24**	**0.8823**

Pada pengujian berikutnya, pengaruh pengujian performa kinerja sistem dengan TF - IDF dan menggunakan *upsampling minority class.* Hasil dapat dilihat pada Tabel 11. Pengujian dilakukan sebanyak 5 kali dengan data uji dengan data uji 20%.

Tabel 11 Hasil Peformansi pada TF - IDF *Upsampling Minority Class*

Logistic Regression		*Support Vector Machine* (SVM)		*Random Forest*	
Akurasi (%)	F1 - Score	Akurasi (%)	F1 - Score	Akurasi (%)	F1 - Score
82,79 (+2,55)	0,8278	86,81 (+3,47)	0,8681	88,33 (+0,10)	0,88321

Selanjutnya, dilakukan scenario pengujian membandingkan peformansi menggunakan *feature expansion* dengan kamus kata data *tweet*, *feature expansion* dengan kamus kata data berita, dan *feature expansion* dengan kamus kata gabungan dari data berita, data *tweet* dan (data berita + *tweet*). Pada setiap scenario, pengujian juga dilakukan pada pembagian ratio data latih dan data uji sebesar 80:20. Pengujian dilakukan untuk tiap pengambilan fitur 1, 5, 10 kata yang mempunyai similaritas tertinggi atau *Top Similarity* dari kamus kata yang telah dibuat. Pada setiap klasifikasi, juga dilakukan percobaan sebanyak 5 kali dan diambil nilai dari rata-rata akurasinya.

Hasil akurasi dan *F1 – score* setelah menggunakan *feature expansion* dengan ukuran fitur 1000 pada masing-masing klasifikasi pada masing masing algoritma dapat ditunjukkan pada Tabel 9-11. Kolom *Corpus Tweet*, *Corpus IndoNews*, *Corpus IndoNews + tweet*, secara berurutan masing masing mendekskripsikan hasil nilai akurasi dari pengujian feture expansion menggunakan kamus data *tweet*, kamus data berita, dan kamus kata gabungan dari berita dan *tweet*.

Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Logistic Regression* dapat dilihat dari Tabel 12. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 10 Similarity* menggunakan *Corpus IndoNews* sebesar 83,33%.

Tabel 12 Hasil Peformansi *Feature Expansion* pada *Logistic Regression* 1000 Fitur

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	80,73	83,10 (+2,94)	83,10 (+2,94)	83,10 (+2,94)
Top 5	80,73	83,01 (+2,82)	82,65 (+2,37)	83,01 (+2,82)
Top 10	80,73	83,00 (+2,81)	**83,33 (+3,22)**	83,22 (+3,08)

Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Support Vector Machine* dapat dilihat dari Tabel 13. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 10 Similarity* menggunakan *Corpus Tweet* sebesar 87,34%.

Tabel 13 Hasil Peformansi *Feature Expansion* pada *Support Vector Machine* 1000 Fitur

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	83,90	87,01 (+3,71)	86,98 (+3,67)	87,19 (+3,92)
Top 5	83,90	87,19 (+3,92)	87,09 (+3,80)	87,09 (+3,79)
Top 10	83,90	**87,34 (+4,10)**	87,27 (+4,01)	87,33 (+4,08)

Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Random Forest* dapat dilihat dari Tabel 14. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 10 Similarity* menggunakan *Corpus IndoNews + Tweet* sebesar 88,75%.

Tabel 14 Hasil Peformansi *Feature Expansion* pada *Random Forest* 1000 Fitur

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	88,24	88,52 (+0,32)	88,38 (+0,16)	88,51 (+0,31)
Top 5	88,24	88,34 (+0,11)	88,28 (+0,05)	88,45 (+0,24)
Top 10	88,24	88,70 (+0,52)	88,65 (+0,47)	**88,75 (+0,58)**

Pada pengujian berikutnya, pengaruh ukuran fitur menjadi 2000 fitur pada *Corpus Tweet, Corpus IndoNews, Corpus IndoNews + Tweet.* Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Logistic Regression* dapat dilihat dari Tabel 15. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 1 Similarity* menggunakan *Corpus IndoNews* sebesar 83,70%.

Tabel 15 Hasil Peformansi *Feature Expansion* pada *Logistic Regression* 2000 Fitur

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	80,73	83,60 (+3,54)	**83,70 (+3,67)**	83,47 (+3,39)
Top 5	80,73	83,40 (+3,31)	83,53 (+3,47)	83,35 (+3,24)
Top 10	80,73	83,62 (+3,57)	83,68 (+3,64)	83,43 (+3,34)

Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Support Vector Machine* dapat dilihat dari Tabel 16. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 10 Similarity* menggunakan *Corpus IndoNews* sebesar 87,63%.

Tabel 16 Hasil Peformansi *Feature Expansion* pada *Support Vector Machine* 2000 Fitur

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	83,90	87,56 (+4,36)	87,47 (+4,25)	87,43 (+4,20)
Top 5	83,90	87,45 (+4,23)	87,43 (+4,20)	87,44 (+4,22)
Top 10	83,90	87,60 (+4,41)	**87,63 (+4,45)**	87,47 (+4,25)

Hasil pengujian peformansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Random Forest* dapat dilihat dari Tabel 17. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 1 Similarity* menggunakan *Corpus IndoNews* sebesar 89,03%.

Tabel 17 Hasil Peformansi *Feature Expansion* pada *Random Forest* 2000 Fitur

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	88,24	88,88 (+0,73)	**89,03 (+0.90)**	88,72 (+0,55)
Top 5	88,24	88,93 (+0,78)	88,75 (+0,58)	88,67 (+0,49)
Top 10	88,24	88,86 (+0,70)	88,99 (+0,86)	88,82 (+0,66)

Pada pengujian berikutnya, pengaruh *hyperparameter* dengan ukuran fitur menjadi 2000 fitur pada *Corpus Tweet, Corpus Indonews, Corpus IndoNews + Tweet*. Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* dari klasifikasi *Logistic Regression* dapat dilihat dari Tabel 18. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 1 Similarity* menggunakan *Corpus IndoNews + Tweet* sebesar 83,96%.

Tabel 18 Hasil Peformansi *Feature Expansion* pada *Logistic Regression Hyperparameter*

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	80,73	83,59 (+3,54)	83,93 (+3,96)	**83,96 (+4,00)**
Top 5	80,73	83,64 (+3,60)	83,91 (+3,93)	83,84 (+3,85)
Top 10	80,73	83,62 (+3,57)	83,64 (+3,60)	83,67 (+3,)

Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* serta *hyperparameter* dari klasifikasi *Support Vector Machine* dapat dilihat dari Tabel 19. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 5 Similarity* menggunakan *Corpus Tweet* sebesar 88,73%.

Tabel 19 Hasil Peformansi *Feature Expansion* pada *Support Vector Machine Hyperparameter*

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	83,90	88,25 (+5,18)	88,59 (+5,58)	88,58 (+5,57)
Top 5	83,90	**88,69 (+5,71)**	88,52 (+5,50)	88,61 (+5,61)
Top 10	83,90	88,57 (+5,56)	88,36 (+5,32)	88,69 (+5,70)

Hasil pengujian performansi untuk nilai akurasi menggunakan *feature expansion* serta *hyperparameter* dari klasifikasi *Random Forest* dapat dilihat dari Tabel 20. Hasilnya secara keseluruhan mengalami peningkatan. Nilai akurasi tertinggi didapatkan pada *Top 5 Similarity* menggunakan *Corpus IndoNews* sebesar 89,53%.

Tabel 20 Hasil Peformansi *Feature Expansion* pada *Random Forest Hyperparameter*

Top Similarity	Akurasi (%)			
	Baseline	*Corpus Tweet*	*Corpus IndoNews*	*Corpus IndoNews + Tweet*
Top 1	88,24	89,12 (+1,00)	89,09 (+0,97)	89,29 (+1,19)
Top 5	88,24	89,41 (+1,33)	**89,53 (+1,46)**	88,96 (+0,82)
Top 10	88,24	88,89 (+0,74)	88,76 (+0,59)	89,23 (+1,12)

4.7 Analisis Hasil Pengujian

Hasilnya, percobaan menggunakan *feature expansion* dengan penggunaan kamus kata dan ukuran fitur yang berbeda, mendapatkan hasil yang berbeda-beda pula. Berdasarkan hasil akurasi dan *F1-Score* yang didapatkan, dapat dilihat bahwa terjadi peningkatan pada sistem menambahkan teknik TF-IDF untuk membobotkan kata. Lalu, ketika sistem mengimplementasikan *feature expansion* dan akurasi juga ikut meningkat.

5. Kesimpulan

Pada penelitian ini, telah dibangun deteksi *hoax* menggunakan ekspansi fitur metode *Word2Vec* dengan metode klasifikasi *Support Vector Machine* (SVM), *Random Forest*, *Logistic Regression*. Ekspansi Fitur *Word2Vec* digunakan pada sistem pendeteksi *hoax* ini bertujuan untuk mengurangi ketidakcocokan kosakata pada kalimat *tweet*. Ekspansi fitur dilakukan dengan menggunakan 3 *corpus Word2Vec* (*Tweet, IndoNews*, dan gabungan) dan juga 3 variasi ekspansi fitur (Top 1, Top 5, Top 10) untuk mencari model terbaik. Pada penelitian ini, model ekspansi fitur ini berhasil meningkatkan nilai akurasi untuk semua metode klasifikasi yang digunakan.

Peningkatan nilai akurasi tertinggi terdapat pada model dengan metode *Random Forest* menggunakan *hyperparameter* dengan *Corpus IndoNews* dan menggunakan top 5 sebesar 89,53% dengan kenaikan sebesar 1,46% dari baseline.

REFERENSI

[1] S. Asur dan B. A. Huberman, "Predicting the future with social media," in *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, 2010, vol. 1, hal. 492–499, doi: 10.1109/WI-IAT.2010.63.

[2] APJII, "Laporan Survei Internet APJII 2019 – 2020," *Asos. Penyelenggara Jasa Internet Indones.*, vol. 2020, hal. 1–146, 2020, [Daring]. Tersedia pada: https://apjii.or.id/survei.

[3] Y. Y. Chen, S.-P. Yong, dan A. Ishak, "Email *Hoax* Detection System Using Levenshtein Distance Method.," *J. Comput.*, vol. 9, no. 2, hal. 441–446, 2014.

[4] S. M. Nainar, "Information on the Web.," *Pediatric dentistry*, vol. 22, no. 4. hal. 298, 2000.

[5] A. Fauzi, E. B. Setiawan, dan Z. K. A. Baizal, "*Hoax* News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," in *Journal of Physics: Conference Series*, 2019, vol. 1192, no. 1, doi: 10.1088/1742-6596/1192/1/012025.

[6] E. B. Setiawan, D. H. Widyantoro, dan K. Surendro, "Feature expansion using word embedding for tweet topic classification," in *Proceeding of 2016 10th International Conference on Telecommunication Systems Services and Applications, TSSA 2016: Special Issue in Radar Technology*, 2017, no. October, doi: 10.1109/TSSA.2016.7871085.

[7] W. Wu, B. Zhang, dan M. Ostendorf, "Automatic generation of personalized annotation tags for Twitter users," *NAACL HLT 2010 - Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist. Proc. Main Conf.*, no. June, hal. 689–692, 2010.

[8] E. Rasywir dan A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita *Hoax* Berbahasa Indonesia Berbasis Pembelajaran Mesin," *J. Cybermatika*, vol. 3, no. 2, hal. 1–8, 2015, [Daring]. Tersedia pada: https://www.mendeley.com/import/.

[9] D. Maulina dan R. Sagara, "Klasifikasi Artikel *Hoax* Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency," *J. Mantik Penusa*, vol. 2, no. 1, hal. 35–40, 2018.

[10] F. Komunikasi, U. M. Surakarta, J. A. Yani, dan T. Pos, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," *J. Tek. Elektro*, vol. 9, no. 1, hal. 24–29, 2017, doi: 10.15294/jte.v9i1.10452.

[11] L. Binarwati, "Untuk Optimalisasi Random Forest Dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru : Studi Kasus Pt . Xyz," *Fak. Mat. dan Ilmu Pengetah. Alam Inst. Teknol. Sepuluh Nop. Surabaya*, no. Juli 2017, hal. 1–109, 2017.

[12] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, dan A. Joulin, "Advances in pre-training distributed word representations," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, no. 1, hal. 52–55.

[13] E. B. Setiawan, D. H. Widyantoro, dan K. Surendro, "Feature expansion for sentiment analysis in twitter," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2018-Octob, hal. 509–513, 2018, doi: 10.1109/EECSI.2018.8752851.

[14] T. Pranckevičius dan V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, hal. 221–232, 2017, doi: 10.22364/bjmc.2017.5.2.05.

[15] V. Kalra dan R. Aggarwal, "Importance of Text Data Preprocessing & Implementation in RapidMiner," *Proc. First Int. Conf. Inf. Technol. Knowl. Manag.*, vol. 14, hal. 71–75, 2018, doi: 10.15439/2017km46.

[16] A. Fitri Niasita, P. P. Adikara, dan S. Adinugroho, "Analisis Sentimen Pembangunan Infrastruktur di Indonesia dengan Automated Lexicon *Word2Vec* dan Naive-Bayes," *J-Ptiik*, vol. 3, no. 3, hal. 2673–2679, 2019, [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id.

[17] J. A. Septian, T. M. Fahrudin, dan A. Nugroho, "Journal of Intelligent Systems and Computation 43," hal. 43–49, [Daring]. Tersedia pada: https://t.co/9WloaWpfD5.

[18] B. Susanto, H. Lina, dan A. R. Chrismanto, "Penerapan Social Network Analysis dalam Penentuan Centrality Studi Kasus Social Network Twitter," *J. Inform.*, vol. 8, no. 1, 2012, doi: 10.21460/inf.2012.81.111.

[19] R. Hendayana, "Application Method of Logistic Regression Analyze the Agricultural Technology Adoption," *Inform. Pertan.*, vol. 22, no. 1, hal. 1–9, 2013, [Daring]. Tersedia pada: http://ejurnal.litbang.pertanian.go.id/index.php/IP/article/view/2271/1970.

[20] I. Ahmad, M. Basheri, M. J. Iqbal, dan A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, hal. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[21] F. S. Jumeilah, "Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 1, hal. 19–25, 2017, doi: 10.29207/resti.v1i1.11.

[22] K. Shah, H. Patel, D. Sanghvi, dan M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 5, no. 1, 2020, doi: 10.1007/s41133-020-00032-0.