

ABSTRAK

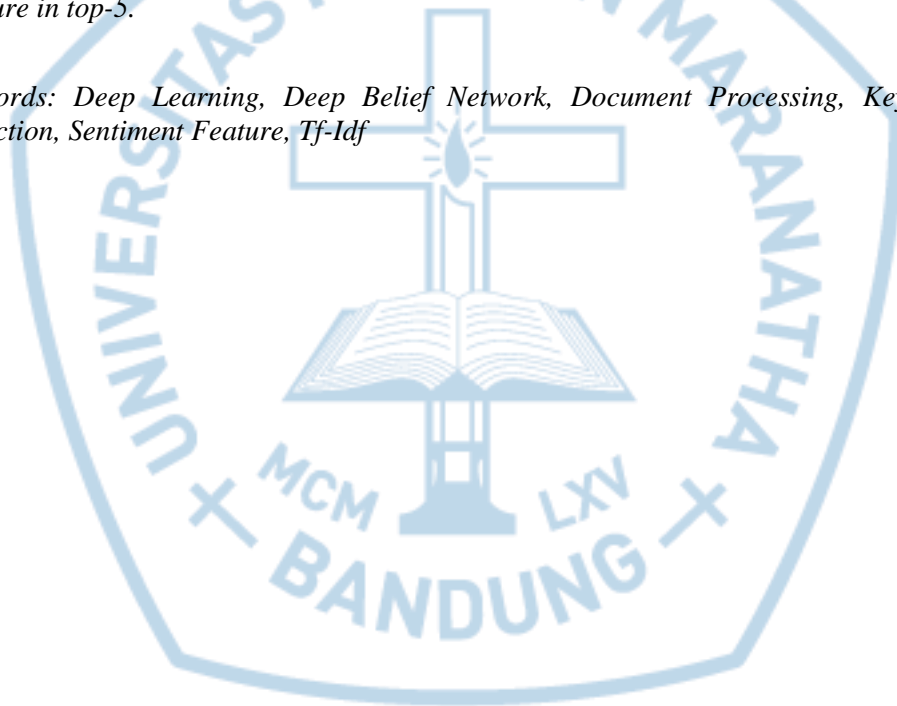
Frasa kunci adalah gabungan kata yang mewakili konsep atau garis besar dari suatu dokumen. Frasa kunci digunakan untuk membantu pembaca dalam mengetahui pokok bahasan dari dokumen. Sayangnya terdapat publikasi ilmiah yang memiliki frasa kunci yang tidak relevan terhadap isi dari dokumen atau tidak memiliki frasa kunci. Berdasarkan permasalahan tersebut maka dalam tugas akhir akan dibuat sistem yang dapat melakukan ekstraksi frasa kunci pada publikasi ilmiah secara otomatis dari pdf. Dalam menentukan frasa kunci pada dokumen, akan diusulkan untuk menggunakan pembobotan tf-idf dan deep belief network sebagai metode pembelajaran dengan nilai sentimen sebagai salah satu fitur pembelajaran. Selain nilai sentimen, akan digunakan posisi section sebagai fitur pembelajaran. Posisi section akan ditentukan dengan menggunakan karakteristik font. Deep belief network diusulkan untuk mengetahui efek dari penggunaan deep learning terhadap ekstraksi frasa kunci. Seluruh pengujian yang dilakukan akan menggunakan dataset milik NUS terkait publikasi ilmiah dengan judul “Keyphrase Extraction in Scientific Publications”. Berdasarkan hasil penelitian didapat hasil bahwa penggunaan deep belief network akan menghasilkan model pembelajaran dengan akurasi yang lebih tinggi dibandingkan dengan menggunakan regresi logistik sebesar 4,33%. Penggunaan analisa sentimen sebagai fitur pembelajaran dapat memberikan peningkatan akurasi terhadap model pembelajaran sebesar 4,17%. Sistem ekstraksi frasa kunci yang dibangun menghasilkan f-measure sebesar 13,22%

Kata kunci: Deep Learning, Deep Belief Network, Ekstraksi Frasa Kunci, Fitur Sentimen, Pemrosesan Dokumen, Tf-Idf

ABSTRACT

Keyphrases are combination of words which represent concept or main idea in document. Keyphrases are used to aid reader's understanding regarding to main topic in document. Unfortunately, there are scientific publications which have keyphrase that doesn't represent content of document or have no keyphrase at all. Based on the problem, in this work will be built an automatic keyphrase extraction system for scientific publication in pdf format. In order to determine keyphrases, proposed to use TF-IDF weighting and deep belief network as learning method with sentiment value as one of the learning feature. Besides sentiment value, will be used section position as learning feature. Section position will be determined using font characteristics. Deep belief network is proposed in order to find out the effect of using deep learning in keyphrase extraction. The entire testing conducted will use dataset belongs to NUS regarding scientific publication titled "Keyphrase Extraction in Scientific Publications. Based on result, using of deep belief network will bring higher accuracy for learning model compared of using logistic regression in 4.33%. The use of sentiment analysis also gives enhancement to the accuracy of learning model in 4.17%. Proposed keyphrase extraction system has 13.22% for f-measure in top-5.

Keywords: Deep Learning, Deep Belief Network, Document Processing, Keyphrase Extraction, Sentiment Feature, Tf-Idf



DAFTAR ISI

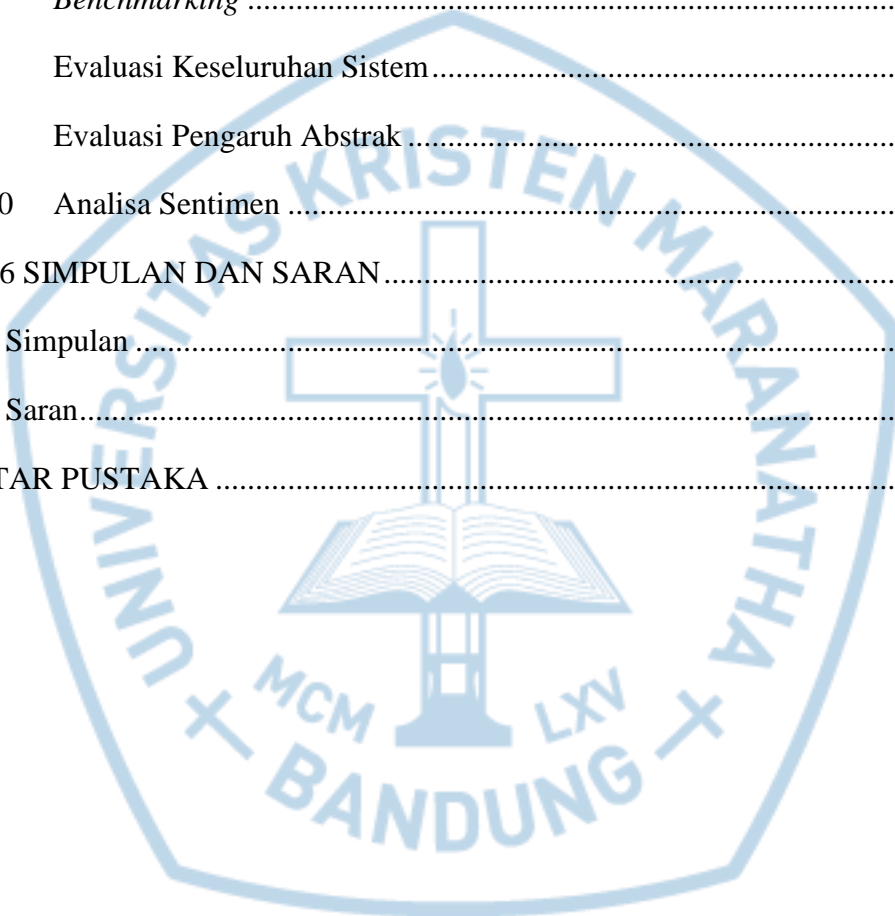
LEMBAR PENGESAHAN	i
PERNYATAAN ORISINALITAS LAPORAN PENELITIAN.....	ii
PERNYATAAN PUBLIKASI LAPORAN PENELITIAN	iii
PRAKATA.....	iv
ABSTRAK	vi
ABSTRACT.....	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR.....	xiii
DAFTAR TABEL.....	xvi
DAFTAR FORMULA	xvii
DAFTAR NOTASI/ LAMBANG.....	xviii
DAFTAR SINGKATAN	xix
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	2
1.3 Tujuan Pembahasan	2
1.4 Ruang Lingkup.....	3
1.5 Sumber Data.....	3
1.6 Sistematika Penyajian	3
BAB 2 KAJIAN TEORI	6
2.1 Temu Balik Informasi	6
2.2 Penguraian Dokumen	7
2.2.1 Tokenisasi	7
2.2.2 <i>Stopping</i>	8

2.2.3 Lemmatization	9
2.2.4 N-Gram	10
2.3 Pembobotan <i>TF-IDF</i>	10
2.4 <i>Part of Speech</i>	11
2.5 Evaluasi Temu Balik Informasi	13
2.6 Pembelajaran Mesin	13
2.7 Jaringan Saraf Tiruan	14
2.8 <i>Backpropagation</i>	16
2.9 Momentum	18
2.10 <i>Deep Neural Network</i>	18
2.11 <i>Restricted Boltzman Machines (RBMs)</i>	19
2.12 <i>Deep Belief Network</i>	23
2.13 <i>Pretraining</i> DBN	24
2.14 <i>Fine Tuning</i> DBN	26
2.15 Fitur-Fitur Pembelajaran Umum Ekstraksi <i>Keyphrase</i>	26
2.16 <i>K-Fold Cross Validation</i>	27
2.17 Accord .NET	28
2.18 Stanford NLP	29
2.19 ITextSharp	29
2.20 Kontribusi Penelitian	29
BAB 3 ANALISIS DAN RANCANGAN SISTEM	34
3.1 Rancangan Metode	34
3.1.1 Persiapan Data	34
3.1.1.1 PoS Tagging	35
3.1.1.2 Tokenisasi	36
3.1.1.3 N-gram	36

3.1.1.4	Identifikasi <i>Noun Phrase</i>	36
3.1.1.5	<i>Stopword</i>	37
3.1.1.6	<i>Lemmanization</i>	37
3.1.2	Pembentukan Model Pembelajaran.....	38
3.1.3	Ekstraksi <i>Keyphrase</i>	40
3.2	Pemodelan Sistem	40
3.2.1	Pemodelan Perangkat Lunak.....	41
3.2.1.1	Use Case Diagram.....	41
3.2.1.1.1	Rancangan Use Case Diagram.....	41
3.2.1.1.2	Deskripsi Use Case Diagram	41
3.2.1.2	Class Diagram	42
3.2.1.3	Activity Diagram.....	44
3.2.1.3.1	Activity Diagram Pembentukan Model	44
3.2.1.3.2	Activity Diagram Ekstraksi <i>Keyphrase</i>	45
3.2.2	Rancangan Antarmuka Pengguna	45
3.2.2.1	Jendela Pembentukan Model	46
3.2.2.2	Jendela Ekstraksi <i>Keyphrase</i>	47
BAB 4	IMPLEMENTASI.....	49
4.1	Implementasi <i>Class</i>	49
4.1.1	<i>Class Phrase</i>	49
4.1.2	<i>Class PhraseHelper</i>	49
4.1.3	<i>Class DeepNeuralNet</i>	50
4.1.4	<i>Class LearningHelper</i>	50
4.1.5	<i>Class PreprocessingMethod</i>	51
4.1.6	<i>Class PreprocessingStep</i>	51
4.1.7	<i>Class StanfordNlpPipe</i>	52

4.2	Implementasi Antarmuka	53
4.1.1	Implementasi Antarmuka Modul Pembentukan Model	53
4.2.1	Implementasi Antarmuka Modul Ekstraksi Keyphrase	53
4.3	Implementasi Algoritma.....	54
4.3.1	<i>Deterministic Finite Automata (DFA)</i>	54
4.3.2	<i>Sentiment Analysis</i>	56
4.3.3	<i>Name Entity Recognition</i>	58
4.3.4	<i>Stopping</i>	58
4.3.5	<i>Lemmanization</i>	59
4.3.6	Pengenalan <i>Section</i>	60
4.3.7	<i>Pretraining</i>	61
4.3.8	<i>Backpropagation</i>	63
4.4	Implementasi Metode.....	64
4.4.1	Persiapan Data.....	64
4.4.1.1	Konversi Pdf	65
4.4.1.2	PoS Tagging.....	67
4.4.1.3	Sentiment Analysis	68
4.4.1.4	Pengambilan Kandidat Frasa Kunci	69
4.4.1.5	Penggabungan Frasa	72
4.4.1.6	Perhitungan Idf	76
4.4.1.7	Ranking Tf-Idf	77
4.4.2	Pembentukan Model.....	80
4.4.3	Ekstraksi Keyphrase.....	83
BAB 5 PENGUJIAN		84
5.1	Rencana Pengujian	84
5.2	Pengujian Black Box.....	84

5.1.1 Pengujian Pembentukan Model Pembelajaran.....	84
5.1.2 Pengujian Ekstraksi Keyphrase.....	87
5.3 Data Pengujian	88
5.4 Pengujian Hasil konversi Pdf ke Teks	89
5.5 Pengujian Fitur Pembelajaran	91
5.6 Pengujian Jumlah Layer dan Jumlah Neuron.....	93
5.7 <i>Benchmarking</i>	98
5.8 Evaluasi Keseluruhan Sistem.....	99
5.9 Evaluasi Pengaruh Abstrak.....	101
5.10 Analisa Sentimen	102
BAB 6 SIMPULAN DAN SARAN.....	104
6.1 Simpulan	104
6.2 Saran.....	104
DAFTAR PUSTAKA	105



DAFTAR GAMBAR

Gambar 2.1 Contoh Tokenisasi [5]	8
Gambar 2.2 Contoh <i>Stopword</i> Bahasa Inggris	9
Gambar 2.3 Contoh <i>Lemmatization</i> Pada Dokumen [5]	9
Gambar 2.4 Contoh <i>n-gram</i> pada dokumen	10
Gambar 2.5 Contoh hasil dari POS tagger [3]	12
Gambar 2.6 Contoh <i>Multilayer Perceptron</i> (MLP) [12].....	14
Gambar 2.7 Perceptron, Fungsi Penjumlahan, dan Fungsi Aktivasi [10].....	15
Gambar 2.8 Fungsi sigmoid [12]	16
Gambar 2.9 Perbedaan ANN dan DNN [12]	19
Gambar 2.10 Struktur <i>Restricted Boltzman Machine</i> [12].....	19
Gambar 2.11 Fase Maju RBM [18]	20
Gambar 2.12 Fase Mundur Pada RBM [18]	20
Gambar 2.13 <i>Gibbs Sampling</i> Pada RBM [12].....	22
Gambar 2.14 RBMs Pada DBN [16].....	24
Gambar 2.15 Algoritma <i>Contrastive Divergence</i> [12].....	25
Gambar 2.16 <i>Learning</i> DBN Pada Jaringan Dengan 3 <i>Hidden Layer</i> [22]	26
Gambar 2.17 Contoh <i>5-Fold Cross Validation</i> [11]	28
Gambar 3.1 Langkah-Langkah Persiapan Data	35
Gambar 3.2 Langkah-Langkah Pembentukan Model	39
Gambar 3.3 Langkah-Langkah Ekstraksi Keyphrase.....	40
Gambar 3.4 Rancangan Use Case Diagram	41
Gambar 3.5 Rancangan Class Diagram Sistem Ekstraksi Keyphrase.....	43
Gambar 3.6 Activity Diagram Pembentukan Model.....	44
Gambar 3.7 Activity Diagram Ekstraksi Keyphrase.....	45
Gambar 3.8 Rancangan Jendela Pembentukan Model.....	46
Gambar 3.9 Rancangan Jendela Ekstraksi <i>Keyphrase</i>	48
Gambar 4.1 <i>Class</i> Phrase	49
Gambar 4.2 <i>Class</i> PhraseHelper	50
Gambar 4.3 <i>Class</i> DeepNeuralNet.....	50
Gambar 4.4 <i>Class</i> LearningHelper.....	51

Gambar 4.5 <i>Class</i> PreprocessingMethod.....	51
Gambar 4.6 <i>Class</i> PreprocessingStep	52
Gambar 4.7 <i>Class</i> StanfordNlpPipe	52
Gambar 4.8 Tampilan Antarmuka Modul Pembentukan Model.....	53
Gambar 4.9 Tampilan Antarmuka Ekstraksi Keyphrase.....	54
Gambar 4.10 Method IsNounPhrase Implementasi DFA	56
Gambar 4.11 Method Sentiment Analysis	57
Gambar 4.12 Method GetNameEntity	58
Gambar 4.13 Method ContainsStopword.....	59
Gambar 4.14 Method Lemmatize	60
Gambar 4.15 Implementasi Pengenalan Section.....	61
Gambar 4.16 Method UnsupervisedPretraining.....	62
Gambar 4.17 Method SupervisedPretraining	63
Gambar 4.18 Method Backpropagation	63
Gambar 4.19 Implementasi Persiapan Data.....	65
Gambar 4.20 Contoh Hasil Konversi Pdf ke Teks.....	66
Gambar 4.21 Contoh Hasil Pemisahan Kata Kunci.....	67
Gambar 4.22 Kode untuk Mengenali Kalimat dan <i>PoS Tagging</i>	67
Gambar 4.23 Contoh Hasil <i>PoS Tagging</i>	68
Gambar 4.24 Contoh Hasil Analisa Sentimen	69
Gambar 4.25 Potongan Kode Step 1	72
Gambar 4.26 Potongan Kode Step 2	73
Gambar 4.27 Potongan Kode Perhitungan Idf.....	76
Gambar 4.28 Contoh Hasil Perhitungan Idf.....	77
Gambar 4.29 Ranking tf-Idf.....	78
Gambar 5.1 Perbandingan Hasil Konversi Pdf oleh Sistem Dengan Nguyen	90
Gambar 5.2 Hasil Konversi Pdf ke Teks Pada Dokumen 51 Oleh Sistem	91
Gambar 5.3 Hasil Konversi Pdf ke Teks Pada Dokumen 51 Oleh Nguyen.....	91
Gambar 5.4 Dampak Fitur Pada Model Pembelajaran.....	92
Gambar 5.5 Dampak Akurasi Penambahan <i>Layer</i> Pada Model Tanpa <i>Pretraining</i>	94
Gambar 5.6 Dampak Akurasi Jumlah <i>Neuron</i> Pada Model Tanpa <i>Pretraining</i> ...	95

Gambar 5.7 Dampak Akurasi Penambahan *Layer* Pada Model Dengan *Pretraining*
..... 96

Gambar 5.8 Dampak Penambahan *Neuron* Pada Model Dengan *Pretraining*..... 96

Gambar 5.9 Perbandingan *Layer* Model Tanpa *Pretraining* dan Dengan *Pretraining*
..... 97

Gambar 5.10 Perbandingan Model Tanpa *Pretraining* dan Dengan *Pretraining*. 98

Gambar 5.11 Perbandingan *Deep Belief Network* dan *Logistic Regeresion*..... 99









DAFTAR TABEL

Tabel 2.1 Kelas dalam Penn Treebank [9]	12
Tabel 2.2 Fitur-Fitur Pembelajaran Ekstraksi <i>Keyphrase</i> [23]	26
Tabel 2.3 Kumpulan Library Accord.NET [23].....	28
Tabel 3.1 Rancangan Fitur-Fitur Pembelajaran	38
Tabel 3.2 Deskripsi Use Case Diagram untuk Pembentukan Model	41
Tabel 3.3 Deskripsi Use Case Diagram untuk Ekstraksi <i>Keyphrase</i>	42
Tabel 4.1 Contoh Hasil Step 1	70
Tabel 4.2 Contoh Hasil Step 2	75
Tabel 4.3 Contoh Hasil Step 3	79
Tabel 4.4 Contoh Hasil Pemilihan Frasa Kunci Relevan dan Tidak Relevan	81
Tabel 5.1 Test Case untuk Pembentukan Model Pembelajaran.....	84
Tabel 5.2 Test Case untuk Ekstraksi <i>Keyphrase</i>	87
Tabel 5.3 Rata-Rata <i>Precision</i> , <i>Recall</i> , dan <i>F-Measure</i> Skenario 1	100
Tabel 5.4 Rata-Rata <i>Precision</i> , <i>Recall</i> , dan <i>F-Measure</i> Skenario 2	100
Tabel 5.5 Rata-Rata <i>Precision</i> , <i>Recall</i> , dan <i>F-Measure</i> Skenario 3	100
Tabel 5.6 Rata-Rata <i>Precision</i> , <i>Recall</i> , dan <i>F-Measure</i> Skenario 4	101
Tabel 5.7 Rata-Rata <i>Precision</i> , <i>Recall</i> , dan <i>F-Measure</i> Dokumen Dengan Abstrak	102
Tabel 5.8 Rata-Rata <i>Precision</i> , <i>Recall</i> , dan <i>F-Measure</i> Dokumen Tanpa Abstrak	102
Tabel 5.9 Korelasi Kalimat dengan Nilai Sentimen.....	103

DAFTAR FORMULA

Formula 2.1 Persamaan TF-IDF [5].....	11
Formula 2.2 Persamaan <i>Precision</i> [5].....	13
Formula 2.3 Persamaan <i>Recall</i> [5].....	13
Formula 2.4 <i>F-Measure</i> dengan <i>Harmonic Mean</i> [5].....	13
Formula 2.5 Fungsi Penjumlahan Jaringan Saraf Tiruan [10].....	15
Formula 2.6 Fungsi Sigmoid Jaringan Saraf Tiruan [10].....	15
Formula 2.7 <i>Error Output Layer</i> Jaringan Saraf Tiruan [10].....	16
Formula 2.8 <i>Error Hidden Layer</i> Jaringan Saraf Tiruan [10].....	17
Formula 2.9 <i>Delta Rule</i> Jaringan Saraf Tiruan [10].....	17
Formula 2.10 Perubahan Bobot <i>Perceptron</i> Jaringan Saraf Tiruan [10].....	17
Formula 2.11 <i>Delta Rule</i> Jaringan Saraf Tiruan Dengan Momentum [10].....	18
Formula 2.12 Energy-Based Model RBM [16].....	21
Formula 2.13 Peluang Untuk <i>Hidden Units</i> [12].....	21
Formula 2.14 Peluang Untuk <i>Visible Units</i> [12].....	21
Formula 2.15 Perubahan Bobot dan Bias <i>Visible</i> dan <i>Hidden Unit</i> [12].....	22
Formula 3.1 Regular Expresion DFA [33].....	37

DAFTAR NOTASI/ LAMBANG

Jenis	Notasi/ Lambang	Nama	Arti
Use Case		Aktor	Menggambarkan aktor atau pengguna aplikasi.
Use Case		Case	Menggambarkan proses atau aksi yang dapat dilakukan oleh aktor pada aplikasi.
Use Case		Association	Menggambarkan komunikasi antara <i>use case</i> dan aktor yang berpartisipasi (asosiasi).
Activity Diagram		Initial Node	Menandakan dimulainya aktivitas pada sebuah sistem.
Activity Diagram		Activity	Menandakan aktivitas apa yang akan dilakukan oleh pengguna aplikasi.
Activity Diagram		Final Node	Menandakan akhir aliran proses sistem

DAFTAR SINGKATAN

ANN	Artificial Neural Network
DNN	Deep Neural Network
MLP	Multilayer Perceptron
PoS	Part of Speech
RBM	Restricted Boltzman Machine
TF-IDF	Term Frequency–Inverse Document Frequency
UML	Unified Modelling Language

