

ABSTRAK

Alfons Richardo:

Skripsi

Chatbot dengan Algoritma *Bidirectional Long Short Term Memory* untuk website

Admisi dan Humas Universitas Kristen Petra

Universitas Kristen Petra (UK Petra) memiliki layanan informasi yang bertujuan untuk membantu masyarakat, terutama mahasiswa baru dan orang tua mahasiswa untuk memperoleh informasi seputar UK Petra. UK Petra sudah memiliki layanan *chatbot* sederhana yang mampu menjawab beberapa pertanyaan umum yang sering ditanyakan masyarakat. Namun apabila pertanyaan yang diajukan terlalu rumit, maka masyarakat akan dihibau untuk langsung berbincang dengan petugas. Hal ini tentu kurang efektif, mengingat jam kerja petugas yang terbatas, belum lagi apabila petugas sedang sibuk melayani orang lain. Penelitian oleh ko Kevin dengan penggunaan metode k-NN dan HMM menunjukkan kurang mampuan *chatbot* dalam menerima pertanyaan yang kompleks (lebih dari 8 kata).

Oleh karena itu, skripsi ini akan menggunakan metode *k-Means* untuk melakukan pengelompokan data, lalu model *Bidirectional Long Short Term Memory* akan digunakan untuk pemberian jawaban akhir. Akan dilakukan juga penelitian dengan menggunakan model *Convolutional Neural Network* sebagai pembandingan. *Word Embedding* yang juga akan dilakukan antara lain adalah *Tokenizer*, *Word2Vec*, *Bag of Words*, dan *GloVe*.

Hasil akurasi rata-rata tertinggi *chatbot* sebesar 55.56%, dengan menggunakan *Word2Vec CBOW Window=5 Min_Count=1 Bidirectional Long Short Term Memory* dengan *Splitting* Data dan *Stopwords* Dihilangkan.

Kata kunci: *chatbot*, *artificial intelligence*, *k-means*, *bidirectional long short term memory*.

ABSTRACT

Alfons Richardo:

Undergraduate Thesis

Chatbot with *Bidirectional Long Short Term Memory Algorithm* for Petra Christian University Admission and Public Relations website

Petra Christian University has an information service that aims to help the community, especially new students and parents, to obtain information about Petra Christian University. PCU already has a simple chatbot service that is able to answer some common questions that people often ask. However, if the questions asked are too complicated, then the students and parents will be encouraged to directly talk to a person. This is of course less effective, considering the worker's limited working hours, not to mention when the workers are busy serving other people. Research by Kevin using the k-NN and HMM methods shows the inability of chatbots to accept complex questions (more than 8 words).

Therefore, this thesis will use the k-Means method to group data, then the Bidirectional Long Short Term Memory model will be used to provide the final answer. Research will also be carried out using the Convolutional Neural Network model as a comparison. Word Embedding that will also be carried out includes Tokenizer, Word2Vec, Bag of Words, and GloVe.

The results of the highest average accuracy of the chatbot is 55.56%, using Word2Vec CBOW Window=5 Min_Count=1 Bidirectional Long Short Term Memory with Data Splitting and Stopwords Removed.

Keywords: *chatbot, artificial intelligence, k-means, bidirectional long short term memory.*

DAFTAR ISI

HALAMAN JUDUL.....	i
LEMBAR PENGESAHAN	ii
LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI.....	iii
KATA PENGANTAR	iv
ABSTRAK	vi
DAFTAR ISI	viii
DAFTAR GAMBAR.....	xiii
DAFTAR TABEL	xiv
DAFTAR SEGMENT PROGRAM	xvi
DAFTAR RUMUS.....	xix
DAFTAR LAMPIRAN	xx
1. PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan Penelitian.....	2
1.4 Ruang Lingkup	2
1.5 Metodologi Penelitian	3
1.6 Sistematika Penulisan.....	5
2. LANDASAN TEORI.....	7
2.1 Tinjauan Pustaka	7
2.1.1 Admisi dan Humas Universitas Kristen Petra	7
2.1.2 <i>Artificial Intelligence</i>	7
2.1.3 <i>Chatbot</i>	7
2.1.4 <i>Natural Language Processing</i>	8
2.1.5 <i>Text Preprocessing</i>	9
2.1.5.1 <i>Tokenization</i>	9
2.1.5.2 <i>Lower Casing</i>	9

2.1.5.3	<i>Stop Words Removal</i>	9
2.1.5.4	<i>Stemming</i>	10
2.1.5.5	<i>Lemmatization</i>	10
2.1.6	<i>Feature Extraction</i>	10
2.1.6.1	<i>Tokenizer</i>	10
2.1.6.2	<i>Bag of Words</i>	11
2.1.6.3	<i>Term Frequency – Inverse Document Frequency</i>	11
2.1.6.4	<i>Word2Vec</i>	12
2.1.7	<i>Machine Learning</i>	14
2.1.7.1	<i>Unsupervised Learning</i>	14
2.1.7.2	<i>Supervised Learning</i>	15
2.1.7.3	<i>Reinforcement Learning</i>	15
2.1.8	<i>Clustering</i>	15
2.1.8.1	<i>Centroid-based Clustering</i>	16
2.1.8.2	<i>Density-based Clustering</i>	16
2.1.8.3	<i>Distribution-based Clustering</i>	17
2.1.8.4	<i>Hierarchical Clustering</i>	17
2.1.9	<i>k-Means</i>	18
2.1.10	<i>Deep Learning</i>	19
2.1.11	<i>Bidirectional Long Short Term Memory</i>	20
2.2	<i>Tinjauan Studi</i>	24
2.2.1	<i>Chatbot untuk Website UK Petra dengan Hidden Markov Model dan k-Nearest Neighbor (Kevin, 2021)</i>	24
2.2.2	<i>A New Chatbot for Customer Service on Social Media (Xu et al, 2018)</i>	25
2.2.3	<i>Automated Thai-FAQ Chatbot using RNN-LSTM (Muangkammuen et al, 2018)</i> .	25
2.2.4	<i>LSTM and Simple RNN Comparison in the Problem of Sequence to Sequence on Conversation Data Using Bahasa Indonesia (Prabowo et al, 2018)</i>	26
2.2.5	<i>Implementasi Natural Language Processing pada Sistem Chatbot Informasi Saham dengan Algoritma Long Short-Term Memory (LSTM) dan Fuzzy String Matching (Azni, 2022)</i>	27
3.	ANALISIS DAN DESAIN SISTEM	29
3.1	<i>Dataset</i>	29
3.2	<i>Pengolahan Awal Data</i>	29
3.3	<i>Analisis Penelitian Sebelumnya</i>	30
3.4	<i>Desain Sistem</i>	31

3.4.1	Alur Sistem	31
3.4.2	<i>Text Preprocessing</i>	32
3.4.3	<i>k-Means</i>	35
3.4.4	<i>Word Embedding</i>	35
3.4.5	<i>Bidirectional Long Short Term Memory</i>	37
3.5	Pengujian	38
3.5.1	Pertanyaan	38
3.5.2	<i>Word Embedding</i>	39
3.5.3	<i>Deep Learning</i>	39
3.5.4	Pengaturan Lain	39
4.	IMPLEMENTASI SISTEM	40
4.1	<i>Load Data</i>	40
4.2	<i>Text Preprocessing</i>	41
4.3	<i>Feature Extraction</i>	42
4.4	<i>Train Test Split</i>	57
4.5	<i>Bidirectional Long Short Term Memory</i>	57
4.6	<i>Convolutional Neural Network</i>	58
4.7	Testing Input User	59
4.8	Perhitungan Akurasi	66
5.	PENGUJIAN	67
5.1	Pengujian <i>Chatbot</i>	67
5.2	Hasil Percobaan	67
5.2.1	Pengujian Akurasi Semua Model dengan Pertanyaan Lama	67
5.2.1.1	Penjelasan <i>Type</i>	67
5.2.1.2	Pengujian dengan <i>Word Embedding Tokenizer</i> dan Model <i>Bidirectional Long Short Term Memory</i>	70
5.2.1.3	Pengujian dengan <i>Word Embedding GloVe</i> dan Model <i>Bidirectional Long Short Term Memory</i>	70
5.2.1.4	Pengujian dengan <i>Word Embedding Word2Vec</i> dan Model <i>Bidirectional Long Short Term Memory</i>	71
5.2.1.5	Pengujian dengan <i>Word Embedding Bag of Words</i> dan Model <i>Bidirectional Long Short Term Memory</i>	72
5.2.1.6	Pengujian dengan <i>Word Embedding Tokenizer</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	73

5.2.1.7	Pengujian dengan <i>Word Embedding Word2Vec</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	74
5.2.1.8	Pengujian dengan <i>Word Embedding Tokenizer</i> dengan Model <i>Convolutional Neural Network</i>	76
5.2.1.9	Pengujian dengan <i>Word Embedding Word2Vec SkipGram (Window=5 & Min_count=2)</i> dengan Model <i>Convolutional Neural Network</i>	76
5.2.1.10	Pengujian dengan <i>Word Embedding GloVe</i> dengan Model <i>1 Layer Convolutional Neural Network</i>	77
5.2.2	Perbandingan Akurasi dengan dan tanpa Penggunaan <i>k-Means</i> Menggunakan Pertanyaan Lama	78
5.2.3	Perbandingan dengan Pengujian Sebelumnya	80
5.2.4	Pengujian Akurasi Semua Model dengan Pertanyaan Baru	81
5.2.4.1	Pengujian dengan <i>Word Embedding Tokenizer</i> dan Model <i>Bidirectional Long Short Term Memory</i>	82
5.2.4.2	Pengujian dengan <i>Word Embedding GloVe</i> dan Model <i>Bidirectional Long Short Term Memory</i>	82
5.2.4.3	Pengujian dengan <i>Word Embedding Word2Vec</i> dan Model <i>Bidirectional Long Short Term Memory</i>	83
5.2.4.4	Pengujian dengan <i>Word Embedding Bag of Words</i> dan Model <i>Bidirectional Long Short Term Memory</i>	84
5.2.4.5	Pengujian dengan <i>Word Embedding Tokenizer</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	85
5.2.4.6	Pengujian dengan <i>Word Embedding Word2Vec</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	85
5.2.4.7	Pengujian dengan <i>Word Embedding Tokenizer</i> dengan Model <i>Convolutional Neural Network</i>	86
5.2.4.8	Pengujian dengan <i>Word Embedding Word2Vec SkipGram (Window=5 & Min_count=2)</i> dengan Model <i>Convolutional Neural Network</i>	87
5.2.4.9	Pengujian dengan <i>Word Embedding GloVe Full</i> dengan Model <i>1 Layer Convolutional Neural Network</i>	88
5.2.4.10	Kesimpulan Percobaan 5.2.4	88
5.2.5	Perbandingan Akurasi dengan dan tanpa Penggunaan <i>k-Means</i> Menggunakan Pertanyaan Baru	88
5.3	Diskusi	90
6.	KESIMPULAN DAN SARAN	92
6.1	Kesimpulan	92
6.2	Saran	92
	DAFTAR PUSTAKA	94

LAMPIRAN.....	97
---------------	----

DAFTAR GAMBAR

Gambar 2.1 Arsitektur <i>Continuous Bag of Words</i>	13
Gambar 2.2 Arsitektur <i>Skip Gram</i>	13
Gambar 2.3 <i>Centroid-based Clustering</i>	16
Gambar 2.4 <i>Density-based Clustering</i>	16
Gambar 2.5 <i>Distribution-based Clustering</i>	17
Gambar 2.6 <i>Hierarchical Clustering</i>	17
Gambar 2.7 <i>Convolutional Neural Network</i>	19
Gambar 2.8 <i>Recurrent Neural Network</i>	19
Gambar 2.9 <i>Long Short Term Memory</i>	21
Gambar 2.10 Tahapan <i>LSTM</i>	21
Gambar 2.11 <i>Bidirectional Long Short Term Memory</i>	24
Gambar 3.1 Gambar <i>dataset</i>	29
Gambar 3.2 Alur Sistem	31
Gambar 3.3 Alur <i>text preprocessing</i>	33
Gambar 3.4 Proses <i>tokenization</i>	33
Gambar 3.5 Proses <i>lower casing</i>	34
Gambar 3.6 Proses <i>stopwords removal</i>	34
Gambar 3.7 Proses <i>stemming</i>	35
Gambar 3.8 Alur <i>k-Means</i>	35
Gambar 3.9 Alur <i>Word Embedding</i>	36
Gambar 3.10 Alur <i>Bidirectional Long Short Term Memory</i>	37
Gambar 5.1 <i>chatbot</i> menerima <i>input</i> pertanyaan <i>user</i>	67
Gambar 5.2 <i>chatbot</i> memberikan <i>response</i> dari pertanyaan <i>user</i>	67

DAFTAR TABEL

Tabel 5.1 Penjelasan <i>Type Word Embedding</i>	68
Tabel 5.2 Kesimpulan Akurasi Percobaan <i>Word Embedding Tokenizer</i> dan Model <i>Bidirectional Long Short Term Memory</i>	70
Tabel 5.3 Kesimpulan Akurasi Percobaan <i>Word Embedding GloVe</i> dan Model <i>Bidirectional Long Short Term Memory</i>	70
Tabel 5.4 Kesimpulan Akurasi Percobaan <i>Word Embedding Word2Vec</i> dan Model <i>Bidirectional Long Short Term Memory</i>	71
Tabel 5.5 Kesimpulan Akurasi Percobaan <i>Word Embedding Bag of Words</i> dan Model <i>Bidirectional Long Short Term Memory</i>	72
Tabel 5.6 Kesimpulan Akurasi Percobaan <i>Word Embedding Tokenizer</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	73
Tabel 5.7 Kesimpulan Akurasi Percobaan <i>Word Embedding Word2Vec</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	74
Tabel 5.8 Kesimpulan Akurasi Percobaan <i>Word Embedding Tokenizer</i> dengan Model <i>Convolutional Neural Network</i>	76
Tabel 5.9 Kesimpulan Akurasi Percobaan <i>Word Embedding Word2Vec SkipGram (Window=5 & Min_count=2)</i> dengan Model <i>Convolutional Neural Network</i>	77
Tabel 5.10 Kesimpulan Akurasi Percobaan <i>Word Embedding GloVe Full</i> dengan Model <i>1 Layer Convolutional Neural Network</i>	78
Tabel 5.11 Kesimpulan akurasi semua model dengan menggunakan <i>k-Means clustering</i>	78
Tabel 5.12 Kesimpulan akurasi semua model tanpa menggunakan <i>k-Means clustering</i>	80
Tabel 5.13 Kesimpulan akurasi dibanding dengan penelitian sebelumnya (Kevin, 2021)	81
Tabel 5.14 Kesimpulan Akurasi Percobaan <i>Word Embedding Tokenizer</i> dan Model <i>Bidirectional Long Short Term Memory</i>	82
Tabel 5.15 Kesimpulan Akurasi Percobaan <i>Word Embedding GloVe</i> dan Model <i>Bidirectional Long Short Term Memory</i>	82

Tabel 5.16 Kesimpulan Akurasi Percobaan <i>Word Embedding Word2Vec</i> dan Model <i>Bidirectional Long Short Term Memory</i>	83
Tabel 5.17 Kesimpulan Akurasi Percobaan <i>Word Embedding Bag of Words</i> dan Model <i>Bidirectional Long Short Term Memory</i>	84
Tabel 5.18 Kesimpulan Akurasi Percobaan <i>Word Embedding Tokenizer</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	85
Tabel 5.19 Kesimpulan Akurasi Percobaan <i>Word Embedding Word2Vec</i> dan <i>k-Means</i> dengan Model <i>Bidirectional Long Short Term Memory</i>	85
Tabel 5.20 Kesimpulan Akurasi Percobaan <i>Word Embedding Tokenizer</i> dengan Model <i>Convolutional Neural Network</i>	86
Tabel 5.21 Kesimpulan Akurasi Percobaan <i>Word Embedding Word2Vec SkipGram (Window=5 & Min_count=2)</i> dengan Model <i>Convolutional Neural Network</i>	87
Tabel 5.22 Kesimpulan Akurasi Percobaan <i>Word Embedding GloVe Full</i> dengan Model <i>1 Layer Convolutional Neural Network</i>	88
Tabel 5.23 Kesimpulan akurasi semua model dengan menggunakan <i>k-Means clustering</i>	89
Tabel 5.24 Kesimpulan akurasi semua model tanpa menggunakan <i>k-Means clustering</i>	89

DAFTAR SEGMENT PROGRAM

Segmen Program 4.1 <i>Load Data</i>	40
Segmen Program 4.2 <i>Text Preprocessing</i>	41
Segmen Program 4.3 <i>Word Embedding Tokenizer</i>	42
Segmen Program 4.4 <i>Word Embedding Tokenizer dan k-Means</i>	42
Segmen Program 4.5 <i>Bag of Words</i>	43
Segmen Program 4.6 <i>Bag of Words Unigram</i>	44
Segmen Program 4.7 <i>Bag of Words Bigram</i>	44
Segmen Program 4.8 <i>Bag of Words Trigram</i>	44
Segmen Program 4.9 <i>Bag of Words Unigram dan min_df = 2</i>	45
Segmen Program 4.10 <i>Word2Vec CBOW window=5 dan min_count=1</i>	45
Segmen Program 4.11 <i>Word2Vec CBOW window=5 dan min_count=2</i>	46
Segmen Program 4.12 <i>Word2Vec CBOW window=10 dan min_count=1</i>	46
Segmen Program 4.13 <i>Word2Vec CBOW window=10 dan min_count=2</i>	46
Segmen Program 4.14 <i>Word2Vec SkipGram window=5 dan min_count=1</i>	47
Segmen Program 4.15 <i>Word2Vec SkipGram window=5 dan min_count=2</i>	48
Segmen Program 4.16 <i>Word2Vec SkipGram window=10 dan min_count=1</i>	48
Segmen Program 4.17 <i>Word2Vec SkipGram window=10 dan min_count=2</i>	48
Segmen Program 4.18 <i>Word2Vec dengan k-Means</i>	48
Segmen Program 4.19 Pembuatan model <i>Word2Vec SkipGram Window=5 dan Min_count=1</i> dalam setiap <i>cluster</i>	49
Segmen Program 4.20 Pembuatan model <i>Word2Vec SkipGram Window=5 dan Min_count=2</i> dalam setiap <i>cluster</i>	51
Segmen Program 4.21 Pembuatan model <i>Word2Vec SkipGram Window=10 dan Min_count=1</i> dalam setiap <i>cluster</i>	51

Segmen Program 4.22 Pembuatan model <i>Word2Vec SkipGram Window=10</i> dan <i>Min_count=2</i> dalam setiap <i>cluster</i>	51
Segmen Program 4.23 Pembuatan model <i>Word2Vec CBOW Window=5</i> dan <i>Min_count=1</i> dalam setiap <i>cluster</i>	51
Segmen Program 4.24 Pembuatan model <i>Word2Vec CBOW Window=5</i> dan <i>Min_count=2</i> dalam setiap <i>cluster</i>	52
Segmen Program 4.25 Pembuatan model <i>Word2Vec CBOW Window=10</i> dan <i>Min_count=1</i> dalam setiap <i>cluster</i>	52
Segmen Program 4.26 Pembuatan model <i>Word2Vec CBOW Window=10</i> dan <i>Min_count=2</i> dalam setiap <i>cluster</i>	53
Segmen Program 4.27 <i>TFIDF</i>	54
Segmen Program 4.28 <i>Function</i> untuk mengambil nilai vektor dari semua kata	54
Segmen Program 4.29 <i>GloVe Mean</i>	55
Segmen Program 4.30 <i>GloVe Min</i>	56
Segmen Program 4.31 <i>GloVe Max</i>	56
Segmen Program 4.32 <i>GloVe Full</i>	56
Segmen Program 4.33 <i>Train Test Split</i>	57
Segmen Program 4.34 Model <i>Bidirectional Long Short Term Memory</i>	57
Segmen Program 4.35 Model <i>Convolutional Neural Network 3 Layer</i>	58
Segmen Program 4.36 Model <i>Convolutional Neural Network 2 Layer</i>	58
Segmen Program 4.37 Model <i>Convolutional Neural Network 1 Layer</i>	59
Segmen Program 4.38 <i>Testing Input User Word Embedding Word2Vec</i>	59
Segmen Program 4.39 <i>Testing Input User Word Embedding Tokenizer</i> menggunakan <i>k-Means</i>	61
Segmen Program 4.40 <i>Testing Input User Word Embedding Tokenizer</i>	61
Segmen Program 4.41 <i>Testing Input User Word Embedding GloVe Min</i>	62
Segmen Program 4.42 <i>Testing Input User Word Embedding GloVe Mean</i>	63
Segmen Program 4.43 <i>Testing Input User Word Embedding GloVe Max</i>	63

Segmen Program 4.44 <i>Testing Input User Word Embedding Word2Vec</i> menggunakan <i>k-Means</i>	64
Segmen Program 4.45 Perhitungan Akurasi Model <i>Chatbot</i>	66

DAFTAR RUMUS

(2.1) Rumus <i>Term Frequency</i>	12
(2.2) Rumus <i>Inverse Document Frequency</i>	12
(2.3) Rumus <i>TF-IDF</i>	12
(2.4) Rumus <i>Euclidean Distance</i>	18
(2.5) Rumus <i>Forget Gate LSTM</i>	22
(2.6) Rumus <i>Sigmoid LSTM</i>	22
(2.7) Rumus <i>Tanh LSTM</i>	22
(2.8) Rumus <i>Update Cell State LSTM</i>	23
(2.9) Rumus <i>Perkalian Output Gate LSTM</i>	23

DAFTAR LAMPIRAN

Pengujian <i>Chatbot</i> dengan Pertanyaan Lama	97
Pengujian <i>Chatbot</i> dengan Pertanyaan Baru	496