

PERBANDINGAN TF-IDF *WEIGHTED* WORD2VEC, FASTTEXT, DAN INDOBERT DALAM KLASIFIKASI SENTIMEN SELEKSI PTN DENGAN BILSTM-DENSE

SHERINDA ENGELINA



**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Perbandingan TF-IDF *Weighted* Word2Vec, FastText, dan IndoBERT dalam Klasifikasi Sentimen Seleksi PTN dengan BiLSTM-Dense” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dalam penyusunan karya ini, penulis menggunakan bantuan kecerdasan buatan Grok, GPT, Perplexity, dan Claude untuk melakukan proses pelabelan pada analisis, penyesuaian komputasi, penyesuaian gambar, pemahaman metode lebih mendalam dan pencarian bahan bacaan yang memiliki akses terbuka. Setelah menggunakan alat/layanan tersebut, penulis meninjau dan menyunting konten sesuai kebutuhan serta bertanggung jawab penuh atas isi karya tugas akhir ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juli 2026

Sherinda Engelina
G1401221067



ABSTRAK

SHERINDA ENGELINA. Perbandingan Performa TF-IDF *weighted* Word2Vec, FastText, dan IndoBERT dalam Klasifikasi Sentimen Seleksi PTN dengan BiLSTM-Dense. Dibimbing oleh KUSMAN SADIK dan SEPTIAN RAHARDIANTORO.

Analisis sentimen merupakan pendekatan komputasional yang digunakan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini, perasaan, atau perilaku seseorang terhadap topik tertentu. Dalam analisis sentimen, input data yang digunakan berbentuk teks, sedangkan model *deep learning* seperti BiLSTM-Dense membutuhkan data dalam bentuk numerik sehingga diperlukan teknik ekstraksi fitur untuk mengubah teks menjadi representasi numerik. Penelitian ini membandingkan tiga metode ekstraksi fitur, yaitu TF-IDF *Weighted* Word2Vec sebagai metode *embedding* terboboti frekuensi, FastText metode *embedding* berdasarkan sub n-gram kata, dan IndoBERT sebagai metode *embedding* berbasis kontekstual. Ekstraksi fitur yang dihasilkan kemudian digunakan sebagai input bagi BiLSTM-Dense yang berperan sebagai metode klasifikasi. Penelitian ini bertujuan untuk membandingkan model *hybrid* berdasarkan metrik *accuracy* dan waktu komputasinya serta menganalisis karakteristik performa setiap model *hybrid*. Dataset penelitian diperoleh melalui proses *scraping* pada platform X dan YouTube yang berkaitan dengan ujian masuk perguruan tinggi selama periode Januari 2023 hingga Desember 2025. Hasil penelitian menunjukkan metode TF-IDF *weighted* Word2Vec-BiLSTM-Dense, FastText-BiLSTM-Dense, dan IndoBERT-BiLSTM-Dense berturut-turut memiliki waktu komputasi yang semakin besar dan hal tersebut juga diiringi dengan peningkatan performa *accuracy* yang semakin tinggi. *Accuracy* antara ketiga metode berbeda signifikan dengan tingkat *accuracy* tertinggi diperoleh melalui metode IndoBERT-BiLSTM-Dense sebesar 74,6% pada data uji.

Kata kunci: BiLSTM, FastText, IndoBERT, TF-IDF, Word2Vec.



ABSTRACT

SHERINDA ENGELINA. Comparison of TF-IDF *weighted* Word2Vec, FastText, and IndoBERT for University Selection Sentiment Classification Using BiLSTM-Dense. Supervised by KUSMAN SADIK and SEPTIAN RAHARDIANTORO.

Sentiment analysis is a computational approach used to identify, extract, and classify an individual's opinions, feelings, or behaviors regarding specific topics. While sentiment analysis relies on text-based input, *deep learning* models like BiLSTM-Dense require numerical data. Consequently, feature extraction techniques are necessary to convert text into numerical representations. This study compares three feature extraction methods, TF-IDF *Weighted* Word2Vec as a frequency-*weighted* embedding method, FastText as an embedding method based on sub-word n-grams, and IndoBERT as a context-based embedding method. The resulting extracted features serve as input for the BiLSTM-Dense model, which functions as the classifier. The study aims to compare these *hybrid* models using *accuracy* metrics and computational time, and to analyze the performance characteristics of each model. The dataset was obtained by *scraping* X and YouTube for content related to university entrance exams between January 2023 and December 2025. The results indicate that the computational times for the TF-IDF *weighted* Word2Vec-BiLSTM-Dense, FastText-BiLSTM-Dense, and IndoBERT-BiLSTM-Dense methods increased progressively in that order, and this increase in computational time was accompanied by a corresponding improvement in *accuracy*. Significant differences in *accuracy* were observed among the three methods, with the IndoBERT-BiLSTM-Dense method achieving the highest *accuracy* at 74.6% on the test data.

Keywords: BiLSTM, FastText, IndoBERT, TF-IDF, Word2Vec.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026¹
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

PERBANDINGAN TF-IDF *WEIGHTED* WORD2VEC, FASTTEXT, DAN INDOBERT DALAM KLASIFIKASI SENTIMEN SELEKSI PTN DENGAN BILSTM-DENSE

SHERINDA ENGELINA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Skripsi:
1. Dr. Ir. Budi Susetyo, M.S



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Skripsi : Perbandingan TF-IDF *Weighted* Word2Vec, FastText,
dan IndoBERT dalam Klasifikasi Sentimen Seleksi
PTN dengan BiLSTM-Dense

Nama : Sherinda Engelina
NIM : G1401221067

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Dr. Kusman Sadik, S.Si., M.Si.

Pembimbing 2:

Dr. Septian Rahardiantoro, S.Stat., M.Si.

Diketahui oleh

Ketua Program Studi:

Dr. Bagus Sartono, S.Si., M.Si.
NIP 197804112005011002

Tanggal Ujian:
16 Juli 2026

Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Tuhan yang Maha Esa atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak Januari 2026 sampai Juli 2026 ini ialah analisis data teks, dengan judul “Perbandingan TF-IDF *weighted* Word2Vec, FastText, dan IndoBERT dalam Klasifikasi Sentimen Seleksi PTN dengan BiLSTM-Dense”.

Terima kasih penulis ucapkan kepada seluruh pihak yang telah memberikan bimbingan, dukungan, serta doa dalam penyusunan karya ilmiah ini, kepada:

1. Alm. Yeye selaku kakek tercinta yang meskipun tidak sempat menanti hingga wisuda tapi senantiasa menyertai penulis dengan doa dan kasihnya.
2. Papa, Mama, Cece, Koko dan Epen selaku keluarga penulis yang senantiasa memberikan dukungan, doa, dan kasih sayang.
3. Komisi pembimbing, Dr. Kusman Sadik, S.Si., M.Si. dan Dr. Septian Rahardianto, S.Stat., M.Si. yang telah dengan sabar memberi bimbingan, saran, dan masukan membangun.
4. Dr. Ir. Budi Susetyo, M.S selaku penguji luar, Ir. Aam Alamudi, M.Si selaku moderator seminar hasil, serta La Ode Abdul Rahman, S.Si., M.Si. selaku moderator kolokium yang telah memberikan masukan dan arahan yang sangat berarti dalam penyusunan karya tulis ini.
5. Seluruh dosen Program Studi Statistika dan Sains Data SSMI IPB yang telah memberikan ilmu yang sangat bermanfaat bagi penulis serta staf Tata Usaha Program Studi Statistika dan Sains Data SSMI yang turut menunjang kebutuhan administrasi selama perkuliahan.
6. Ria, Excel, Indryani, teman-teman sejawat yang bersedia untuk membantu penulis meninjau serta memberikan inspirasi bagi penulis dalam menyusun karya tulis ini.
7. KMB IPB yang telah menjadi keluarga kedua dalam mendukung hardskill, softskill, hingga spiritual penulis selama menjalani perkuliahan.
8. Teman-teman TSA IPB yang menjadi lingkungan positif untuk penulis terus mengembangkan diri.
9. Teman-teman statistika 59 yang telah kebersamai dan saling menyemangati selama menjalani perkuliahan ini.
10. Ko Gilbetch yang di setiap kondisi telah dengan tulus menemani, memberikan fasilitas, dan dukungan penuh dalam penyusunan karya tulis ini.
11. Tanoto Foundation yang telah memberikan dukungan finansial dan program pengembangan diri sehingga penulis dapat menjalani perkuliahan tanpa kendala finansial dan mengembangkan diri secara optimal.
12. Statify sebagai badan konsultan data yang telah memberikan kepercayaan dan bimbingan kepada penulis untuk menangani penelitian klien sehingga memperkaya pengalaman penulis dalam melakukan analisis data

Penulis menyadari bahwa skripsi ini masih memiliki banyak keterbatasan, dengan rendah hati penulis memohon maaf terdapat kekurangan dalam penulisan ini. Meskipun begitu, semoga skripsi ini dapat bermanfaat bagi banyak pihak dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2026

Sherinda Engelina



DAFTAR ISI

DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xii
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Manfaat	2
II TINJAUAN PUSTAKA	3
2.1 Seleksi Nasional Penerimaan Mahasiswa Baru	3
2.2 TF-IDF <i>Weighted</i> Word2Vec	3
2.3 FastText	6
2.4 IndoBERT	7
2.5 Bidirectional Long Short- Term Memory-Dense (BiLSTM-Dense)	9
2.6 <i>Repeated Measures</i> ANOVA	11
2.7 Uji Beda Nyata Terkecil (BNT)	11
2.8 Uji Shapiro-Wilk	12
2.9 Uji Mauchly	13
2.10 Metrik Evaluasi Klasifikasi	14
III METODE	15
3.1 Perangkat	15
3.2 Data	15
3.3 Analisis Data	15
IV HASIL DAN PEMBAHASAN	18
4.1 Pelabelan Data	18
4.2 Praproses Data	18
4.3 Eksplorasi Data	19
4.4 Pemodelan TF-IDF <i>Weighted</i> Word2Vec- BiLSTM-Dense	21
4.5 Pemodelan model <i>hybrid</i> FastText-BiLSTM-Dense	26
4.6 Pemodelan model <i>hybrid</i> IndoBERT-BiLSTM-Dense	29
4.7 Uji Signifikansi <i>Accuracy</i> Model <i>Hybrid</i>	34
V SIMPULAN DAN SARAN	36
5.1 Simpulan	37
5.2 Saran	38
DAFTAR PUSTAKA	39
LAMPIRAN	43
RIWAYAT HIDUP	48



DAFTAR TABEL

1	Parameter TF-IDF	5
2	Parameter Word2Vec	6
3	Parameter FastText	7
4	Parameter IndoBERT	8
5	Parameter BiLSTM-Dense	10
6	Spesifikasi <i>Google Colaboratory</i>	15
7	Arsitektur BiLSTM-Dense	16
8	Ilustrasi proses pelabelan	18
9	Ilustrasi tahap praproses data	19
10	Arsitektur ekstraksi fitur TF-IDF	22
11	Arsitektur ekstraksi fitur Word2Vec	22
12	Ilustrasi hasil ekstraksi fitur TF-IDF <i>Weighted</i> Word2Vec	23
13	Bentuk input BiLSTM-Dense ekstraksi fitur TF-IDF <i>Weighted</i> Word2Vec	23
14	Arsitektur ekstraksi fitur FastText	26
15	Ilustrasi hasil ekstraksi fitur FastText	26
16	Bentuk input BiLSTM-Dense hasil ekstraksi fitur FastText	27
17	Arsitektur ekstraksi fitur IndoBERT	30
18	Ilustrasi hasil ekstraksi fitur IndoBERT	30
19	Bentuk input BiLSTM-Dense hasil ekstraksi fitur IndoBERT	31
20	Perbandingan rata-rata <i>accuracy</i> pelatihan dan waktu komputasi	34
21	<i>Repeated measures</i> ANOVA	34
22	Hasil uji asumsi	35
23	Hasil uji lanjut BNT perbandingan <i>mean accuracy</i> antar model <i>hybrid</i>	35



DAFTAR GAMBAR

1	Arsitektur model pelatihan pada Word2Vec	4
2	Proporsi amatan setiap kelas	19
3	<i>Wordcloud</i> setiap kelas	20
4	(A) sebaran panjang teks dan 4(B) Jumlah kata unik setiap kelas	21
5	<i>Heatmap</i> performa parameter TF-IDF Word2Vec-BiLSTM-Dense	24
6	Grafik <i>learning curve</i> TF-IDF <i>weighted</i> Word2Vec-BiLSTM-Dense	25
7	Perbandingan metrik evaluasi TF-IDF-Word2Vec BiLSTM-Dense	25
8	<i>Heatmap</i> performa parameter FastText-BiLSTM-Dense	28
9	Grafik <i>learning curve</i> FastText-BiLSTM-Dense	28
10	Perbandingan nilai metrik evaluasi FastText-BiLSTM-Dense	29
11	<i>Heatmap</i> performa parameter IndoBERT-BiLSTM-Dense	32
12	Grafik <i>learning curve</i> IndoBERT-BiLSTM-Dense	33
13	Perbandingan nilai metrik evaluasi IndoBERT-BiLSTM-Dense	33

DAFTAR LAMPIRAN

1	Lampiran 1 hasil <i>hyperparameter</i> tuning TF-IDF Word2Vec-BiLSTM-Dense	44
2	Lampiran 2 hasil <i>hyperparameter</i> tuning FastText-BiLSTM-Dense	45
3	Lampiran 3 hasil <i>hyperparameter</i> tuning IndoBERT-BiLSTM-Dense	45
4	Lampiran 4 <i>Syntax</i> pelabelan dengan Grok AI	46