

**PERBANDINGAN AKURASI METODE WORD EMBEDDING
TF-IDF DAN WORD2VEC MENGGUNAKAN RECURRENT
NEURAL NETWORK UNTUK ANALISIS SENTIMEN TWEET
VAKSINASI COVID-19**

SKRIPSI



Oleh :

NADIA RISTYA DEWI

18081010071

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"

JAWA TIMUR

2021

**PERBANDINGAN AKURASI METODE WORD EMBEDDING
TF-IDF DAN WORD2VEC MENGGUNAKAN RECURRENT
NEURAL NETWORK UNTUK ANALISIS SENTIMEN TWEET
VAKSINASI COVID-19**

SKRIPSI

Diajukan untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Komputer

Program Studi Informatika



Oleh :

NADIA RISTYA DEWI

18081010071

**PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
JAWA TIMUR
2021**

LEMBAR PENGESAHAN SKRIPSI

Judul : PERBANDINGAN AKURASI METODE WORD EMBEDDING TF-IDF DAN
WORD2VEC MENGGUNAKAN RECURRENT NEURAL NETWORK
UNTUK ANALISIS SENTIMEN TWEET VAKSINASI COVID-19
Oleh : NADIA RISTYA DEWI
NPM : 18081010071

Telah Diseminarkan Dalam Ujian Skripsi Pada :

Mengetahui

Dosen Pembimbing

1.



Eva Yulia Pusaningrum, S.Kom, M.Kom

NIP. 19890705 2021212 002

2.



Hendra Maulana, S.Kom, M.Kom

NPT. 201198 31 223248

Dosen Penguji

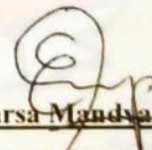
1.



Yisti Vita Yia, S.ST, M.Kom.

NIP. 19860425 2021212 001

2.



Eka Prakarsa Mandasrtha, S.T., M.Kom.

NIP. 19880525|2018031 001

Menyetujui

Dekan

Fakultas Ilmu Komputer



Dr. Ir. Ni Ketut Sari, MT

NIP. 19650731 199203 2 001

Koordinator Program Studi

Teknik Informatika



Budi Nugroho, S.Kom, M.Kom

NIP. 19800907 2021211 005

SURAT PERNYATAAN ANTI PLAGIAT

Saya, mahasiswa program studi Informatika UPN "Veteran" Jawa Timur,
yang bertanda tangan di bawah ini:

Nama : Nadia Ristya Dewi

NPM : 18081010071

Menyatakan bahwa judul skripsi yang saya ajukan dan kerjakan yang berjudul :

**“PERBANDINGAN AKURASI METODE WORD EMBEDDING TF-IDF
DAN WORD2VEC MENGGUNAKAN RECURRENT NEURAL
NETWORK UNTUK ANALISIS SENTIMEN TWEET VAKSINASI
COVID-19**

Bukan merupakan plagiat dari skripsi/tugas akhir/penelitian orang lain dan juga bukan merupakan produk dan atau software yang saya beli daripihak lain. Saya juga menyatakan bahwa skripsi ini adalah pekerjaan saya sendiri, kecuali yang dinyatakan dalam daftar pustaka dan tidak pernah diajukan untuk syarat memperoleh gelar di UPN "Veteran" Jawa Timur maupun di institusi pendidikan lain.

Jika ternyata di kemudian hari pernyataan ini terbukti tidak benar, maka Saya siap menerima segala konsekuensinya.

Surabaya, 20 Januari 2022

Penulis,




NADIA RISTYA DEWI

NPM. 18081010071

PERBANDINGAN AKURASI METODE WORD EMBEDDING TF-IDF DAN WORD2VEC MENGGUNAKAN RECURRENT NEURAL NETWORK UNTUK ANALISIS SENTIMEN TWEET VAKSINASI COVID-19

Nama Mahasiswa : Nadia Ristya Dewi
NPM : 18081010071
Program Studi : Informatika
Dosen Pembimbing : Eva Yulia Puspaningrum, S.Kom, M.Kom
Hendra Maulana, S.Kom, M.Kom

ABSTRAK

Sosial media merupakan sarana komunikasi maya yang saat ini banyak digunakan oleh manusia. Di sosial media banyak terdapat informasi yang mengandung opini, hal ini menyebabkan sering menyebabkan terjadinya ketimpangan antara informasi yang sedang dibicarakan dan informasi yang diinterpretasikan serta banyaknya penyebaran informasi yang tidak benar adanya (hoax) sehingga tidak benar-benar bisa dipahami sentimen yg ingin disampaikan, terlebih pada dua tahun terakhir ini sangat diperlukan informasi yang benar mengenai topik kesehatan. Penelitian ini bertujuan untuk mengetahui akurasi dari percobaan pembelajaran mesin dengan menerapkan metode word embedding TF-IDF dan Word2Vec yang akan melakukan analisis terhadap sentimen mengenai tweet dengan topik vaksinasi Covid-19. Penelitian ini dilakukan dengan menerapkan kedua metode word embedding tersebut pada algoritma Recurrent Neural Network (RNN) dengan menggunakan 6490 data tweet. Data yang dimiliki dibagi dengan perbandingan 7:3 untuk data training dan data testing. Percobaan menggunakan RNN-Word2Vec menghasilkan akurasi sebesar 51.71%, sedangkan percobaan menggunakan RNN-Word2Vec menghasilkan akurasi sebesar 50.73%.

Kata kunci: Ekstraksi fitur, Word2Vec, TD-IDF, RNN, Sentimen Analisis.

KATA PENGANTAR

Dengan mengucapkan puji dan syukur kehadirat Allah SWT yang telah memberikan petunjuk, pertolongan serta seluruh nikmat-Nya kepada kita, dengan izin dan ridho-Nya lah penulis dapat menyelesaikan laporan skripsi ini yang berjudul “Perbandingan Akurasi Metode Word Embedding TF-IDF dan Word2vec Menggunakan Recurrent Neural Network untuk Analisis Sentimen Tweet Vaksinasi Covid-19”

Banyak dukungan dan bantuan yang didapatkan selama melakukan penelitian hingga akhirnya mampu menyelesaikan penulisan laporan skripsi ini, Dengan rasa hormat, penulis mengucapkan terima kasih kepada seluruh pihak terkait yang turut membantu dan terlibat dalam penyusunan laporan ini dari awal hingga akhir.

Penulis menyadari mungkin masih ada kekurangan pada laporan skripsi ini mengingat keterbatasan pengetahuan dan kemampuan. Oleh karena itu, penulis sangat mengharapkan masukan, kritik, dan saran yang dapat membangun dan mengembangkan laporan ini.

Surabaya, 20 Januari 2022

Penyusun

UCAPAN TERIMA KASIH

Dalam penyusunan dan penulisan skripsi ini tidak terlepas dari bantuan, bimbingan, serta dukungan dari beberapa pihak. Oleh karena itu dalam kesempatan ini penulis menyampaikan terima kasih kepada Allah SWT yang selalu ada dan memberi kelancaran, kemudahan, kekuatan, dan sumber dari segala pencerahan. Tidak lupa penulis berterima kasih kepada yang terhormat :

1. Bapak Prof. Dr. Ir. Akhmad Fauzi, M.MT selaku Rektor Universitas Pembangunan Nasional “Veteran” Jawa Timur.
2. Ibu Dr. Ir. Ni Ketut Sari, M.T. selaku Dekan Fakultas Ilmu Komputer Universitas Pembangunan Nasional “Veteran” Jawa Timur.
3. Bapak Budi Nugroho, S.Kom., M.Kom., selaku Ketua Program Studi Teknik Informatika Universitas Pembangunan Nasional “Veteran” Jawa Timur.
4. Ibu Eva Yulia Puspaningrum, S.Kom, M.Kom. selaku dosen pembimbing I dan Bapak Hendra Maulana, S.Kom, M.Kom. selaku dosen pembimbing II, yang telah sabar, ikhlas meluangkan waktu, tenaga dan pikiran untuk memberikan bimbingan, motivasi, arahan, dan saran-saran yang sangat berharga kepada penulis selama penyusunan skripsi ini.
5. Staff dan dosen Teknik Informatika Universitas Pembangunan Nasional “Veteran” Jawa Timur yang telah memberikan ilmu, arahan, dan pengalamannya selama perkuliahan.
6. Kedua orang tua penulis, A. Haris Sudjoko dan Listyawati yang selalu memberi semangat, motivasi, dan mendoakan untuk kebaikan, keberhasilan maupun kesuksesan penulis.
7. Adik dan kakak kandung penulis, Octgi Ristya Perdana dan Desya Ristya Putri yang senantiasa mendoakan dan mendukung penulis hingga penyusunan skripsi ini selesai.
8. Helna Frecenta, Taufiqur Rohman, dan Bagas Cakra Wiradana yang telah sabar berjuang bersama dalam suka maupun duka dengan saling menguatkan dan saling memberikan dukungan selama perkuliahan di

Teknik Informatika Universitas Pembangunan Nasional “Veteran” Jawa Timur.

9. Teman-teman Teknik Informatika angkatan 2018, yang sudah bersedia membantu, mengisi, dan menjadi bagian dalam cerita hidup dan perkuliahan penulis.
10. Semua pihak yang tidak dapat penulis ucapkan yang juga turut mendoakan dan menyemangati penulis dalam menyelesaikan skripsi ini.

Terimakasih atas segala bantuan, dukungan, dan doa-doanya, semoga Allah SWT memberikan balasan yang setimpal kepada semua pihak yang telah membantu penulis sehingga penulis mampu dan berhasil menyelesaikan penelitian dan proses penyusunan laporan penelitian skripsi dengan baik dan tepat

Surabaya, 20 Januari 2022

Penulis

DAFTAR ISI

LEMBAR PENGESAHAN SKRIPSI	Error! Bookmark not defined.
ABSTRAK.....	iii
KATA PENGANTAR	iv
UCAPAN TERIMA KASIH.....	v
DAFTAR ISI.....	vii
BAB I.....	1
PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	4
1.3 Batasan Masalah	5
1.4 Tujuan Penelitian	5
1.5 Manfaat Penelitian	6
BAB II.....	7
TINJAUAN PUSTAKA.....	7
2.1 Penelitian Pendahulu	7
2.2 Dasar Teori	9
BAB III	18
METODOLOGI	18
3.1 Tahapan pengerjaan tugas akhir	18
3.2 Studi Literatur.....	19
3.3 Analisa dan Desain	19
3.4 Akuisisi Data dan Labelling	20
3.5 Preprocessing dan Ekstraksi Fitur	24
3.6 Klasifikasi Data	29

3.7	Pengujian Sistem dengan Confusion Matrix	30
BAB IV	31
HASIL DAN PEMBAHASAN	31
4.1.	Lingkungan Penelitian	31
4.2.	Proses pengambilan data.....	32
4.3.	Preprocessing data	36
4.4.	Ekstraksi Fitur.....	42
4.5.	Split train dan test	49
4.6.	Proses Klasifikasi.....	50
4.7.	Proses Pengujian.....	64
BAB V	75
PENUTUP	75
5.1	Kesimpulan.....	75
5.2	Saran	75
DAFTAR PUSTAKA	77

DAFTAR TABEL

Tabel 2.1 Case Folding	10
Tabel 2.2 Tokenizing	10
Tabel 2.3 Stopword removal.....	11
Tabel 2.4 Stemming	11
Tabel 2.5 Multiclass Confusion Matrix	16
Tabel 3.1 Tahapan Pengerjaan Tugas Akhir.....	18
Tabel 3.2 Label pada tweet	23
Tabel 4.1 Spesifikasi perangkat keras yang digunakan untuk penelitian	31
Tabel 4.2 Spesifikasi perangkat lunak yang digunakan untuk penelitian	32
Tabel 4.3 epoch, akurasi, dan waktu komputasi setiap skenario	62
Tabel 4.4 Confusion matrix RNN-TF-IDF	65
Tabel 4.5 Confusion matrix RNN-Word2Vec	66
Tabel 4.6 Classification Report RNN-TF-IDF	67
Tabel 4.7 Confusion matrix RNN-Word2Vec	67

DAFTAR GAMBAR

Gambar 2.1 CBOW Word2Vec	13
Gambar 2.2 Skip-gram Word2Vec	14
Gambar 2.3 Recurrent Neural Network	15
Gambar 3.1 Flowchart sistem	19
Gambar 3.2 Tahapan Akuisisi Data & Labelling.....	20
Gambar 3.3 Grafik tren pencarian keyword	21
Gambar 3.4 Crawling Tweet.....	22
Gambar 3.5 Labelling	24
Gambar 3.6 Pemisahan data text untuk tahap preprocessing.....	25
Gambar 3.7 Tahapan Preprocessing	25
Gambar 3.8 Tweet berdasarkan kelas	26
Gambar 3.0.9 Tweet per kelas setelah dilakukan case folding	26
Gambar 3.10 Tweet per kelas setelah dilakukan tokenizing.....	27
Gambar 3.11 Tweet per kelas setelah dilakukan stopword removal.....	27
Gambar 3.12 Tweet per kelas setelah dilakukan stemming.....	27
Gambar 3.13 File csv berisi tweet setelah preprocessing	28
Gambar 3.14 Tahapan Klasifikasi.....	30
Gambar 4.1 API token Twitter Developer	33
Gambar 4.2 Kumpulan file csv hasil crawling tweet.....	35
Gambar 4.3 Proses pelabelan data tweet.....	35
Gambar 4.4 List yang akan diproses.....	36
Gambar 4.5 Data setelah melalui proses case folding	38
Gambar 4.6 Data setelah melalui proses tokenizing	39
Gambar 4.7 Daftar stopword.....	40
Gambar 4.8 Data setelah melalui proses stopword removal.....	40
Gambar 4.9 Data setelah melalui proses stemming	41
Gambar 4.10 File csv berisi data setelah preprocessing	42
Gambar 4.0.11 Data text dalam bentuk list untuk diterapkan TF-IDF	43
Gambar 4.12 Vector TF-IDF dari data text.....	44
Gambar 4.13 Numpy array integer data text.....	45
Gambar 4.14 Data text dalam bentuk list untuk diterapkan Word2Vec	46

Gambar 4.15 Vector Word2Vec dari data text	48
Gambar 4.16 Data label	49
Gambar 4.17 Numpy array dari data label	49
Gambar 4.18 Hasil pengujian Skenario 1	52
Gambar 4.19 Grafik pelatihan model Skenario 1	52
Gambar 4.20 Hasil pengujian Skenario 2	53
Gambar 4.21 Grafik pelatihan model Skenario 2	53
Gambar 4.22 Hasil pengujian Skenario 3	54
Gambar 4.23 Grafik pelatihan model Skenario 3	54
Gambar 4.24 Hasil pengujian Skenario 4	55
Gambar 4.25 Grafik pelatihan model Skenario 4	55
Gambar 4.26 Hasil pengujian Skenario 5	56
Gambar 4.27 Grafik pelatihan model Skenario 5	56
Gambar 4.28 Hasil pengujian Skenario 6	57
Gambar 4.29 Grafik pelatihan model Skenario 1	57
Gambar 4.30 Grafik percobaan 5 epoch pada masing-masing skenario.....	58
Gambar 4.31 Grafik percobaan 10 epoch pada masing-masing skenario.....	59
Gambar 4.32 Grafik percobaan 15 epoch pada masing-masing skenario.....	60
Gambar 4.33 Akurasi test model scenario	61
Gambar 4.34 Waktu Komputasi tiap scenario	61
Gambar 4.35 Grafik akurasi Skenario 6 pada setiap epoch	63
Gambar 4.36 Grafik akurasi Skenario 6 pada setiap epoch	63
Gambar 4.37 Akurasi testing model RNN-TF-IDF	64
Gambar 4.38 Akurasi testing model RNN-Word2Vec	64
Gambar 4.39 Classification Report RNN-TF-IDF.....	66
Gambar 4.40 Classification Report RNN-Word2Vec	66
Gambar 4.41 Waktu koputasi scenario 1	69
Gambar 4.42 Waktu koputasi scenario 2	69
Gambar 4.43 Waktu koputasi scenario 3	70
Gambar 4.44 Waktu koputasi scenario 4	71
Gambar 4.45 Waktu koputasi scenario 5	72
Gambar 4.46 Waktu koputasi scenario 6	73