



**EVALUASI PERFORMA EKSTRAKSI FITUR BERBASIS N-GRAM DAN
WORD2VEC PADA DATA TWITTER ANALISIS SENTIMEN COVID-19**

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer**

Oleh

MOHAMMAD RYAN FADHILLAH

NIM 1911016210015

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

JUNI 2023



**EVALUASI PERFORMA EKSTRAKSI FITUR BERBASIS N-GRAM DAN
WORD2VEC PADA DATA TWITTER ANALISIS SENTIMEN COVID-19**

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer**

Oleh

MOHAMMAD RYAN FADHILLAH

NIM 1911016210015

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

JUNI 2023

SKRIPSI

EVALUASI PERFORMA EKSTRAKSI FITUR BERBASIS N-GRAM DAN WORD2VEC PADA DATA TWITTER ANALISIS SENTIMEN COVID-19

Oleh:

MOHAMMAD RYAN FADHILLAH

NIM 1911016210015

Telah dipertahankan di depan Dosen Penguji pada tanggal 15 Juni 2023.

Susunan Dosen Penguji:

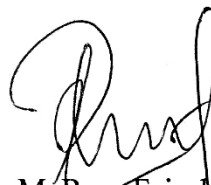
Pembimbing I



Muhammad Itqan Mazdadi, S.Kom, M.Kom.

NIP. 199006122019031013

Dosen Penguji I



M. Reza Faisal, S.T., M.T., Ph.D.

NIP. 197612202008121001

Pembimbing II



Triando Hamonangan Saragih, S.Kom., M.Kom.

NIP. 199308242019031012

Dosen Penguji II



Rudy Herteno, S.Kom., M.Kom.

NIP. 198809252022031003

Banjarbaru, 21 Juni 2023

Koordinator Program Studi Ilmu Komputer



Ryan Sulaiman, S.T., M.Kom

NIP. 197703252008121001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam Daftar Pustaka.

Banjarbaru, 15 Juni 2023

Yang Menyatakan,



Mohammad Ryan Fadhillah

NIM. 1911016210015

ABSTRAK

EVALUASI PERFORMA EKSTRAKSI FITUR BERBASIS N-GRAM DAN WORD2VEC PADA DATA TWITTER ANALISIS SENTIMEN COVID-19

(Oleh : Mohammad Ryan Fadhillah; Pembimbing: Muhammad Itqan Mazdadi, S.Kom, M.Kom. dan Triando Hamonangan Saragih, S.Kom., M.Kom.; 2023; 55 halaman)

Data teks tidak dapat diproses langsung oleh algoritma pembelajaran mesin karena masih berupa data tidak terstruktur, sehingga perlu terlebih dahulu diubah menjadi data terstruktur melalui proses yang disebut ekstraksi fitur agar selanjutnya dapat dilakukan klasifikasi. Masih belum ada penelitian yang mencoba mengetahui pengaruh dari perubahan jumlah dataset yang digunakan pada tahap ekstraksi fitur tersebut terhadap hasil akurasi dari klasifikasi data. Oleh karena itu, perlu diketahui berapa besar pengaruh variasi jumlah data terhadap ekstraksi fitur dalam melakukan klasifikasi data. Evaluasi performa ekstraksi fitur dilakukan dengan menggunakan COVID-19 Indonesian Tweet Dataset. Variasi jumlah data yang digunakan adalah 400 data, 800 data, 1600 data, dan 3200 data. Penelitian ini menggunakan algoritma ekstraksi fitur N-Gram yaitu Unigram, Bigram, dan Trigram serta Word2Vec dengan algoritma klasifikasi Naïve Bayes Classifier. Algoritma klasifikasi membagi dataset tersebut menjadi dua kelas, yaitu positif dan negatif. Hasil penelitian menunjukkan hasil akurasi yang didapatkan tidak selalu bertambah seiring dengan bertambahnya jumlah data yang digunakan. Nilai akurasi tertinggi diperoleh ekstraksi fitur Unigram dengan menggunakan 3200 data yaitu sebesar 78.75%. Bigram dan Trigram juga memperoleh hasil akurasi tertinggi di 3200 data, sedangkan Word2Vec meraih hasil akurasi tertinggi di 800 data. Penelitian ini membuktikan bahwa variasi jumlah data yang lebih besar belum bisa menjamin bahwa nilai akurasi yang didapatkan akan lebih baik.

Kata kunci: Evaluasi Performa, Klasifikasi, N-Gram, Word2Vec, *Naïve Bayes Classifier*

ABSTRACT

PERFORMANCE EVALUATION OF N-GRAM AND WORD2VEC BASED FEATURE EXTRACTION ON COVID-19 SENTIMENT ANALYSIS TWITTER DATA (By : Mohammad Ryan Fadhillah; Supervisor: Muhammad Itqan Mazdadi, S.Kom, M.Kom. and Triando Hamonangan Saragih, S.Kom., M.Kom.; 2023; 55 pages)

Text data cannot be processed directly by machine learning algorithms because it is still unstructured data, so it needs to be converted into structured data through a process called feature extraction so it can be classified. There is still no research that tries to determine the effect of changing the datasets number used at the feature extraction stage on the accuracy results of data classification. Therefore, it is necessary to know how much influence the amount of data variation has on feature extraction in classifying data. Evaluation of feature extraction performance is using the COVID-19 Indonesian Tweet Dataset. Variations in the amount of data used are 400 data, 800 data, 1600 data, and 3200 data. This study uses the N-Gram feature extraction algorithm, specifically Unigram, Bigram, and Trigram as well as Word2Vec with the Naïve Bayes Classifier classification algorithm. The classification algorithm divides the dataset into two classes, namely positive and negative. The results of the study show that the accuracy results obtained do not always increase with the increase in the amount of data used. The highest accuracy was obtained by the Unigram feature extraction using 3200 data that is 78.75%. Bigram and Trigram also obtained the highest accuracy results on 3200 data, while Word2Vec achieved the highest accuracy results on 800 data. This study proves that a larger variation in the amount of data cannot guarantee that the accuracy value obtained will be better.

Keywords: *Performance Evaluation, Classification, N-Gram, Word2Vec, Naïve Bayes Classifier*

PRAKATA

Puji syukur penulis panjatkan ke Tuhan kita Yang Maha Esa karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul “EVALUASI PERFORMA EKSTRAKSI FITUR BERBASIS N-GRAM DAN WORD2VEC PADA DATA TWITTER ANALISIS SENTIMEN COVID-19” untuk memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat.

Pada lembar ini penulis ingin menyampaikan ucapan terimakasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut :

1. Ayah saya, Yunea Rahman dan ibu saya, Ruhlinidyanti yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian skripsi ini.
2. Bapak Muhammad Itqan Mazdadi, S.Kom, M.Kom. selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
3. Bapak Triando Hamonangan Saragih, S.Kom., M.Kom. selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Irwan Budiman S.T., M.Kom selaku Ketua Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
5. Seluruh Dosen dan staff Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
6. Teman-teman dan sahabat-sahabat keluarga Ilmu Komputer angkatan 2019 yang memberikan dukungan dan selalu mengingatkan serta mendoakan dalam proses mengerjakan skripsi ini.
7. Ucapan terima kasih kepada Nurul Afna yang memberikan bantuan dan dukungan selama proses penyelesaian skripsi ini.
8. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna, namun penulis mengharapkan bantuan serupa berupa saran dan kritik yang membangun dari semua pihak demi kesempurnaan dan mutu penulisan skripsi ini.

Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya serta mendapat keridhaan Allah SWT.

Banjarbaru, 15 Juni 2023



Mohammad Ryan Fadhillah

DAFTAR ISI

	Halaman
HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	Kesalahan! Bookmark tidak ditentukan.
HALAMAN PERNYATAAN	iii
ABSTRAK	iv
<i>ABSTRACT</i>	v
PRAKATA.....	vi
DAFTAR ISI.....	viii
DAFTAR TABEL.....	x
DAFTAR GAMBAR	xi
DAFTAR LAMPIRAN.....	xii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan	3
1.4 Manfaat Penelitian.....	4
1.5 Batasan Masalah	4
BAB II TINJAUAN PUSTAKA.....	5
2.1 Kajian Terdahulu	5
2.2 Landasan Teori	12
2.2.1 Data Mining	12
2.2.2 N-Gram	13
2.2.3 Word2Vec	15
2.2.4 TF-IDF	16
2.2.5 <i>Naïve Bayes Classifier</i>	17
2.2.6 <i>Confusion Matrix</i>	18
BAB III METODE PENELITIAN.....	20
3.1 Bahan Penelitian	20
3.2 Alat Penelitian	20
3.3 Prosedur Penelitian	20
BAB IV HASIL DAN PEMBAHASAN	24

4.1	Hasil.....	24
4.1.1	Pengumpulan Dataset.....	24
4.1.2	<i>Preprocessing Data</i>	25
4.1.3	Pembagian <i>Data Training</i> dan <i>Data Testing</i>	29
4.1.4	Pembobotan TF-IDF	30
4.1.5	Ekstraksi Fitur	30
4.1.6	Klasifikasi <i>Naïve Bayes Classifier</i>	34
4.1.7	Evaluasi.....	37
4.2	Pembahasan	42
BAB V PENUTUP.....		53
5.1	Kesimpulan.....	53
5.2	Saran	53
DAFTAR PUSTAKA		54
LAMPIRAN.....		56

DAFTAR TABEL

Tabel	Halaman
Tabel 1. Keaslian Penelitian.....	9
Tabel 2. Confusion Matrix	19
Tabel 3. Contoh Data	21
Tabel 4. Nama Fitur	24
Tabel 5. Contoh <i>COVID-19 Tweet Indonesia Dataset</i>	24
Tabel 6. Tahapan <i>Preprocessing Remove URLs, Symbols, dan Emojis</i>	26
Tabel 7. Tahapan <i>Preprocessing Remove Hastags dan Mention</i>	27
Tabel 8. Tahapan <i>Preprocessing Remove Punctuation dan Case Folding</i>	28
Tabel 9. Tahapan <i>Preprocessing Convert Slangword</i>	28
Tabel 10. Tahapan <i>Preprocessing Remove Stopword dan Stemming</i>	29
Tabel 11. Hasil perhitungan TF-IDF.....	30
Tabel 12. Hasil Perhitungan Unigram.....	31
Tabel 13. Hasil Perhitungan Bigram.....	32
Tabel 14. Hasil Perhitungan Trigram.....	32
Tabel 15. Hasil perhitungan ekstraksi fitur Word2Vec	33
Tabel 16. Contoh Hasil Klasifikasi 400 Data dengan Metode <i>Naïve Bayes Classifier</i> dengan Ekstraksi Fitur Word2Vec	34
Tabel 17. Hasil Klasifikasi 800 Data dengan Metode <i>Naïve Bayes Classifier</i> dengan Ekstraksi Fitur Unigram.....	35
Tabel 18. Hasil Klasifikasi 1600 Data dengan Metode <i>Naïve Bayes Classifier</i> dengan Ekstraksi Fitur Bigram	36
Tabel 19. Hasil Klasifikasi 3200 Data dengan Metode <i>Naïve Bayes Classifier</i> dengan Ekstraksi Fitur Trigram.....	37
Tabel 20. Hasil Akurasi dengan Variasi Jumlah 400 Data	38
Tabel 21. Hasil Akurasi dengan Variasi Jumlah 800 Data	39
Tabel 22. Hasil Akurasi dengan Variasi Jumlah 1600 Data	40
Tabel 23. Hasil Akurasi dengan Variasi Jumlah 3200 Data	41
Tabel 24. Fitur dalam Data Training dengan Ekstraksi Fitur Unigram	51
Tabel 25. Fitur dalam Data Training dengan Ekstraksi Fitur Bigram.....	51
Tabel 26. Fitur dalam Data Training dengan Ekstraksi Fitur Trigram	51

DAFTAR GAMBAR

Gambar	Halaman
Gambar 1. Proses Data Mining	12
Gambar 2. Model N-Gram	13
Gambar 3. Jenis-Jenis N-Gram	14
Gambar 4. Word2Vec (CBOW and Skip-gram)	15
Gambar 5. Input dan Output dari Algoritma Word2Vec	15
Gambar 6. Alur Klasifikasi <i>Naïve Bayes Classifier</i>	17
Gambar 7. Alur Penelitian.....	21
Gambar 8. Alur Preprocessing	26
Gambar 9. Diagram Hasil Akurasi dengan Variasi Jumlah 400 Data.....	38
Gambar 10. Diagram Hasil Akurasi dengan Variasi Jumlah 800 Data.....	39
Gambar 11. Diagram Hasil Akurasi dengan Variasi Jumlah 1600 Data.....	40
Gambar 12. Diagram Hasil Akurasi dengan Variasi Jumlah 3200 Data.....	41
Gambar 13. Hasil akurasi dengan ekstraksi fitur Word2Vec.....	43
Gambar 14. Hasil akurasi dengan ekstraksi fitur Unigram	44
Gambar 15. Hasil akurasi dengan ekstraksi fitur Bigram	45
Gambar 16. Hasil akurasi dengan ekstraksi fitur Trigram	46
Gambar 17. Nilai Akurasi Tertinggi	47
Gambar 18. Rata-Rata Nilai Akurasi Berdasarkan Variasi Jumlah Data.....	48
Gambar 20. Perbandingan hasil akurasi dari ekstraksi fitur Word2Vec dan Unigram	49

DAFTAR LAMPIRAN

- Lampiran 1 Perhitungan Manual Sampel Data menggunakan N-Gram
- Lampiran 2. Perhitungan Manual Sampel Data menggunakan Word2Vec
- Lampiran 3. Riwayat Hidup Penulis