



**PERBANDINGAN EKSTRAKSI FITUR BERBASIS *N-GRAM* DENGAN
PEMBOBOTAN TF-IDF DAN ALGORITMA KLASIFIKASI *RANDOM
FOREST* DAN SVM UNTUK DETEKSI PELAPORAN GEJALA COVID-19
DARI TWITTER**

Skripsi

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

**Oleh
ALMADEA RUSELLAWATI
NIM. 1611016220002**

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

JUNI 2023



**PERBANDINGAN EKSTRAKSI FITUR BERBASIS *N-GRAM* DENGAN
PEMBOBOTAN TF-IDF DAN ALGORITMA KLASIFIKASI *RANDOM
FOREST* DAN SVM UNTUK DETEKSI PELAPORAN GEJALA COVID-19
DARI TWITTER**

Skripsi

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

ALMADEA RUSELLAWATI

NIM. 1611016220002

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

JUNI 2023

SKRIPSI

PERBANDINGAN EKSTRAKSI FITUR BERBASIS *N-GRAM* DENGAN
PEMBOBOTAN TF-IDF DAN ALGORITMA KLASIFIKASI *RANDOM*
FOREST DAN SVM UNTUK DETEKSI PELAPORAN GEJALA COVID-19
DARI TWITTER

Oleh:

ALMADEA RUSELLAWATI

NIM.1611016220002

Telah dipertahankan di depan Dosen Penguji pada tanggal 23 Juni 2023

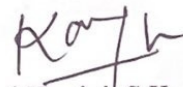
Susunan Dosen Penguji:

Pembimbing I



Mohammad Reza Falsal, S.T, M.T, Ph.D
NIP.197612202008121001

Dosen Penguji I



Dwi Kartini, S.Kom, M.Kom
NIP.198704212012122003

Pembimbing II



Muhammad Itqan Mazdadi, M.Kom.
NIP. 199006122019031013

Dosen Penguji II



Triando Hamonangan Saragih,
S.Kom., M.Kom.
NIP. 199308242019031012

Banjarbaru, Juni 2023

Ketua Program Studi Ilmu Komputer

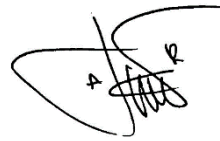


Iwan Budiman, S.T., M.Kom
NIP.197703252008121001

PERYATAAN

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 diberikan orang lain, kecuali yang secara tertulis diacu didalam naskah ini dan disebutkan dalam Daftar Pustaka.

Banjarbaru, 23 Juni 2023

A handwritten signature in black ink, appearing to be 'Almadea Rusellawati', with a stylized flourish above it.

Almadea Rusellawati
NIM.1611016220002

ABSTRAK

PERBANDINGAN EKSTRAKSI FITUR BERBASIS *N-GRAM* DENGAN PEMBOBOTAN TF-IDF DAN ALGORITMA KLASIFIKASI *RANDOM FOREST* DAN SVM UNTUK DETEKSI PELAPORAN GEJALA COVID-19 DARI TWITTER (Oleh : Almadea Rusellawati; Pembimbing : Mohammad Reza Faisal, S.T, M.T, Ph.D dan Muhammad Itqan Mazdadi, M.Kom.; 2023; 63 halaman)

Virus Covid-19 hingga saat ini masih sangat banyak menimbulkan gejala-gejala kepada masyarakat. Meski tidak separah dulu, tetapi Covid-19 juga belum kunjung hilang. Saat ini masyarakat sudah dengan sangat terbuka menceritakan tentang diri mereka yang sedang terkena atau memiliki gejala Covid-19 di berbagai media sosial mereka, salah satunya adalah Twitter. Dataset yang digunakan untuk penelitian ini adalah data gejala Covid-19 dari penelitian sebelumnya oleh Sari (2022). Pembobotan TF-IDF digunakan untuk menghitung jumlah kemunculan term pada data, dilakukan pengujian kembali menggunakan *N-Gram* sebagai fitur pemisah kata. Dimana *N-Gram* tersebut dibagi menjadi 3 yaitu *unigram*, *bigram* & *trigram*. Penelitian ini menggunakan klasifikasi *Random Forest* dan *Support Vector Machine* yang melakukan pengujian pada kernel *Linear*, *Radial Basic Function*, dan *Polyomial*. Performa hasil akurasi tertinggi diperoleh dari pengujian *Unigram* dengan *Support Vector Machine* kernel *Linear* sebesar 88,8%.

Kata Kunci : Ekstraksi Fitur, Gejala Covid-19, *N-Gram*, TF-IDF, Klasifikasi, *Random Forest*, *Support Vector Machine*

ABSTRACT

COMPARISON OF N-GRAM BASED FEATURE EXTRACTION WITH TF-IDF WEIGHTING AND RANDOM FOREST AND SVM CLASSIFICATION ALGORITHMS FOR DETECTION OF COVID-19 SYMPTOM REPORTING FROM TWITTER (By : Almadea Rusellawati; Pembimbing : Mohammad Reza Faisal, S.T, M.T, Ph.D dan Muhammad Itqan Mazdadi, M.Kom.; 2023; 63 pages)

The Covid-19 virus is still very much causing symptoms to the community. Although not as severe as before, Covid-19 has not disappeared. Currently, people are very openly telling about themselves who are affected or have symptoms of Covid-19 on their various social media, one of which is Twitter. The dataset used for this research is data on Covid-19 symptoms from previous research by Sari (2022). TF-IDF weighting is used to calculate the number of occurrences of terms in the data, testing again using *N-Gram* as a word separation feature. Where the *N-Gram* is divided into 3 namely *unigram*, *bigram* and *trigram*. This research uses *Random Forest* and *Support Vector Machine* classification which tests the *Linear*, *Radial Basic Function*, and *Polynomial* kernels. The highest accuracy performance is obtained from testing *Unigram* with *Support Vector Machine Linear* kernel of 88.8%.

Keywords : Feature Extraction, Covid-19 Symptoms, N-Gram, TF-IDF, Classification, Random Forest, Support Vector Machine

PRAKATA

Assalamualaikum Warahmatullahi Wabarakatuh.

Puji syukur penulis panjatkan kepada Allah SWT karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul **Perbandingan Ekstraksi Fitur Berbasis *N-Gram* dengan Pembobotan TF-IDF pada Algoritma Klasifikasi *Random Forest* dan SVM untuk Deteksi Pelaporan Gejala Covid-19 dari Twitter** untuk memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat. Tak lupa pula penulis panjatkan sholawat dan salam ke hadirat Rasulullah Muhammad SAW beserta para sahabat, keluarga, dan pengikut beliau hingga yaumul qiama.

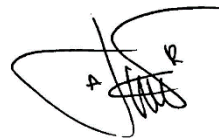
Pada lembar ini penulis ingin menyampaikan ucapan terimakasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, diantaranya:

1. Diri saya sendiri yang tidak pernah patah semangat walaupun banyak menemui kesulitan baik disebabkan oleh diri sendiri maupun dari hal lain.
2. Keluarga terutama kedua orang tua saya bapak Rusmawardi dan Ibu Suswati, dan juga adik-adik saya Almaulida Azzahra dan M. Raffa Almadani yang selalu memberikan dukungan, doa, dan bantuan dalam proses penyelesaian skripsi ini.
3. Bapak Mohammad Reza Faisal, S.T., M.T., Ph.D selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Muhammad Itqan Mazdadi, M.Kom. selaku dosen pembimbing pendamping dan Koordinator Program Studi Ilmu Komputer FMIPA ULM yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
5. Ibu Dwi Kartini, S.Kom, M.Kom dan bapak Triando Hamonangan Saragih, S.Kom., M.Kom. selaku dosen penguji yang banyak memberi saran dan meluangkan waktu untuk penyelesaian skripsi ini.
6. Seluruh Dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.

7. Teman-teman saya yang telah bersedia membantu memecahkan kebingungan-kebingungan dan memberikan saran, dukungan, semangat, serta motivasi, selama proses penyelesaian skripsi.
8. Teman-teman keluarga Ilmu Komputer angkatan 2016 yang memberikan dukungan dan bantuan semasa kuliah dan selama proses penyelesaian skripsi.
9. 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 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, 23 Juni 2023

A handwritten signature in black ink, appearing to be 'Almadea Rusellawati', with a stylized flourish above it.

Almadea Rusellawati

DAFTAR ISI

LEMBAR JUDUL	i
LEMBAR PENGESAHAN	ii
PERYATAAN	iii
ABSTRAK	iv
ABSTRACT	v
PRAKATA	vi
DAFTAR ISI	viii
DAFTAR TABEL	x
DAFTAR GAMBAR	xiii
DAFTAR LAMPIRAN	xiii
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan.....	2
1.4 Manfaat Penelitian.....	3
1.5 Batasan Masalah.....	3
BAB II TINJAUAN PUSTAKA	4
2.1 Kajian Terdahulu	4
2.2 Keaslian Penelitan	5
2.3 Twitter	8
2.4 Covid-19.....	9
2.5 <i>Text Mining</i>	10
2.6 <i>Text Pre-Processing</i>	11
2.7 Ekstraksi Fitur	13
2.8 <i>N-Gram</i>	14
2.9 TF-IDF.....	17
2.10 <i>Random Forest</i>	18
2.11 <i>Support Vector Machine</i>	23

2.12	<i>Cross Validation</i>	25
2.13	Confusion Matrix	26
BAB III METODE PENELITIAN		29
3.1	Alat Penelitian	29
3.2	Bahan Penelitian.....	29
3.3	Prosedur Penelitian.....	29
BAB IV HASIL DAN PEMBAHASAN		33
4.1	Hasil.....	33
4.1.1	Pengumpulan Data	33
4.1.2	Preprocessing Data.....	34
4.1.3	Ekstraksi Fitur	40
4.1.4	Klasifikasi	44
4.2	Pembahasan	55
BAB V PENUTUP.....		60
5.1	Kesimpulan.....	60
5.2	Saran.....	60
DAFTAR PUSTAKA		61
LAMPIRAN.....		64
RIWAYAT HIDUP		69

DAFTAR TABEL

Tabel 1. Keaslian Penelitian.....	6
Tabel 2. Perancangan Penelitian	7
Tabel 3. Membentuk <i>Bigram</i>	16
Tabel 4. Confusion Matrix	27
Tabel 5. Data Gejala Covid-19.....	30
Tabel 6. Contoh data tweet gejala covid-19.....	33
Tabel 7. Jumlah data tweet gejala Covid-19	34
Tabel 8. Penerapan cleansing.....	34
Tabel 9. Penerapan <i>case folding</i>	35
Tabel 10. Kamus <i>Slangword</i> (Pahrul, 2018).....	36
Tabel 11. Penerapan konversi <i>slang word</i>	37
Tabel 12. Kamus <i>Stopword</i> (Pahrul, 2018).....	38
Tabel 13. Kamus <i>Stopword</i> (Tala & Wibisono, 2016).....	38
Tabel 14. Penerapan <i>stopword removal</i>	39
Tabel 15. Penerapan <i>tokenizing</i>	40
Tabel 16. Proses <i>N-Gram Unigram</i>	41
Tabel 17. Proses <i>N-Gram Bigram</i>	41
Tabel 18. Proses <i>N-Gram Trigram</i>	42
Tabel 19. Proses TF-IDF <i>Unigram</i>	42
Tabel 20. Proses TF-IDF <i>Bigram</i>	43
Tabel 21. Proses TF-IDF <i>Trigram</i>	43
Tabel 22. <i>10 K-fold cross validation</i>	44
Tabel 23. <i>Confusion Matrix Random Forest</i> pada <i>Unigram</i>	46
Tabel 24. <i>Confusion Matrix Random Forest</i> pada <i>Bigram</i>	46
Tabel 25. <i>Confusion Matrix Random Forest</i> pada <i>Trigram</i>	46
Tabel 26. Perhitungan Akurasi <i>Random Forest</i>	46
Tabel 27. Contoh perhitungan <i>Support Vector Machine</i>	47
Tabel 28. <i>Confusion Matrix SVM Linear</i> pada <i>Unigram</i>	51
Tabel 29. <i>Confusion Matrix SVM Linear</i> pada <i>Bigram</i>	51

Tabel 30. <i>Confusion Matrix SVM Linear</i> pada <i>Trigram</i>	51
Tabel 31. Perhitungan Akurasi SVM <i>Linear</i>	52
Tabel 32. <i>Confusion Matrix SVM RBF</i> pada <i>Unigram</i>	52
Tabel 33. <i>Confusion Matrix SVM RBF</i> pada <i>Bigram</i>	52
Tabel 34. <i>Confusion Matrix SVM RBF</i> pada <i>Trigram</i>	53
Tabel 35. Perhitungan Akurasi SVM RBF	53
Tabel 36. <i>Confusion Matrix SVM Polynomial</i> pada <i>Unigram</i>	54
Tabel 37. <i>Confusion Matrix SVM Polynomial</i> pada <i>Bigram</i>	54
Tabel 38. <i>Confusion Matrix SVM Polynomial</i> pada <i>Trigram</i>	54
Tabel 39. Perhitungan Akurasi SVM <i>Polynomial</i>	54

DAFTAR GAMBAR

Gambar 1. Tahapan Proses Text Mining.....	10
Gambar 2. <i>Random Forest</i>	21
Gambar 3. Hyperplane yang memisahkan dua kelas positif (+1) dan negatif(-1)	23
Gambar 4. Ilustrasi <i>Cross validation</i>	26
Gambar 5. Alur penelitian.....	30
Gambar 6. Grafik kinerja <i>Random Forest</i> dan <i>N-Gram</i> TF-IDF	57
Gambar 7. Grafik kinerja SVM <i>Linear</i> dan <i>N-Gram</i> TF-IDF	57
Gambar 8. Grafik kinerja SVM RBF dan <i>N-Gram</i> TF-IDF	58
Gambar 9. Grafik kinerja SVM <i>Polynomial</i> dan <i>N-Gram</i> TF-IDF.....	58
Gambar 10. Perbandingan hasil akurasi.....	59

DAFTAR LAMPIRAN

Lampiran 1. Source code cleansing dan case folding	64
Lampiran 2. Source code slangword.....	65
Lampiran 3. Source code stopwords.....	65
Lampiran 4. Source code Function Ekstraksi Fitur N-Gram dan TF-IDF	66
Lampiran 5. Source code ekstraksi fitur <i>N-Gram</i>	66
Lampiran 6. Source code ekstraksi fitur TF-IDF	66
Lampiran 7. Source code klasifikasi <i>Random Forest Unigram</i>	67
Lampiran 8. Source code klasifikasi <i>Support Vector Machine Linear Unigram</i>	67
Lampiran 9. Source code klasifikasi <i>Support Vector Machine RBF Unigram</i>	68
Lampiran 10. Source code klasifikasi <i>Support Vector Machine Polyomial Unigram</i>	68