



**PENERAPAN *LEXICON BASED-NAÏVE BAYES CLASSIFIER* UNTUK  
ANALISIS SENTIMEN PUBLIK TERHADAP KESEHATAN  
MENTAL GENERASI Z**

**SKRIPSI**

**Untuk memenuhi persyaratan  
dalam menyelesaikan program sarjana Strata-1 Statistika**

**Oleh  
AIDA SAFITRI  
NIM. 2111017220004**

**PROGRAM STUDI S-1 STATISTIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMBUNG MANGKURAT  
BANJARBARU  
JUNI 2025**



**PENERAPAN *LEXICON BASED-NAÏVE BAYES CLASSIFIER* UNTUK  
ANALISIS SENTIMEN PUBLIK TERHADAP KESEHATAN  
MENTAL GENERASI Z**

**SKRIPSI**

**Untuk memenuhi persyaratan  
dalam menyelesaikan program sarjana Strata-1 Statistika**

**Oleh  
AIDA SAFITRI  
NIM. 2111017220004**

**PROGRAM STUDI S-1 STATISTIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS LAMBUNG MANGKURAT  
BANJARBARU  
JUNI 2025**

**SKRIPSI**

**PENERAPAN LEXICON BASED-NAIVE BAYES CLASSIFIER UNTUK  
ANALISIS SENTIMEN PUBLIK TERHADAP KESEHATAN MENTAL  
GENERASI Z**

Oleh:  
**Aida Safitri**  
**2111017220004**

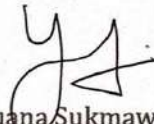
Telah dipertahankan pada hari Kamis, tanggal 26 Juni 2025 dan disetujui oleh dosen pembimbing dan dosen penguji sebagai berikut:

**Pembimbing I**



Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D  
NIP. 198303282005012001

**Penguji I**



Yuana Sukmawaty, S.Si., M.Si  
NIP. 198810152015042002

**Pembimbing II**



Oni Soesanto, S.Si., M.Si  
NIP. 197301262005011003

**Penguji II**



Selvi Annisa, S.Si., M.Si  
NIP. 199212262022032016

Banjarbaru, 2 Juli 2025

Mengetahui,  
Koordinator Program Studi Statistika  
MIPA ULM



Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D  
NIP. 198303282005012001

## 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, 23 Juni 2025



Aida Safitri

NIM. 2111017220004

## ABSTRAK

**Penerapan *Lexicon Based-Naive Bayes Classifier* untuk Analisis Sentimen Publik terhadap Kesehatan Mental Generasi Z** (Oleh: Aida Safitri; Pembimbing: Dewi Anggraini dan Oni Soesanto, 2025; 95 halaman)

Generasi Z adalah generasi yang lahir antara tahun 1997-2012 dan tumbuh dalam era digitalisasi sehingga teknologi menjadi salah satu bagian penting dari Generasi Z. Penggunaan teknologi yang berlebihan dapat menimbulkan dampak negatif, seperti kecanduan yang berpotensi menyebabkan ketidakstabilan emosi serta risiko masalah kesehatan mental, termasuk depresi dan kecemasan. Masalah kesehatan mental Generasi Z saat ini menjadi isu cukup serius dan banyak dibicarakan di media sosial X. Opini publik mengenai isu ini penting untuk dianalisis karena mencerminkan persepsi, pemahaman, dan sikap masyarakat terhadap kesehatan mental. Penelitian ini bertujuan untuk menganalisis sentimen pengguna media sosial X terhadap kesehatan mental Generasi Z menggunakan metode *Lexicon Based*, serta membandingkan kinerja model *Bernoulli* dan *Multinomial Naïve Bayes* dalam mengklasifikasikan sentimen. Hasil penelitian menunjukkan 80.6% pengguna media sosial X cenderung memberikan sentimen negatif, menggambarkan kekhawatiran publik terhadap kondisi kesehatan mental Generasi Z yang dinilai kurang sehat dan berdampak pada performa dalam kehidupan sehari-hari. Dari perbandingan kinerja, model *Multinomial Naive Bayes* menunjukkan kinerja lebih baik dibandingkan model *Bernoulli Naive Bayes* dengan sensitivitas sebesar 0.64318442, spesifisitas sebesar 0.81078431, dan *G-Mean* sebesar 0.71558742. Hasil penelitian ini, diharapkan membantu lembaga pendidikan dan perusahaan dalam merumuskan kebijakan serta program yang efektif, sekaligus meningkatkan kesadaran Generasi Z akan pentingnya menjaga kesehatan mental.

Kata Kunci: Kesehatan Mental, Generasi Z, Analisis Sentimen, *Lexicon Based*, *Naive Bayes Classifier*

## ABSTRACT

**Implementation of Lexicon Based-Naive Bayes Classifier for Public Sentiment Analysis on Generation Z's Mental Health** (By: Aida Safitri; Advisors: Dewi Anggraini and Oni Soesanto, 2023; 95 page)

Generation Z is a generation born between 1997-2012 and grew up in an era of digitalization making technology an important part of Generation Z. Excessive use of technology has the potential to cause negative impacts, such as addiction that causes emotional instability and the risk of mental health problems, including depression and anxiety. Generation Z's mental health problems are currently a serious issue and are widely discussed on social media X. Public opinion on this issue is important to analyze as it reflects the public's perception, understanding, and attitude toward mental health. This research aims to analyze the sentiment of users on social media X towards Generation Z's mental health using a Lexicon Based method and comparing the performance of Bernoulli and Multinomial Naïve Bayes models in classifying sentiment. The results show that 80.6% of users on social media X tend to give negative sentiments, showing the public's concern about the mental health condition of Generation Z, which is considered unhealthy and has an impact on performance in daily life. From the performance comparison, the Multinomial Naive Bayes model shows better performance than the Bernoulli Naive Bayes model with a sensitivity of 0.64318442, specificity of 0.81078431, and G-Mean of 0.71558742. The results of this research are expected to help educational institutions and companies formulate effective policies and programs, while increasing Generation Z's awareness of the importance of maintaining mental health.

Keywords: Mental Health, Generation Z, Sentiment Analysis, Lexicon Based, Naive Bayes Classifier

## PRAKATA

Puji dan syukur penulis panjatkan ke hadirat Allah SWT yang telah memberikan rahmat, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul “Penerapan *Lexicon Based-Naive Bayes Classifier* untuk Analisis Sentimen Publik terhadap Kesehatan Mental Generasi Z”. Penyusunan Tugas Akhir ini bertujuan untuk memenuhi salah satu syarat dalam menyelesaikan program sarjana S-1 di Program Studi Statistika FMIPA ULM.

Dalam proses penyusunan Tugas Akhir ini, penulis mendapatkan berbagai bantuan, bimbingan, dukungan, dan perhatian dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Orang tua dan keluarga yang selalu memberikan doa, dukungan, dan motivasi yang tiada henti selama masa perkuliahan hingga penyelesaian Tugas Akhir.
2. Ibu Prof. Dewi Anggraini, S.Si., M.App.Sci., Ph.D dan Bapak Oni Soesanto, S.Si., M.Si selaku dosen pembimbing yang telah memberikan bimbingan, arahan, dan dukungan selama proses penyusunan Tugas Akhir.
3. Ibu Yuana Sukmawaty, S.Si., M.Si dan Ibu Selvi Annisa, S.Si., M.Si selaku dosen penguji yang telah memberikan masukan dan saran selama proses penyusunan Tugas Akhir.
4. Bapak/Ibu dosen pengajar dan staf Program Studi Statistika FMIPA ULM yang telah memberikan ilmu, motivasi, dan dukungan selama masa perkuliahan.
5. Sahabat dan teman-teman yang selalu menemani, mendukung, memberikan semangat, serta mendengarkan keluh kesah penulis selama masa perkuliahan hingga penyusunan Tugas Akhir, khususnya Rizky Saputra B, Cindy Ayudia Irwanto, “Sobat Bendungan”, dan “Penjoki Handal”.
6. Teman-teman Program Studi Statistika angkatan 2021 yang telah berjuang bersama selama masa perkuliahan.

Penulis menyadari bahwa Tugas Akhir ini masih jauh dari sempurna dan masih terdapat kekurangan baik dari segi penulisan maupun hasil, sehingga kritik dan saran yang membangun dari semua pihak sangat diharapkan dalam membantu untuk penyempurnaan Tugas Akhir ini. Akhir kata, penulis berharap Tugas Akhir ini dapat memberikan manfaat untuk semua pihak.

Banjarbaru, 23 Juni 2025

Aida Safitri

## DAFTAR ISI

PERNYATAAN .....	ii
ABSTRAK.....	iii
ABSTRACT .....	iv
PRAKATA.....	v
DAFTAR ISI.....	vi
DAFTAR GAMBAR.....	viii
DAFTAR TABEL.....	ix
DAFTAR LAMPIRAN.....	x
DAFTAR ISTILAH, LAMBANG, DAN SINGKATAN.....	xi
BAB I PENDAHULUAN .....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Tujuan Penelitian.....	4
1.4 Manfaat Penelitian.....	4
BAB II TINJAUAN PUSTAKA.....	5
2.1 Kajian Penelitian Terdahulu.....	5
2.2 Kajian Teori.....	7
2.2.1 Kesehatan Mental.....	7
2.2.2 Generasi Z.....	8
2.2.3 Analisis Sentimen.....	9
2.2.4 Data <i>Crawling</i> .....	10
2.2.5 <i>Text Pre-Processing</i> .....	10
2.2.6 <i>Naïve Bayes Classifier</i> .....	12
2.2.7 <i>Bernoulli Naive Bayes</i> .....	13
2.2.8 <i>Multinomial Naïve Bayes</i> .....	14
2.2.9 <i>Lexicon Based</i> .....	16
2.2.10 <i>Term Frequency-Inverse Document Frequency (TF-IDF)</i> .....	17
2.2.11 <i>Synthetic Minority Oversampling Technique (SMOTE)</i> .....	18
2.2.12 Evaluasi Model.....	20
BAB III METODE PENELITIAN.....	23
3.1 Sumber Data.....	23
3.2 Variabel Penelitian.....	23
3.3 Prosedur Penelitian.....	23
BAB IV HASIL DAN PEMBAHASAN .....	28
4.1 Pengumpulan Data.....	28
4.2 <i>Text Pre-Processing</i> .....	28
4.2.1 <i>Data Cleaning</i> .....	28
4.2.2 <i>Case Folding</i> .....	29
4.2.3 <i>Tokenization</i> .....	30
4.2.4 <i>Normalization</i> .....	31
4.2.5 <i>Stopword Removal</i> .....	32
4.2.6 <i>Stemming</i> .....	32
4.3 Penerapan <i>Lexicon Based</i> .....	33
4.4 Eksplorasi Data.....	34
4.4.1 Analisis Deskriptif.....	34
4.4.2 <i>Word Cloud</i> Sentimen Negatif.....	36
4.4.3 <i>Word Cloud</i> Sentimen Positif.....	37
4.4.4 <i>Word Cloud</i> Sentimen Netral.....	37

4.5	Pembagian Data <i>Training</i> dan Data <i>Testing</i> .....	38
4.6	Ekstraksi Fitur TF-IDF.....	39
4.7	Teknik <i>Resampling</i> SMOTE.....	42
4.8	Klasifikasi Model <i>Bernoulli Naive Bayes</i> .....	44
4.9	Klasifikasi Model <i>Multinomial Naive Bayes</i> .....	47
4.10	Perbandingan Kinerja Model Klasifikasi .....	51
BAB V PENUTUP .....		53
5.1	Kesimpulan.....	53
5.2	Saran.....	53
DAFTAR PUSTAKA.....		55
LAMPIRAN .....		60
RIWAYAT HIDUP.....		95

PRODI STATISTIKA

## DAFTAR GAMBAR

Gambar 1. 1 Gangguan Mental Remaja di Indonesia .....	2
Gambar 2. 1 Ilustrasi Cara Kerja SMOTE .....	19
Gambar 3. 1 Prosedur Penelitian .....	27
Gambar 4. 1 Proporsi Kategori Sentimen .....	35
Gambar 4. 2 <i>Word Cloud</i> Sentimen Negatif.....	36
Gambar 4. 3 <i>Word Cloud</i> Sentimen Positif .....	37
Gambar 4. 4 <i>Word Cloud</i> Sentimen Netral .....	38

PRODI STATISTIKA

## DAFTAR TABEL

Tabel 2. 1 Kajian Penelitian Terdahulu.....	5
Tabel 2. 2 Contoh Penentuan Kategori .....	17
Tabel 2. 3 <i>Multiclass Confusion Matrix 3x3</i> .....	20
Tabel 3. 1 Variabel Penelitian.....	23
Tabel 4. 1 Contoh Data <i>Cleaning</i> .....	29
Tabel 4. 2 Contoh <i>Case Folding</i> .....	30
Tabel 4. 3 Contoh <i>Tokenization</i> .....	30
Tabel 4. 4 Contoh <i>Normalization</i> .....	31
Tabel 4. 5 Contoh <i>Stopword Removal</i> .....	32
Tabel 4. 6 Contoh <i>Stemming</i> .....	33
Tabel 4. 7 Contoh Penentuan Kategori pada <i>Dataset</i> .....	33
Tabel 4. 8 Pembagian Data .....	38
Tabel 4. 9 Data Ketiga Dalam Data <i>Training</i> Skenario Pertama .....	39
Tabel 4. 10 Hasil TF-IDF Data Ketiga.....	40
Tabel 4. 11 Contoh Data Kelas Minoritas.....	42
Tabel 4. 12 Contoh Titik Data Perhitungan Jarak <i>Euclidean</i> .....	42
Tabel 4. 13 Contoh Data Sintesis .....	43
Tabel 4. 14 Proporsi Kelas Data <i>Training</i> Sebelum dan Sesudah <i>Resampling</i> .....	43
Tabel 4. 15 Kombinasi Parameter Terbaik <i>Bernoulli Naive Bayes</i> .....	44
Tabel 4. 16 <i>Confusion Matrix</i> Model Terbaik BNB .....	44
Tabel 4. 17 Kombinasi Parameter Terbaik <i>Multinomial Naive Bayes</i> .....	48
Tabel 4. 18 <i>Confusion Matrix</i> Model Terbaik MNB .....	48
Tabel 4. 19 Perbandingan Kinerja Model Klasifikasi .....	51

## DAFTAR LAMPIRAN

Lampiran 1. Proses <i>Crawling</i> Data.....	60
Lampiran 2. <i>Keywords</i> dan Jumlah Data Hasil <i>Crawling</i> .....	61
Lampiran 3. Data Hasil <i>Crawling</i> .....	62
Lampiran 4. Contoh Duplikat Data .....	63
Lampiran 5. Contoh Data Tidak Relevan .....	64
Lampiran 6. Data Penelitian .....	65
Lampiran 7. Kamus <i>Unique Colloquial Words</i> .....	66
Lampiran 8. Kamus Prakoso (2017).....	67
Lampiran 9. Kamus Tambahan Normalisasi.....	68
Lampiran 10. Kamus ID- <i>Stopwords</i> .....	69
Lampiran 11. Kamus Tambahan <i>Stopwords</i> .....	70
Lampiran 12. Kamus Positif InSet <i>Lexicon</i> .....	71
Lampiran 13. Kamus Negatif InSet <i>Lexicon</i> .....	72
Lampiran 14. Data Penelitian dengan Kategori Sentimen .....	73
Lampiran 15. Hasil TF-IDF untuk Skenario 1 (90% dan 10%).....	74
Lampiran 16. Hasil TF-IDF untuk Skenario 2 (80% dan 20%).....	75
Lampiran 17. Hasil TF-IDF untuk Skenario 3 (70% dan 30%).....	76
Lampiran 18. Kombinasi Parameter Model <i>Bernoulli Naive Bayes</i> .....	77
Lampiran 19. Evaluasi Kinerja Model <i>Bernoulli Naive Bayes</i> .....	78
Lampiran 20. Kombinasi Parameter Model <i>Multinomial Naive Bayes</i> .....	79
Lampiran 21. Evaluasi Kinerja Model <i>Multinomial Naive Bayes</i> .....	80
Lampiran 22. <i>Syntax Python</i> untuk <i>Install Library, Import Library, Import</i> Data, dan <i>Text Pre-processing</i> .....	81
Lampiran 23. <i>Syntax Python</i> untuk <i>Lexicon Based</i> dan Eksplorasi Data .....	84
Lampiran 24. <i>Syntax Python</i> untuk Pembagian Data <i>Training</i> dan Data <i>Testing</i> , TF-IDF Setiap Skenario, Penerapan SMOTE, dan Kombinasi Parameter .....	86
Lampiran 25. <i>Syntax Python</i> untuk Model <i>Bernoulli Naive Bayes</i> .....	89
Lampiran 26. <i>Syntax Python</i> untuk Model <i>Multinomial Naive Bayes</i> .....	92

## DAFTAR LAMBANG DAN SINGKATAN

$c$	Variabel target
$d$	Data dengan kelas yang belum diketahui
$P(c d)$	Peluang dari $c$ dengan $d$ diketahui, disebut juga <i>posterior</i>
$P(d c)$	Peluang dari $d$ dengan $c$ diketahui, disebut juga <i>likelihood</i>
$P(c)$	Peluang $c$ , disebut juga <i>prior</i>
$P(d)$	Peluang $d$ , disebut juga <i>evidence</i>
$t$	<i>Term</i> (kata)
$t_k$	<i>Term</i> ke- $k$
$V$	Himpunan kata dalam data <i>training</i>
$n_d$	Jumlah total kata pada data $d$
$e_i$	Fitur ke- $i$
$M$	Jumlah total fitur
$P(t_k c)$	Peluang dari kata ( <i>term</i> ) $t_k$ terdapat dalam data kelas $c$
$N_c$	Jumlah data dalam kelas $c$ pada data <i>training</i>
$N$	Jumlah total data pada data <i>training</i>
$T_{ct}$	Jumlah kemunculan kata $t$ dalam data <i>training</i> pada kelas $c$
$B$	Jumlah kata ( <i>term</i> ) unik pada semua kelas
$M$	Banyaknya fitur
$P(t_i c)$	Peluang dari kata ( <i>term</i> ) $t_i$ terdapat dalam data kelas $c$
$N_{ct}$	Banyaknya data yang memuat <i>term</i> $t$ dalam data kelas $c$ pada data <i>training</i>
$W_{t,d}$	Bobot kata atau <i>term</i> ( $t$ ) terhadap data ( $d$ )
$tf_{t,d}$	Jumlah kemunculan kata atau <i>term</i> ( $t$ ) dalam data ( $d$ )
$D$	Jumlah total data
$df_t$	Jumlah data yang mengandung kata ( $t$ )
$\propto$	Proporsional
NBC	<i>Naïve Bayes Classifier</i>
BNB	<i>Bernoulli Naïve Baye</i>
MNB	<i>Multinomial Naïve Bayes</i>
NLP	<i>Natural language Processing</i>
ADHD	<i>Attention Deficit Hyperactivity Disorder</i>
PTSD	<i>Post-Traumatic Stress Disorder</i>
API	<i>Application Programming Interface</i>
URL	<i>Uniform Resource Locator</i>
TP	<i>True Positive</i>
TN	<i>True Negative</i>
FP	<i>False Positive</i>
FN	<i>False Negative</i>