SKRIPSI

OPTIMISASI REGRESI BINOMIAL NEGATIF LASSO BERBASIS AIC PADA FITUR TEKSTUAL PROMOSI DI PLATFORM X (STUDI KASUS: @SHOPEEID)

DEA ISWARI
22106010078

PROGRAM STUDI MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
YOGYAKARTA

2026

OPTIMISASI REGRESI BINOMIAL NEGATIF LASSO BERBASIS AIC PADA FITUR TEKSTUAL PROMOSI DI PLATFORM X (STUDI KASUS: @SHOPEEID)

Skripsi

Untuk memenuhi sebagian persyaratan

mencapai derajat Sarjana S-1

Program Studi Matematika

diajukan oleh

DEA ISWARI

22106010078

Kepada

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA

YOGYAKARTA

2026

SURAT PERSETUJUAN SKRIPSI/TUGAS AKHIR

Hal : Persetujuan Skripsi / Tugas Akhir
Lamp :

Kepada
Yth. Dekan Fakultas Sains dan Teknologi
UIN Sunan Kalijaga Yogyakarta
di Yogyakarta

Assalamu'alaikum wr. wb.

Setelah membaca, meneliti, memberikan petunjuk dan mengoreksi serta mengadakan perbaikan seperlunya, maka kami selaku pembimbing berpendapat bahwa skripsi Saudara:

Nama	: Dea iswari
NIM	: 22106010078
Judul Skripsi	: Optimisasi Regresi Binomial Negatif LASSO Berbasis AIC pada Fitur Tekstual Promosi di Platform X (Studi Kasus: @ShopeeID)

sudah dapat diajukan kembali kepada Program Studi Matematika Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta sebagai salah satu syarat untuk memperoleh gelar Sarjana Strata Satu dalam Program Studi Matematika.

Dengan ini kami mengharap agar skripsi/tugas akhir Saudara tersebut di atas dapat segera dimunaqasyahkan. Atas perhatiannya kami ucapkan terima kasih.

Wassalamu'alaikum wr. wb.

Yogyakarta, 22 Mei 2026
Pembimbing

Sri Utami Zuliana
19741003 200003 2 002

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN KALIJAGA
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Marsda Adisucipto Telp. (0274) 540971 Fax. (0274) 519739 Yogyakarta 55281

PENGESAHAN TUGAS AKHIR

Nomor : B-1167/Un.02/DST/PP.00.9/06/2026

Tugas Akhir dengan judul : Optimisasi Regresi Binomial Negatif LASSO Berbasis AIC pada Fitur Tekstual Promosi di Platform X (Studi Kasus: @ShopeeID)

yang dipersiapkan dan disusun oleh:

Nama : DEA ISWARI
Nomor Induk Mahasiswa : 22106010078
Telah diujikan pada : Selasa, 02 Juni 2026
Nilai ujian Tugas Akhir : A

dinyatakan telah diterima oleh Fakultas Sains dan Teknologi UIN Sunan Kalijaga Yogyakarta

TIM UJIAN TUGAS AKHIR

Ketua Sidang

Sri Utami Zuliana, S.Si., M.Sc., Ph.D.
SIGNED

Valid ID: 6a21032ef0643

Penguji I

Dr. Mohammad Farhan Qudratullah, S.Si., M.Si
SIGNED

Valid ID: 6a20d7b6d3cc8

Penguji II

Aulia Khifah Futhona, M.Sc.
SIGNED

Valid ID: 6a210067ca186

Yogyakarta, 02 Juni 2026
UIN Sunan Kalijaga
Dekan Fakultas Sains dan Teknologi

Prof. Dr. Dra. Hj. Khurul Wardati, M.Si.
SIGNED

Valid ID: 6a21306fe9e5d

SURAT PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini:

Nama	: Dea Iswari
NIM	: 22106010078
Program Studi	: Matematika
Fakultas	: Sains dan Teknologi

Dengan ini menyatakan bahwa isi skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar sarjana di suatu Perguruan Tinggi dan sesungguhnya skripsi ini merupakan hasil pekerjaan penulis sendiri sepanjang pengetahuan penulis, bukan duplikasi atau saduran dari karya orang lain kecuali bagian tertentu yang penulis ambil sebagai bahan acuan. Apabila terbukti pernyataan ini tidak benar, sepenuhnya menjadi tanggung jawab penulis.

Yogyakarta, 23 Mei 2026

Dea Iswari

HALAMAN PERSEMBAHAN

SETIAP KATA DALAM SKRIPSI INI MUNGKIN AKAN USAI, NAMUN TIDAK UNTUK RASA TERIMA KASIH PADA LEMBAR PERSEMBAHAN,

KARYA INI SAYA PERSEMBAHKAN UNTUK DIRI SENDIRI YANG TELAH BERTAHAN DAN TERUS MELANGKAH, SERTA SEBAGAI TANDA CINTA DAN TERIMA KASIH YANG MENDALAM UNTUK ORANG TUA, KAKAK, DAN SIMBAH, ATAS LIMPAHAN DOA, DUKUNGAN, SERTA KASIH SAYANG YANG TAK TERHINGGA,

UNTUK KELUARGA BESAR, SAHABAT, DAN TEMAN–TEMAN YANG SELALU HADIR MEWARNAI HIDUP SERTA SEMUA ORANG BAIK YANG TELAH MENGIRINGI SELURUH PROSES INI,

TAK LUPA, PERSEMBAHAN INI JUGA UNTUK ALMAMATER TERCINTA **UNIVERSITAS ISLAM NEGERI SUNAN KALIIJAGA YOGYAKARTA** YANG TELAH MENJADI RUANG TUMBUH DAN BELAJAR SELAMA 4 TAHUN MASA STUDI.

MOTTO

"Sesungguhnya sesudah kesulitan itu ada kemudahan. Sesungguhnya sesudah kesulitan itu ada kemudahan."

(QS. Al-Insyirah: 5-6)

"Tidak ada yang mustahil bagi mereka yang mau mencoba."

(Alexander The Great)

"Satu-satunya hal yang paling menakutkan adalah rasa takut itu sendiri, maka menjadi berani adalah pilihan terbaik untuk tumbuh."

(Anonym)

"Man Jadda Wajada"

(Barangsiapa bersungguh-sungguh, maka dia akan berhasil)

Trust yourself, you know the way.

(2226)

PRAKATA

Alhamdulillahirabbil'alamin, puji syukur kehadirat Allah SWT yang telah melimpahkan rahmat dan karunianya yang tak ternilai harganya berupa keimanan, kesabaran, kekuatan, dan kelancaran. Shalawat serta salam semoga selalu tercurahkan kepada Nabi Muhammad SAW sehingga penulis dapat menyelesaikan skripsi ini yang berjudul "Optimisasi Regresi Binomial Negatif LASSO Berbasis AIC pada Fitur Tekstual Promosi di Platform X (Studi Kasus: @ShopeeID)"

Penulis menyadari bahwa penyusunan skripsi ini tidak terlepas dari dukungan, bantuan, dan doa dari berbagai pihak. Segala tantangan yang dihadapi selama proses penelitian menjadi lebih ringan berkat motivasi serta perhatian yang diberikan. Oleh karena itu, dengan penuh rasa hormat dan ketulusan, penulis menyampaikan rasa terima kasih dan penghargaan sebesar-besarnya kepada:

1. Bapak Prof. Noorhaidi, S.Ag., M.A., M.Phil., Ph.D., selaku Rektor UIN Sunan Kalijaga Yogyakarta.
2. Ibu Prof. Dr. Dra. Hj. Khurul Wardati, M.Si., selaku Dekan Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
3. Ibu Dr. Epha Diana Supandi, S.Si., M.Sc., selaku Ketua Program Studi Matematika Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Kalijaga Yogyakarta.
4. Ibu Sri Utami Zuliana, S.Si., M.Sc., Ph.D., selaku Dosen Pembimbing, terima kasih atas bimbingan, dorongan, serta ilmu yang senantiasa Ibu bagikan kepada penulis. Segala arahan Ibu tidak hanya membantu penyelesaian

skripsi ini, namun juga menjadi pelajaran berharga yang akan selalu penulis kenang. Penulis juga memohon maaf atas segala kesalahan dan kekeliruan yang penulis lakukan selama masa penyusunan skripsi ini.

5. Bapak Dr. Mohammad Farhan Qudratullah, S.Si., M.Si, selaku Dosen Penguji I, yang telah meluangkan waktu untuk menguji, memberikan masukan, serta saran yang bermanfaat bagi penyempurnaan skripsi ini.
6. Ibu Aulia Khifah Futhona, M.Sc., selaku Dosen Penguji II, yang telah memberikan kritik, saran, dan arahan yang membangun sehingga penulis dapat memperbaiki dan menyempurnakan skripsi ini.
7. Bapak Deddy Rahmadi, M.Sc., selaku Dosen Pembimbing Akademik yang telah banyak meluangkan waktu untuk membimbing dan mengarahkan penulis sejak awal perkuliahan hingga akhir masa studi. Terima kasih atas komunikasinya yang sangat baik. Semoga kebaikan dan ilmu yang Bapak bagikan menjadi amal jariyah yang tak terputus.
8. Seluruh dosen Program Studi Matematika dan staf Fakultas Sains dan Teknologi yang senantiasa memberikan ilmu dan layanan terbaik selama masa studi.
9. Kedua orang tua tercinta, sosok yang sangat penulis sayangi dan menjadi alasan utama penulis untuk terus berjuang. Untuk Bapak, terima kasih atas segala usaha, kepercayaan, dan dukungan yang diberikan. Terima kasih karena senantiasa mengupayakan yang terbaik serta percaya pada setiap langkah yang penulis pilih. Kepada Mamak, terima kasih atas kesabaran, ketulusan, serta kasih sayang dalam menemani, melindungi, serta menjadi tempat berbagi cerita. Tidak ada kata yang mampu menggambarkan betapa

berharganya kehadiran Bapak dan Mamak bagi penulis. Semoga Allah SWT senantiasa melimpahkan kesehatan, kebahagiaan, dan perlindungan kepada Bapak dan Mamak agar dapat terus membersamai perjalanan penulis di masa mendatang.

10. Teruntuk Mas Damar, sosok kakak yang selalu menjadi pendukung terbaik dalam setiap langkah penulis. Terima kasih karena selalu peduli, memfasilitasi, dan menyayangi penulis dengan tulus. Terima kasih selalu percaya dengan kemampuan penulis bahkan di saat penulis meragukan diri sendiri. Dukungan serta kehadiran Mas dan Mba Vian, kakak ipar, sangat berarti dalam perjalanan ini. Terima kasih untuk kebersamaan dan perhatian yang tidak pernah putus. Semoga Allah SWT senantiasa melimpahkan keberkahan serta rezeki dalam hidup Mas dan Mba.

11. Tiga manusia menyebalkan yang penulis sayangi, Dhara aka Ardha, Jessferellll aka Jaqueline, dan Náanana aka Nanda. Terima kasih telah menjadi sobat terbaik selama ini. Banyak cerita yang telah terukir, mulai dari tawa, kesal, marah, hingga tangis. Meski begitu, selalu ada dukungan dan rasa saling menguatkan yang membuat setiap proses terasa ringan. Terima kasih karena selalu ada, siap direpotkan, dan menjadi tempat menampung cerita. Doa penulis, semoga kebahagiaan, kesuksesan, dan perlindungan Allah SWT senantiasa menyertai setiap langkah.

12. Najih dan Nafal, rekan bertumbuh dan berproses yang telah mengenalkan penulis pada dunia pemrograman. Terima kasih atas berbagai perspektif baru, dorongan untuk terus konsisten dan berkembang, serta bantuan dan dukungan yang diberikan selama proses penyusunan skripsi ini. Terima kasih juga karena telah memfasilitasi penulis dalam melaksanakan simulasi presentasi

bersama teman-teman komunitas Pejuang Pemrograman. Bersama kalian, perjalanan ini terasa lebih penuh semangat dan optimis. Semoga kesuksesan senantiasa menyertai langkah kita.

13. Rekan satu bimbingan, terkhusus Robit dan Yasmin. Terima kasih atas dukungan, semangat, dan kebersamaan sepanjang penyusunan skripsi ini.
14. Ainun Salma Nurbaiti, sobat sedari bayi. Terima kasih telah selalu menguatkan, menjadi teman berkeluh, hingga berjuang bersama dalam menyelesaikan skripsi.
15. Mba Linggar, terima kasih karena selalu meluangkan waktu untuk menjawab setiap pertanyaan, memberikan koreksi, serta arahan yang sangat membantu.
16. Sobat Icil, Fiesna, Assa, Aini, dan Hurun. Terima kasih atas semangat, dukungan, dan kesediaan selalu ada untuk merayakan setiap proses.
17. Teman-teman angkatan 2022, terima kasih telah menjadi bagian dari perjalanan penulis selama di bangku perkuliahan. Terima kasih atas semangat dan kenangan indah yang telah kita lalui bersama.
18. Seluruh pihak yang telah membantu dalam penyusunan skripsi ini, baik secara langsung maupun tidak langsung. Terima kasih atas segala dukungan dan doa yang diberikan. Semoga kebaikan kalian senantiasa dibalas dengan keberkahan oleh Allah SWT.

Yogyakarta, 25 Mei 2026

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMBANG

Y	= variabel respons
y_i	= nilai observasi variabel respons pada data ke-i
$\hat{Y}$	= nilai prediksi variabel respons
$E(Y)$	= nilai harapan variabel respons
$Var(Y)$	= variansi dari variabel respons
$\mathbf{X}$	= variabel prediktor
x_{ij}	= nilai variabel prediktor (fitur) ke-j pada observasi ke-i
x'_{ij}	= nilai variabel prediktor setelah normalisasi
$TermFreq_{ij}$	= frekuensi kemunculan kata ke-j pada dokumen ke-i
$DocFreq_j$	= jumlah dokumen yang mengandung kata ke-j
β	= parameter atau koefisien regresi
β_0	= intersep atau konstanta model
$\hat{\beta}$	= nilai estimasi parameter regresi
μ	= rata-rata variabel respons
η	= prediktor linear
α	= parameter dispersi pada distribusi binomial negatif
λ	= parameter penalti
ϵ	= komponen galat (*error*)

$g(\cdot)$	= fungsi *link*
$g^{-1}(\cdot)$	= invers fungsi *link*
$L(\cdot)$	= fungsi *likelihood*
$\ell(\cdot)$	= fungsi *log-likelihood*
$\ln$	= logaritma natural
e	= bilangan euler ($\approx 2,718$)
$Q(\cdot)$	= fungsi objektif yang diminimalkan
$P_\lambda(\cdot)$	= fungsi penalti pada metode regularisasi
$R^2_{McFadden}$	= koefisien determinasi semu
ϕ	= rasio dispersi
χ^2	= Statistik uji *Pearson Chi-square*
$\Gamma(\cdot)$	= fungsi gamma
$\psi(\cdot)$	= fungsi digamma
$f(\cdot)$	= fungsi massa probabilitas
v_i	= peubah acak yang mengikuti distribusi gamma
$\mathbf{H}$	= matriks hessian (turunan parsial kedua)
$g(\theta)$	= vektor gradien
θ	= vektor parameter gabungan (β dan α)
n	= jumlah total observasi (dokumen)
p	= jumlah variabel prediktor (fitur kata)
p^*	= jumlah variabel prediktor terpilih
df_{resid}	= derajat bebas residu

i = indeks observasi

j = indeks variabel prediktor

k = jumlah parameter yang diestimasi

m = indeks iterasi algoritma Newton-Raphson

t = parameter tuning

∂ = operator turunan parsial

T = transpose matriks atau vektor

$\sum$ = operator penjumlahan beruntun

$\prod$ = operator perkalian beruntun

$!$ = faktorial

INTISARI

OPTIMISASI REGRESI BINOMIAL NEGATIF LASSO BERBASIS AIC PADA FITUR TEKSTUAL PROMOSI DI PLATFORM X (STUDI KASUS: @ShopeeID)

Dea Iswari

22106010078

Overdispersi pada *count data* menyebabkan pelanggaran asumsi equidispersi sehingga distribusi Poisson tidak lagi mampu memodelkan variabilitas data secara memadai. Regresi Binomial Negatif digunakan sebagai pendekatan alternatif melalui penambahan parameter dispersi untuk mengakomodasi variansi yang melebihi nilai rata-rata. Namun, pemodelan menjadi kompleks ketika melibatkan variabel prediktor berupa fitur tekstual hasil ekstraksi TF-IDF yang berpotensi menimbulkan redundansi dan multikolinearitas antarfitur. Untuk menghasilkan model yang parsimoni, diterapkan regularisasi LASSO dengan penalti L_1 yang mampu menyusutkan koefisien variabel dengan kontribusi kecil hingga tepat bernilai nol. Penentuan parameter penalti λ dilakukan melalui minimalisasi kriteria AIC guna memperoleh keseimbangan antara kompleksitas dan performa model. Implementasi metode ini pada fitur tekstual promosi akun @ShopeeID di platform X menunjukkan bahwa regresi Binomial Negatif LASSO berbasis AIC mampu mereduksi dimensi model dengan mengeliminasi 30 variabel redundan. Meskipun model menjadi lebih ringkas, kemampuan penjelasan model tetap relatif konsisten yang ditunjukkan oleh penurunan nilai McFadden's Pseudo R^2 yang sangat kecil, yaitu sebesar 0,0013 dibandingkan model penuh. Hasil penelitian menunjukkan bahwa integrasi regresi Binomial Negatif dan regularisasi LASSO efektif dalam meningkatkan interpretabilitas model tanpa mengorbankan kualitas statistik secara signifikan.

Kata Kunci: Binomial Negatif, Overdispersi, Regularisasi LASSO, AIC, TF-IDF

ABSTRACT

AIC-BASED NEGATIVE BINOMIAL LASSO REGRESSION OPTIMIZATION FOR TEXTUAL PROMOTIONAL FEATURES ON THE X PLATFORM (CASE STUDY: @ShopeeID)

Dea Iswari

22106010078

Overdispersion in count data leads to a violation of the assumption of equidispersion, meaning that the Poisson distribution is no longer capable of adequately modeling the variability of the data. Negative Binomial regression is used as an alternative approach by adding a dispersion parameter to accommodate variance that exceeds the mean. However, modeling becomes complex when it involves predictor variables in the form of textual features extracted via TF-IDF, which may lead to redundancy and multicollinearity among features. To produce a parsimonious model, LASSO regularization with an L_1 penalty is applied, which shrinks the coefficients of variables with small contributions to exactly zero. The penalty parameter λ is determined by minimizing the AIC criterion to achieve a balance between model complexity and performance. The implementation of this method on textual features from the @ShopeeID account's promotional posts on the X platform demonstrates that AIC-based Negative Binomial LASSO regression effectively reduces model dimensionality by eliminating 30 redundant variables. Although the model becomes more compact, its explanatory power remains relatively consistent, as indicated by a very small decrease in McFadden's Pseudo R^2, specifically, 0.0013 compared to the full model. The research results show that the integration of Negative Binomial regression and LASSO regularization is effective in improving model interpretability without significantly sacrificing statistical quality.

Keywords: Negative Binomial, Overdispersion, LASSO Regularization, AIC, TF-IDF

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Regresi merupakan metode dalam statistika yang banyak digunakan untuk memodelkan hubungan antara variabel respons dan satu atau lebih variabel penjelas (Montgomery et al., 2021). Melalui analisis regresi, pola keterkaitan antarvariabel dapat dipahami sekaligus digunakan untuk prediksi. Model regresi linear merupakan bentuk regresi yang paling umum digunakan, dengan asumsi variabel respons berdistribusi normal dan memiliki hubungan linear dengan variabel penjelas. Namun, asumsi tersebut tidak selalu terpenuhi, terutama ketika variabel respons berupa *count data*. *Count data* merupakan data yang menunjukkan jumlah kejadian suatu peristiwa dan bersifat diskrit, sehingga memerlukan pendekatan model khusus untuk mengakomodasi sifat distribusinya (Cameron & Trivedi, 2013).

Pada pemodelan *count data*, distribusi Poisson menjadi model dasar yang umum digunakan. Regresi Poisson menghubungkan rata-rata kejadian dengan variabel penjelas melalui fungsi penghubung (*link*) logaritmik. Secara konseptual, *count data* merepresentasikan jumlah kejadian dalam suatu interval waktu atau ruang, sehingga regresi Poisson lebih sesuai dibandingkan regresi linear karena mampu menghasilkan prediksi tetap berada pada rentang yang sesuai (Cameron & Trivedi, 2013).

Namun, model Poisson memiliki asumsi equidispersi, yaitu nilai rata-rata

sama dengan variansi. Dalam praktiknya, data sering menunjukan variansi yang lebih besar dari rata-rata (overdispersi) atau banyaknya nilai nol (Agresti, 2018). Jika diabaikan, kondisi ini dapat menyebabkan estimasi *standard error* menjadi bias dan kesimpulan menjadi tidak valid (Cameron & Trivedi, 2013). Oleh karena itu, regresi Binomial Negatif digunakan sebagai pengembangan dari model Poisson karena memiliki parameter dispersi tambahan yang mampu mengakomodasi variansi berlebih (Hilbe, 2011).

Selain permasalahan overdispersi pada variabel respons, tantangan lain juga muncul dari karakteristik variabel prediktor yang berasal dari data teks. Data teks tidak dapat langsung digunakan dalam model regresi, sehingga diperlukan transformasi melalui teknik *text mining* untuk mengubahnya menjadi representasi numerik (Rosa, 2010). Salah satu tahapannya adalah ekstraksi fitur, yaitu proses mengubah teks menjadi variabel numerik melalui metode pembobotan seperti TF-IDF, yang mengukur tingkat kepentingan suatu kata dalam dokumen.

Representasi TF-IDF menghasilkan fitur yang relatif banyak dan bersifat *sparse*, di mana tidak semua fitur relevan terhadap variabel respons (Rosa, 2010). Kondisi ini meningkatkan risiko *overfitting*, yaitu model terlalu menyesuaikan diri terhadap *noise* pada data latih akibat kompleksitas parameter (Hastie et al., 2015). Oleh karena itu, diperlukan metode regularisasi untuk mengontrol kompleksitas model dengan memberikan penalti pada koefisien regresi.

Salah satu teknik regularisasi adalah regresi *Ridge*, yang efektif mengatasi multikolinearitas namun tidak dapat menghasilkan koefisien nol. Sebagai pengembangannya, metode *Least Absolute Shrinkage and Selection Operator* (LASSO) menggunakan penalti L_1 yang memungkinkan koefisien menyusut hingga tepat nol, sehingga berfungsi sebagai mekanisme seleksi variabel otomatis

(Tibshirani, 1996). Hal ini menjadikan LASSO sesuai untuk data tekstual dengan jumlah fitur yang besar.

Efektivitas LASSO bergantung pada pemilihan parameter penalti (λ) yang optimal. Parameter ini mengontrol tingkat penyusutan koefisien dan menentukan keseimbangan antara akurasi model dan kompleksitasnya. Dalam penelitian ini, *Akaike Information Criterion* (AIC) digunakan untuk memilih model terbaik dengan mempertimbangkan kecocokan model dan jumlah parameter (Burnham & Anderson, 2002).

Perkembangan teknologi digital telah mengubah berbagai aspek aktivitas masyarakat, termasuk dalam bidang pemasaran dan komunikasi bisnis. Pada era digital saat ini, media sosial tidak hanya berfungsi sebagai sarana komunikasi personal, tetapi juga menjadi media bagi perusahaan untuk melakukan promosi secara cepat, interaktif, dan masif. Strategi pemasaran yang sebelumnya banyak dilakukan melalui media konvensional kini bergeser pada pemasaran digital karena dinilai lebih efektif dalam menjangkau konsumen secara langsung melalui platform *online*.

Salah satu platform media sosial yang banyak dimanfaatkan dalam aktivitas promosi adalah X. Karakteristik platform X yang berbasis teks singkat memungkinkan perusahaan menyampaikan informasi promosi secara ringkas, cepat, dan mudah disebarluaskan kepada pengguna. Selain itu, arus informasi yang berlangsung secara *real-time* menjadikan platform ini relevan digunakan sebagai media komunikasi pemasaran digital, khususnya bagi perusahaan *e-commerce* yang aktif berinteraksi dengan konsumen melalui konten promosi berbasis teks.

Pada konteks pemasaran digital, efektivitas suatu promosi umumnya diukur melalui tingkat keterlibatan pengguna (*user engagement*), seperti jumlah *likes*,

replies, dan *reposts*. Di antara berbagai bentuk interaksi tersebut, penelitian ini menetapkan jumlah *likes* sebagai variabel respons utama karena dianggap mampu merepresentasikan respons positif pengguna secara lebih spontan dan konsisten terhadap suatu konten promosi. Penelitian De Vries et al. (2012) menunjukkan bahwa jumlah *likes* merupakan indikator utama popularitas sebuah unggahan, di mana elemen konten secara signifikan memengaruhi minat pengguna untuk berinteraksi.

Selain itu, karakteristik konsumsi informasi pada media sosial yang serba cepat menyebabkan pengguna cenderung memberikan respons dengan beban kognitif yang rendah. Menurut Appel et al. (2020), *likes* merupakan bentuk respons instan yang paling mampu merepresentasikan impresi awal pengguna terhadap suatu konten. Hal ini menjadi relevan dalam konteks promosi pada platform X, di mana pengguna umumnya melakukan *skimming* terhadap aliran informasi yang sangat dinamis. Di sisi lain, penelitian Li & Xie (2020) menunjukkan bahwa pada konten promosi atau komersial, pengguna cenderung lebih aktif memberikan *likes* dibandingkan melakukan *reposts* atau *replies* karena interaksi tersebut dapat dilakukan tanpa keterlibatan sosial yang tinggi.

Berdasarkan fenomena tersebut, penelitian ini menggunakan data unggahan promosi dari akun resmi @ShopeeID pada platform X. Sebagai salah satu perusahaan *e-commerce* besar di Indonesia, @ShopeeID secara aktif memanfaatkan media sosial untuk menyampaikan berbagai kampanye promosi melalui teks-teks pemasaran yang dipublikasikan secara rutin. Namun, data tersebut menghadirkan tantangan analisis tersendiri karena jumlah *likes* sebagai variabel respons termasuk ke dalam *count data* yang cenderung mengalami overdispersi, sementara representasi fitur teks hasil TF-IDF menghasilkan ruang prediktor berdimensi tinggi dan bersifat

sparse.

Oleh karena itu, penelitian ini bertujuan untuk menerapkan regresi Binomial Negatif dengan regularisasi LASSO yang dioptimasi menggunakan *Akaike Information Criterion* (AIC) guna memperoleh model terbaik. Pendekatan ini dipilih untuk mengontrol kompleksitas parameter melalui seleksi fitur otomatis sehingga model tidak hanya akurat secara prediktif, tetapi juga mampu mengidentifikasi kata-kata promosi yang paling berpengaruh terhadap *engagement* pengguna.

1.2. Rumusan Masalah

1. Bagaimana pembentukan model regresi Binomial Negatif dengan regularisasi LASSO yang dioptimisasi menggunakan kriteria AIC?
2. Bagaimana penerapan model regresi Binomial Negatif dengan regularisasi LASSO pada data teks promosi akun @ShopeeID untuk menentukan nilai parameter penalti (λ) optimal serta dampaknya terhadap reduksi fitur tekstual?
3. Bagaimana interpretasi fitur tekstual yang signifikan dalam model optimal terhadap jumlah *likes* sebagai indikator *engagement*?

1.3. Batasan Masalah

Agar penelitian ini lebih terarah, maka ditetapkan batasan masalah sebagai berikut:

1. Data yang digunakan adalah *tweet* promosi akun @ShopeeID periode 1 Januari 2025 hingga 27 November 2025 di platform X, dengan variabel respons berupa jumlah *likes*.

2. Variabel prediktor merupakan fitur tekstual hasil ekstraksi menggunakan metode TF-IDF.

3. Penelitian dibatasi pada penggunaan model Regresi Binomial Negatif dengan regularisasi LASSO.

4. Pemilihan model terbaik dalam penelitian ini menggunakan kriteria AIC.

5. Analisis dilakukan menggunakan bahasa pemrograman `Python` versi `3.12.12` melalui platform Google Colab.

1.4. Tujuan Penelitian

1. Menganalisis pembentukan model regresi Binomial Negatif dengan regularisasi LASSO yang dioptimisasi menggunakan kriteria AIC.

2. Mengaplikasikan model tersebut pada data teks promosi akun @ShopeeID untuk menentukan nilai parameter penalti (λ) optimal serta menganalisis dampaknya terhadap reduksi fitur tekstual.

3. Menganalisis interpretasi fitur tekstual yang signifikan dalam model optimal terhadap jumlah *likes* sebagai indikator *engagement*.

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoretis maupun praktis. Secara teoretis, penelitian ini diharapkan dapat memperkaya kajian mengenai penerapan Regresi Binomial Negatif pada *count data*, khususnya dalam konteks analisis interaksi pengguna pada media sosial. Selain itu, penelitian ini juga menunjukkan penerapan regularisasi LASSO dalam melakukan penyusutan koefisien (*shrinkage*) dan seleksi fitur tekstual, serta penggunaan *Akaike*

Information Criterion (AIC) sebagai kriteria dalam pemilihan model yang optimal dan parsimoni.

Secara praktis, penelitian ini diharapkan dapat memberikan gambaran mengenai karakteristik fitur tekstual promosi yang berkaitan dengan jumlah *likes* pada akun X @ShopeeID. Hasil penelitian ini juga diharapkan dapat menjadi referensi bagi penelitian selanjutnya yang berkaitan dengan analisis data media sosial dan pemodelan *count data*.

1.6. Tinjauan Pustaka

Count data banyak dijumpai dalam analisis media sosial, khususnya pada variabel *engagement* seperti jumlah *likes*, komentar, atau *retweets*. Namun, karakteristik data tersebut umumnya tidak memenuhi asumsi equidispersi, sebagaimana disyaratkan pada distribusi Poisson. Dalam praktiknya, data *engagement* sering kali mengalami overdispersi, sehingga model Poisson menjadi kurang sesuai. Bhattacharya et al. (2017) membuktikan bahwa data *engagement* di Facebook cenderung mengalami overdispersi, sehingga model Binomial Negatif dengan seleksi berbasis *Akaike Information Criterion* (AIC) terbukti lebih akurat. Fenomena serupa juga ditemukan oleh Shen et al. (2016) pada data ulasan Yelp, di mana pendekatan regularisasi diperlukan untuk memperoleh model yang parsimoni dan tetap interpretatif pada *count data* yang *sparse*.

Tantangan utama muncul ketika data memiliki dimensi tinggi dan potensi multikolinearitas, terutama pada analisis teks. Lehman & Archer (2019) memperkenalkan algoritma *Negative Binomial Generalized Monotone Incremental Forward Stagewise* (NB-GMIFS) untuk mengestimasi model Binomial Negatif (NB) dengan penalti L_1 (LASSO) pada kondisi $P > N$. Meskipun dikembangkan

untuk data genomik, prinsip regularisasi ini sangat relevan untuk menangani karakteristik *sparsity* pada data teks. Efektivitas LASSO dalam meningkatkan performa prediksi dan eliminasi variabel tidak relevan pada model NB juga divalidasi oleh Liu & Pitt (2017) dalam bidang aktuaria, meskipun mereka cenderung menggunakan *cross-validation* dibandingkan kriteria informasi seperti AIC.

Di sisi lain, analisis teks umumnya melibatkan representasi *Bag-of-Words* (BoW) yang menghasilkan banyak fitur dengan frekuensi kemunculan rendah. Cichosz (2023) menekankan bahwa representasi tersebut menghasilkan matriks desain yang didominasi nilai nol, sehingga berpotensi menimbulkan ketidakstabilan estimasi apabila tidak disertai mekanisme seleksi fitur. Meskipun metode *machine learning* seperti *Random Forest* dan *Support Vector Machines* (SVM) mampu menghasilkan akurasi prediksi yang tinggi, model tersebut cenderung kurang interpretatif dalam mengukur kontribusi masing-masing fitur tekstual. Oleh karena itu, penggunaan bobot TF-IDF yang dipadukan dengan regularisasi menjadi krusial untuk menangani permasalahan *sparsity* tersebut.

Pada konteks penelitian *engagement e-commerce* di Indonesia, studi oleh Andariesta & Wasesa (2023) serta Madnure & Kada (2025) lebih berfokus pada peningkatan akurasi prediksi menggunakan pendekatan *machine learning* seperti *Random Forest* dan *Gradient Boosting*. Namun, penelitian tersebut belum secara eksplisit mempertimbangkan karakteristik overdispersi pada *count data* maupun kebutuhan interpretasi koefisien regresi dalam mengukur pengaruh masing-masing kata promosi.

Oleh karena itu, penelitian ini hadir untuk mengisi celah tersebut dengan mengintegrasikan regresi Binomial Negatif dan penalti LASSO yang dioptimisasi

menggunakan AIC. Pendekatan ini ditujukan untuk menangani overdispersi dan *sparsity* pada fitur tekstual promosi di platform X, sekaligus mempertahankan interpretabilitas model statistik yang tidak dimiliki oleh metode *machine learning* konvensional.

Ringkasan perbandingan antara penelitian-penelitian tersebut dengan posisi penelitian ini disajikan dalam Tabel 1.1 sebagai berikut:

Tabel 1.1 Ringkasan Penelitian Terdahulu dan Posisi Penelitian

No	Peneliti (Tahun)	Metode dan Objek	Posisi / Kebaruan Penelitian
1	Lehman & Archer (2019)	NB-GMIFS (LASSO) pada Data Genetik	Mengembangkan pendekatan NB-LASSO pada data tekstual.
2	Liu & Pitt (2017)	BNBR & LASSO pada Data Asuransi	Menggunakan model *univariate* dan fokus pada optimisasi berbasis kriteria AIC.
3	Bhattacharya et al. (2017)	*Hurdle* NB pada *Engagement* Facebook	Mengganti seleksi manual dengan LASSO untuk menyaring ribuan kata promosi secara otomatis.
4	Cichosz (2023)	*NLP & Machine Learning* pada Artikel Medis	Berfokus pada model Regresi untuk mengukur pengaruh tiap kata secara statistik pada data hitung.
5	Andariesta & Wasesa (2023)	*Machine Learning* Klasifikasi pada *Engagement* X	Mempertahankan data hitung asli (bukan kategori) untuk menangani overdispersi secara presisi.
7	**Penelitian ini**	**AIC-Based NB LASSO pada @ShopeeID**	**Optimisasi penalti LASSO pada fitur tekstual *sparse* dan data overdispersi berbasis AIC.**

1.7. Sistematika Penulisan

Sistematika penulisan pada penelitian ini adalah sebagai berikut:

BAB I : Bab ini membahas latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, tinjauan pustaka, serta sistematika penulisan.

BAB II : Bab ini membahas landasan teori yang menjadi dasar penelitian, meliputi konsep regresi, *count data*, GLM, regresi Poisson, overdispersi, regresi Binomial Negatif, regularisasi LASSO, kriteria AIC, serta *text mining*.

BAB III : Bab ini membahas metode penelitian yang digunakan, meliputi jenis dan sumber data, variabel penelitian, *Software* yang digunakan, langkah-langkah analisis, dan diagram analisis data *(flowchart)*

BAB IV : Bab ini menyajikan hasil analisis data dan pemodelan yang telah dilakukan, meliputi eksplorasi data, pembentukan model Regresi Binomial Negatif, optimisasi model menggunakan penalti LASSO, pemilihan parameter penalti optimal berdasarkan AIC, serta interpretasi hasil model.

BAB V : Bab ini membahas kesimpulan yang diperoleh dari hasil penelitian serta saran yang dapat diberikan untuk pengembangan penelitian selanjutnya.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan, diperoleh beberapa kesimpulan sebagai berikut:

1. Pembentukan model regresi Binomial Negatif dengan regularisasi LASSO dilakukan dengan mengintegrasikan penalti L_1 ke dalam fungsi *log-likelihood* Binomial Negatif untuk melakukan seleksi fitur secara simultan. Model yang terbentuk tersebut kemudian dioptimisasi memanfaatkan kriteria *Akaike Information Criterion* (AIC) guna menyeimbangkan antara *goodness-of-fit* dan kompleksitas parameter model. Proses ini menghasilkan sebuah kerangka pemodelan yang secara simultan mampu mengestimasi parameter sekaligus melakukan seleksi fitur.
2. Penerapan model regresi Binomial Negatif dengan regularisasi LASSO pada 1.063 data teks promosi akun @ShopeeID berhasil menentukan nilai parameter penalti (λ) optimal sebesar $7{,}4438$ melalui metode *grid search*. Penerapan parameter penalti optimal ini memberikan dampak signifikan terhadap reduksi dimensi teks, di mana jumlah fitur berhasil dipangkas dari 75 fitur menjadi 45 fitur, tanpa mengorbankan kemampuan model dalam menjelaskan variabilitas data. Model optimal yang terbentuk menghasilkan nilai AIC sebesar $12.291{,}60$, lebih efisien dibandingkan model penuh tanpa regularisasi yang memiliki nilai AIC sebesar $12.333{,}42$.

3. Fitur tekstual yang signifikan dalam model optimal menunjukkan bahwa kata-kata yang berkaitan dengan promosi, keuntungan ekonomi, serta gaya komunikasi yang akrab memiliki pengaruh positif terhadap peningkatan jumlah *likes*. Hal ini ditunjukkan oleh nilai *Incidence Rate Ratio* (IRR) yang tinggi pada fitur seperti *"good"* (IRR = 13,22), *"shopping"* (IRR = 4,81), *"senilai"* (IRR = 3,95), dan "Rp0" (IRR = 3,93). Temuan ini mengindikasikan bahwa konten yang memuat *giveaway*, promo harga, maupun kampanye belanja tematik cenderung lebih mampu menarik interaksi audiens. Selain itu, fitur dengan nuansa percakapan sehari-hari seperti "sih" dan "nih" juga terbukti mendukung peningkatan *engagement*. Sebaliknya, fitur seperti "join", "bayar", dan "youtube" memiliki nilai IRR di bawah satu, yang menunjukkan bahwa konten dengan nuansa instruksional atau pengalihan ke platform eksternal cenderung menurunkan ekspektasi jumlah *likes*.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa pendekatan regresi Binomial Negatif dengan regularisasi LASSO efektif dalam mengidentifikasi fitur tekstual yang relevan serta meningkatkan performa model dalam menganalisis data *engagement* media sosial.

5.2. Saran

Setelah penelitian ini dilakukan, penulis menyampaikan beberapa saran yang diharapkan dapat bermanfaat bagi penelitian selanjutnya:

1. Penelitian ini menggunakan representasi teks berbasis TF-IDF dengan pendekatan unigram. Penelitian selanjutnya dapat mempertimbangkan penggunaan n-gram (bigram atau trigram) untuk menangkap pola frasa

promosi yang lebih kontekstual. Selain itu, pendekatan berbasis *word embedding* dapat dieksplorasi untuk membandingkan performa model, dengan tetap mempertimbangkan implikasinya terhadap tingkat *sparsity* dan kompleksitas komputasi.

2. Pemilihan parameter penalti dalam penelitian ini didasarkan pada minimalisasi AIC. Penelitian selanjutnya disarankan untuk melakukan komparasi menggunakan kriteria informasi lain seperti BIC atau pendekatan *Cross-Validation* guna mengevaluasi konsistensi fitur terpilih serta stabilitas model terhadap data baru.

3. Penelitian ini hanya menggunakan fitur tekstual sebagai variabel prediktor. Penelitian selanjutnya dapat menambahkan variabel lain seperti waktu unggah, jenis media (gambar atau video), serta karakteristik akun. Penambahan variabel tersebut diharapkan dapat meningkatkan kemampuan model dalam menjelaskan variasi *engagement* di media sosial.

DAFTAR PUSTAKA

Agresti, A. (2018). *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

Akaike, H. (1987). Factor analysis and aic. *Psychometrika*, 52(3):317–332.

Andariesta, D. T. & Wasesa, M. (2023). Machine learning models to predict the engagement level of twitter posts: Indonesian e-commerce case study. *Procedia Computer Science*, 227:823–832.

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing science*, 48(1):79–95.

Bhattacharya, S., Srinivasan, P., & Polgreen, P. (2017). Social media engagement analysis of us federal health agencies on facebook. *BMC medical informatics and decision making*, 17(1):49.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Cameron, A. C. & Trivedi, P. K. (2013). *Regression analysis of count data*. Number 53. Cambridge university press.

Cichosz, P. (2023). Bag of words and embedding text representation methods for medical article classification. *International Journal of Applied Mathematics and Computer Science*, 33(4):603–621.

Creswell, J. W. & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

De Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of interactive marketing*, 26(2):83–91.

Feldman, R. & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*, volume 103. Springer.

Lehman, R. R. & Archer, K. J. (2019). Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: Application predicting micronuclei frequency. *PloS one*, 14(1):e0209923.

Li, Y. & Xie, Y. (2020). Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of marketing research*, 57(1):1–19.

Liu, F. & Pitt, D. (2017). Application of bivariate negative binomial regression model in analysing insurance count data. *Annals of Actuarial Science*, 11(2):390–411.

Madnure, V. V. & Kada, P. A. (2025). Predictive modeling of user engagement patterns on social media using data mining approaches. *IJSAT-International Journal on Science and Technology*, 16(4).

Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing,.

McFadden, D. (1977). Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. Discussion Paper 474, Cowles Foundation for Research in Economics, Yale University.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Nugraha, J. (2014). *Pengantar analisis data kategorik: Metode dan aplikasi menggunakan program R*. Deepublish.

Rosa, G. J. (2010). The elements of statistical learning: Data mining, inference, and prediction by hastie, t., tibshirani, r., and friedman, j.

Salinas Ruíz, J., Montesinos López, O. A., Hernández Ramírez, G., & Crossa Hiriart, J. (2023). *Generalized linear mixed models with applications in agriculture and biology*. Springer Nature.

Seabold, S. & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 57–61.

Shen, R., Shen, J., Li, Y., & Wang, H. (2016). Predicting usefulness of yelp reviews with localized linear regression models. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 189–192. IEEE.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.