PENERAPAN *WEB SCRAPING* DENGAN PENDEKATAN N-GRAM UNTUK EKSTRAKSI INFORMASI

DARWIS YUSTOSIO
H1091211009

SKRIPSI

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

PENERAPAN *WEB SCRAPING* DENGAN PENDEKATAN N-GRAM UNTUK EKSTRAKSI INFORMASI

DARWIS YUSTOSIO
H1091211009

Skripsi
Sebagai salah satu syarat untuk memperoleh gelar
Sarjana Statistika pada Program Studi Statistika

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

LEMBAR PENGESAHAN

Judul Tugas Akhir	: Penerapan *Web Scraping* Dengan Pendekatan N-Gram Untuk Ekstraksi Informasi
Nama Mahasiswa	: Darwis Yustosio
NIM	: H1091211009
Jurusan/Program Studi	: Matematika/Statistika
Tanggal Lulus	: 20 Maret 2025
SK Pembimbing	: No. 1568/UN22.8/TD.06/2024/Tanggal 2 September 2024
SK Penguji	: No. 825/ UN22.8/TD.06/2025/Tanggal 11 Maret 2025

Dosen Pembimbing

Pembimbing I	Pembimbing II
Dr. Evy Sulistianingsih, M.Sc.	Hendra Perdana, M.Sc.
NIP 198502172008122006	NIP 198810102019031020

Dosen Penguji

Ketua Penguji	Anggota Penguji
Shantika Martha, M.Si.	Neva Satyahadewi, M.Sc., CRA., CRP., CRMP.
NIP 198403082008122003	NIP 198212042005012001

Pimpinan Sidang (merangkap anggota penguji)	Sekretaris Sidang (merangkap anggota penguji)
Dr. Evy Sulistianingsih, M.Sc.	Hendra Perdana, M.Sc.
NIP 198502172008122006	NIP 198810102019031020

Mengesahkan
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Tanjungpura

Prof. Dr. Gusrizal, M.Si.
NIP 197108022000031001

PERNYATAAN INTEGRITAS AKADEMIK

Saya yang bertanda tangan di bawah ini:

Nama	: Darwis Yustosio
NIM	: H1091211009
Program Studi/ Jurusan	: Statistika/Matematika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Dengan ini menyatakan bahwa dokumen ilmiah Tugas Akhir yang disajikan ini tidak mengandung unsur pelanggaran integritas akademik sesuai Peraturan Menteri Pendidikan, Kebudayaan, Riset, Dan Teknologi Republik Indonesia Nomor 39 Tahun 2021. Apabila di kemudian hari dokumen ilmiah Tugas Akhir ini mengandung unsur pelanggaran integritas akademik sesuai ketentuan perundangan tersebut, maka saya bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Demikian pernyataan ini untuk digunakan sebagaimana mestinya.

) Maret 2025



Darwis Yustosio
NIM H1091211009

PENERAPAN *WEB SCRAPING* DENGAN PENDEKATAN N-GRAM UNTUK EKSTRAKSI INFORMASI

Abstrak

Web scraping merupakan teknik otomatisasi yang digunakan untuk mengekstraksi data dari halaman *web*. Teknik ini memungkinkan pengambilan informasi dalam jumlah besar secara efisien tanpa harus melakukan penyalinan data secara manual. Dalam penelitian ini, *web scraping* diterapkan dengan pendekatan N-Gram untuk mengekstraksi dan menganalisis ulasan pelanggan terhadap pelayanan Hotel Maestro Pontianak. Data yang diperoleh kemudian digunakan untuk mengidentifikasi pola kata dan sentimen yang terkandung dalam ulasan pelanggan. Banyaknya pengunjung yang menginap di hotel ini menyebabkan banyaknya ulasan yang diberikan oleh pengunjung melalui *website* Traveloka. Penelitian ini bertujuan untuk menerapkan *web scraping* dalam pengambilan data ulasan pengunjung mengenai pelayanan Hotel Maestro Pontianak dari *website* Traveloka serta menganalisis informasi yang diperoleh menggunakan pendekatan N-gram untuk memahami opini pelanggan terhadap layanan hotel . Penelitian ini menggunakan teknik *Web Scraping* untuk mengumpulkan data ulasan, yang kemudian dilanjutkan dengan analisis sentimen menggunakan metode *Lexicon Based* dan N-Gram. Sebanyak 285 ulasan Hotel Maestro Pontianak dikumpulkan melalui *Web Scraping*. Data ulasan diproses menggunakan *Text Preprocessing* untuk menyeragamkan bentuk kata sehingga data lebih terstruktur. Selanjutnya, hasil *Text Preprocessing* diekstraksi fitur N-Gram kemudian diproses menggunakan metode *Lexicon Based* dengan bantuan *Library* VADER *Sentiment.* Hasil penerapan menggunakan pendekatan N-Gram, dari 285 data ulasan yang sudah diproses melalui *text preprocessing* dan diperoleh nilai *score polarity*, menunjukan hasil persentase setiap N-Gram. Dimana hasil persentase menunjukkan ulasan tersebut memiliki banyak opini positif mengenai pelayanan Hotel Maestro Pontianak. Informasi yang didapatkan dari hotel tersebut adalah kamar yang bersih, pelayanan yang baik dan ramah dari pihak hotel. Adapun keluhan dan perbaikan kedepannya yang perlu dilakukan oleh pihak hotel adalah masalah tempat parkir yang terbatas dan juga kebocoran platfon yang perlu dibenahi.

Kata kunci: N-Gram, *Unigram, Bigram, Trigram,* VADER.

APPLICATION OF WEB SCRAPING WITH N-GRAM APPROACH FOR INFORMATION EXTRACTION

Abstract

Web scraping is an automation technique used to extract data from web pages. This technique enables the retrieval of large amounts of information efficiently without the need for manual data copying. In this study, web scraping was applied with the N-Gram approach to extract and analyze customer reviews regarding the services of Hotel Maestro Pontianak. The collected data was then used to identify word patterns and sentiments contained in customer reviews. The high number of visitors staying at this hotel has led to many reviews being provided by guests through the Traveloka website. This study aims to apply web scraping to collect customer reviews about the services of Hotel Maestro Pontianak from Traveloka and analyze the obtained information using the N-Gram approach to understand customer opinions about the hotel services. This research employs Web Scraping techniques to collect review data, followed by sentiment analysis using the Lexicon-Based method and N-Gram approach. A total of 285 reviews of Hotel Maestro Pontianak were collected through web scraping. The review data underwent Text Preprocessing to standardize word forms, making the data more structured. Subsequently, the results of text preprocessing were extracted using the N-Gram feature and processed with the Lexicon-Based method using the VADER Sentiment library. The application of the N-Gram approach resulted in sentiment polarity scores from the 285 processed reviews, showing the percentage distribution for each N-Gram. The results indicate that the reviews contain many positive opinions regarding the services of Hotel Maestro Pontianak. The extracted information highlights aspects such as clean rooms, good service, and friendly staff. However, some complaints and areas for improvement include limited parking space and ceiling leaks that need to be addressed by the hotel management.

Keywords: *N-Gram, Unigram, Bigram, Trigram, VADER.*

PRAKATA

Puji dan syukur atas kehadirat Tuhan Yang Maha Esa karena berkat dan karunia-Nya lah penulis dapat menyelesaikan penulisan skripsi yang berjudul "Penerapan *Web Scraping* Dengan Pendekatan N-Gram Untuk Ekstraksi Informasi". Tujuan penulisan skripsi ini adalah untuk memenuhi syarat dalam memperoleh gelar Sarjana Statistika pada Program Studi Statistika Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Tanjungpura Pontianak.

Dalam penyusunan skripsi ini, tentu saja tidak lepas dari bimbingan, arahan, dan motivasi dari berbagai pihak. Untuk kesempatan ini, penulis ingin mengucapkan rasa terima kasih yang sebesar – besarnya kepada :

1. Kedua orang tua penulis, yaitu Bapak Saiful dan Ibu Halimah yang telah memberikan banyak dukungan baik materi ataupun dalam berbagai hal kepada penulis dan menjadi motivasi terbesar dalam kehidupan penulis.
2. Ibu Dr. Evy Sulistianingsih, M.Sc., selaku Dosen Pembimbing I yang telah memberikan arahan dan dukungan untuk menyelesaikan penulisan skripsi ini.
3. Bapak Hendra Perdana, M.Sc., selaku Dosen Pembimbing II yang telah memberikan arahan, dukungan, dan motivasi untuk menyelesaikan penulisan skripsi ini.
4. Ibu Shantika Martha, M.Si., selaku Dosen Penguji I yang telah memberikan saran dan masukan untuk perbaikan skripsi ini sehingga menjadi lebih baik.
5. Ibu Neva Satyahadewi, M.Sc., CRA., CRP., CRMP., selaku Dosen Penguji II yang telah memberikan saran dan masukan untuk perbaikan skripsi ini sehingga menjadi lebih baik.
6. Saudara Sahabat yang telah banyak membantu mengurus berkas agar penulis dapat menyelesaikan penulisan skripsi ini.

Pontianak, 20 Maret 2025

Darwis Yustosio

DAFTAR ISI

LEMBAR PENGESAHAN ii
PERNYATAAN INTEGRITAS AKADEMIK iii
Abstrak iv
Abstract **..... v**
DAFTAR ISI vii
DAFTAR TABEL ix
DAFTAR GAMBAR x
DAFTAR LAMPIRAN xi
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Batasan Masalah 4
1.5 Tinjauan Pustaka 4
1.6 Metodologi Penelitian 5
BAB II LANDASAN TEORI 7
2.1 Data *Mining* 7
2.2 *Text Mining* 7
2.3 *Browser Chrome* 8
2.4 *Python* 8
2.4.1 *Google Colab* 9
2.5 Analisis Sentimen 9
2.6 *Natural Language Processing* 10
2.7 *Machine Learning* 10
2.8 Kamus *Lexicon* 11
BAB III METODE LEXICON BASED DAN N-GRAM 12
3.1 *Web Scraping* 12
3.2 *Translate* Data 12
3.3 *Text Preprocessing* 13
3.3.1 *Cleaning* 13

3.3.2 *Spelling Normalization* 13
3.3.3 *Case Folding* 13
3.3.4 *Tokenizing* 14
3.3.5 *Stopword Removal* 14
3.3.6 *Stemming* 14
3.4 N-GRAM 14
3.5 *Lexicon Based* 15
3.5.1 *Library* VADER *Sentiment* 15
3.6 Visualisasi 16
3.6.1 Sentimen Positif Negatif Netral 16
BAB IV HASIL DAN PEMBAHASAN 17
4.1 Pengambilan Data Menggunakan Teknik *Web Scraping* 17
4.2 *Preprocessing* Data Ulasan 21
4.2.1 *Translate* Data Ulasan 21
4.2.2 *Cleaning* Data Ulasan 22
4.2.3 *Spelling Normalization* Data Ulasan 23
4.2.4 *Case Folding* Data Ulasan 24
4.2.5 *Tokenizing* Data Ulasan 24
4.2.6 *Stopword Removal* Data Ulasan 25
4.2.7 *Stemming* Data Ulasan 25
4.3 Ekstraksi N-Gram 26
4.4 Pelabelan Data Menggunakan *Lexicon Based* 26
4.4.1 Menentukan Kelas Polaritas dan *Score Polarity* 26
4.4 Visualisasi Data Ulasan 27
4.4.1 Visualisasi Sentimen Positif 28
4.4.2 Visualisasi Sentimen Negatif 30
4.4.3 Visualisasi Sentimen Netral 32
BAB V PENUTUP 35
5.1 Kesimpulan 35
5.2 Saran 35
DAFTAR PUSTAKA 37

DAFTAR TABEL

Tabel 4.1 Ilustrasi Data Ulasan Hasil *Web Scraping* 20
Tabel 4.2 Ilustrasi Data Hasil *Translate* 22
Tabel 4.3 Ilustrasi Hasil *Cleansing* 23
Tabel 4.4 Ilustrasi Hasil *Spelling Normalization* 23
Tabel 4.5 Ilustrasi Hasil *Case Folding* 24
Tabel 4.6 Ilustrasi Hasil *Tokenizing* 24
Tabel 4.7 Ilustrasi Hasil *Stopword Removal* 25
Tabel 4.8 Ilustrasi Hasil *Stemming* 25
Tabel 4.9 Ilustrasi Hasil N-Gram 26
Tabel 4.10 Ilustrasi Nilai *Compound Score Unigram* dan Kelas Sentimen 27

DAFTAR GAMBAR

Gambar 1.1 *Flowchart* Penelitian..... 7
Gambar 3.1 Tahap *Text Preprocessing* 13
Gambar 4.1 Menu Pencarian Pada *Website Traveloka*..... 17
Gambar 4.2 *Sitemap* Informasi *Link* Hotel Maestro Pontianak..... 18
Gambar 4.3 Proses *Selector Scrolldown*..... 18
Gambar 4.4 Proses *Selector Pagination* 19
Gambar 4.5 Proses *Selector Review* 19
Gambar 4.6 Variabel Data Hotel Maestro Pontianak 20
Gambar 4.7 *Bar Chart* N-Grams 28
Gambar 4.8 *Word Cloud* Kelas Positif *Unigram*..... 29
Gambar 4.9 *Word Cloud* Kelas Positif *Bigram* 29
Gambar 4.10 *Word Cloud* Kelas Positif *Trigram*..... 30
Gambar 4.11 *Word Cloud* Kelas Negatif *Unigram* 31
Gambar 4.12 *Word Cloud* Kelas Negatif *Bigram*..... 31
Gambar 4.13 *Word Cloud* Kelas Negatif *Trigram* 32
Gambar 4.14 *Word Cloud* Kelas Netral *Unigram* 33
Gambar 4.15 *Word Cloud* Kelas Netral *Bigram*..... 33
Gambar 4.16 *Word Cloud* Kelas Netral *Trigram* 34

DAFTAR LAMPIRAN

Lampiran 1 Tampilan Menu *Website Traveloka* 41
Lampiran 2 Hasil *Web Scraping* 42
Lampiran 3 *Syntax Google Colab Translate* Data 43
Lampiran 4 *Syntax Google Colab Text Preprocessing* 45
Lampiran 5 *Syntax Lexicon Based* dan N-Gram 48
Lampiran 6 *Syntax* Visualisasi 51

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dalam era digital yang semakin berkembang, internet telah menjadi sumber utama informasi bagi masyarakat Indonesia maupun luar negeri. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) mengumumkan bahwa jumlah pengguna internet di Indonesia tahun 2024 telah mencapai 221 juta jiwa dari total populasi 278 juta jiwa di Indonesia pada tahun 2023. Dari angka tersebut, 79% masyarakat Indonesia menggunakan internet untuk mengakses jejaring sosial. Pesatnya perkembangan teknologi informasi dan komunikasi menyebabkan banyaknya masyarakat dengan mudah mengakses informasi yang ada, salah satunya informasi terkait adalah sektor pariwisata. Sektor pariwisata merupakan sektor yang memiliki potensi untuk diperluas karena menciptakan lapangan kerja dan meningkatkan pendapatan yang cukup signifikan. Salah satu yang mendukung pembangunan pada sektor pariwisata adalah industri perhotelan (Bappenas, 2011).

Dalam industri pariwisata, hotel memiliki peran yang penting dalam membantu wisatawan yang berkunjung dengan menyediakan penginapan terutama bagi wisatawan dari luar kota. Publikasi Statistik Tingkat Penghunian Kamar Hotel Kota Pontianak 2023 memuat data Tingkat Penghunian Kamar (TPK), Tingkat Pemakaian Tempat Tempat Tidur (TPTT), Rata-rata Lama Menginap Tamu (RLMT), rata-rata tamu per kamar per bulan dari Januari sampai Desember 2023 untuk hotel berbintang dan hotel *non* bintang. Pada Publikasi ini, cakupan hotel yang menjadi unit observasi adalah hotel yang terdiri dari 23 hotel bintang dan beberapa hotel *non* bintang yang ada di Kota Pontianak (BPS, 2023). Salah satu hotel berbintang tiga di Pontianak adalah Hotel Maestro. Hotel Maestro Pontianak merupakan perusahaan yang bergerak dalam bidang pelayanan jasa perhotelan yang kegiatan usahanya menyediakan kamar bagi wisatawan dalam kota maupun dari luar kota.

Di era *Society* 5.0 sekarang, masyarakat dapat dengan mudah mengakses informasi terkait hotel yang ingin dikunjungi melalui jaringan internet, karena lebih

menghemat waktu dan tenaga. Salah satu *website* atau aplikasi yang menyediakan layanan pemesanan tiket pesawat dan hotel secara *online* adalah *Traveloka.* Ulasan *online* memiliki dampak yang signifikan terhadap persepsi dan keputusan pembelian seorang konsumen (Rachmawan, Swari, dan Putra, 2023). Pada situs penyedia tiket hotel seperti *Traveloka*, ulasan sangat mempengaruhi keputusan calon pengunjung baru. Faktor-faktor seperti fasilitas, kenyamanan, harga, dan ulasan pelanggan sebelumnya mempengaruhi keputusan saat memesan hotel.

Data ulasan pengunjung Hotel Maestro Pontianak pada *website Traveloka* dikumpulkan dalam bentuk teks yang berisi ulasan dengan menggunakan metode *Web Scraping* pada *browser Chrome* melalui *website Traveloka*. *Web scraping* (ekstraksi *web*) yaitu proses pengambilan sebuah informasi dokumen dari sebuah *web* dan menganalisis dokumen tersebut untuk mengambil data dari halaman tersebut yang digunakan bagi kepentingan yang lain (Ramadoni, Umaidah, dan Sari, 2020). Data hasil *Scraping* belum dapat digunakan secara langsung untuk dianalisis karena data yang diperoleh dari hasil *Web Scraping* memiliki bentuk yang tidak terstruktur sehingga perlu dilakukan *Text Preprocessing* untuk menyeragamkan bentuk kata pada ulasan.

Setelah data ulasan pengunjung Hotel Maestro Pontianak mengalami *Text Preprocessing*, kemudian dilakukan pendekatan dengan N-Gram. N-Gram merupakan teknik yang digunakan untuk memetakan hubungan antara kata-kata dalam sebuah teks dengan membagi teks menjadi urutan kata sebanyak N kata berturut-turut. N-Gram adalah kombinasi kata yang sering muncul untuk mengekspresikan suatu sentimen dan penggunaan N-Gram dianggap penting karena dalam bahasa Indonesia banyak frasa yang tidak hanya terdiri dari satu kata (Marga, Isnain, dan Alita, 2022). Penelitian ini menggunakan jenis token dalam N-Gram, yaitu *unigram, bigram,* dan *trigram*. *Unigram* adalah token yang terdiri dari satu kata, *bigram* adalah token yang terdiri dari dua kata, dan *trigram* adalah token yang terdiri dari tiga kata.

Ulasan diproses menggunakan metode *Lexicon Based.* Metode *Lexicon Based* memanfaatkan *machine learning* dengan cara membuat kamus berisikan kosakata yang mengandung opini (*lexicon*) terlebih dahulu. Kata-kata yang terdapat pada

kamus tersebut digunakan untuk mengidentifikasi apakah suatu kalimat mengandung opini atau tidak. Metode *Lexicon Based* mengklasifikasikan teks opini secara otomatis dengan menggunakan daftar kata (*lexicon*) yang telah diberi sentimen. Pendekatan *Lexicon Based* didasarkan pada orientasi kontekstual sentimen dengan jumlah orientasi sentimen pada setiap kata atau kalimat (Setiawan, 2014).

Penelitian ini memanfaatkan metode *Lexicon Based* dan pendekatan N-Gram untuk melakukan eksplorasi opini sentimen masyarakat terhadap pelayanan di Hotel Maestro Pontianak. *Sentiment Analysis* merupakan sebuah penggambaran polaritas pada suatu teks atau kata dan merupakan perpaduan dari data mining dan *text mining*, suatu teknik untuk menganalisa pendapat, sentimen, evaluasi, sikap, penilaian, perasaan dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik ataupun kegiatan tertentu (Cahyono dan Saprudin, 2019). Dengan analisis sentimen, ulasan Hotel Maestro Pontianak yang bersifat netral, positif atau negatif dapat diketahui.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, maka rumusan masalah yang berhubungan dengan penelitian ini adalah

1. Bagaimana penerapan *Web Scraping* untuk mengambil data ulasan pengunjung hotel dari *website Traveloka*?
2. Apa saja informasi ulasan yang dapat diekstraksi mengenai pelayanan hotel menggunakan pendekatan N-Gram?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang diambil, tujuan dalam penelitian ini adalah sebagai berikut.

1. Melakukan proses *Web Scraping* pada *website Traveloka* untuk mengambil data ulasan pengunjung mengenai pelayanan hotel.
2. Melakukan pengumpulan informasi dari data ulasan pengunjung dengan pendekatan N-Gram untuk opini pengunjung terhadap pelayanan hotel.

1.4 Batasan Masalah

Untuk tidak memperluas area pembahasan, perlu adanya batasan masalah untuk menyederhanakan permasalahan yang diteliti. Pembahasan dalam masalah ini sebagai berikut.

1. Pengambilan data berupa komentar atau ulasan dari *website Traveloka* menggunakan bantuan *browser Chrome* yang terhubung dengan internet.
2. Data yang diambil berupa komentar publik terhadap pelayanan Hotel Maestro Pontianak pada *website Traveloka* menggunakan metode *Web Scraping.*
3. Data diambil pada tanggal 26 Desember 2022 sampai dengan 27 September 2024 berjumlah 285 data ulasan.
4. Ulasan yang diambil adalah ulasan yang menggunakan Bahasa Indonesia yang berisi opini, dan analisis mengenai fasilitas Hotel Maestro Pontianak.
5. Ulasan dibagi menjadi ulasan positif, ulasan negatif, dan ulasan netral. Ulasan positif berisi penghargaan, kepuasan, atau pujian terhadap fasilitas dan kinerja pegawai Hotel Maestro Pontianak, ulasan negatif adalah kata-kata yang berisi kritik, ketidakpuasan, dan kesan buruk terhadap fasilitas dan kinerja pegawai Hotel Maestro Pontianak, sedangkan ulasan netral adalah yang tidak memihak atau tidak mengandung opini atau emosi yang kuat.

1.5 Tinjauan Pustaka

Hernikawati (2021) mengimplementasikan penggunaan *Lexicon Based* untuk melihat persentase tanggapan masyarakat Indonesia di media sosial *twitter* terhadap vaksin sinovac. Data yang dikumpulkan berjumlah 500 *tweets* yang diambil dari *twitter* melalui proses *crawling* pada tanggal 3 Mei 2021. Dilakukan *preprocessing* data pada seluruh teks untuk mendapatkan teks yang bersih dari pencilan (*noise*). Setelah diperoleh teks yang bersih, selanjutnya diolah dengan menggunakan metode *Lexicon Based* untuk mendapatkan nilai sentimen. Hasil yang didapatkan dari analisis sentimen masyarakat Indonesia terhadap vaksin sinovac adalah positif sebesar 35,4%, negatif sebesar 27%, dan netral sebesar 37,6%.

Anjani dan Fauzan (2021) mengimplementasikan N-Gram dalam menganalisis sentimen masyarakat Daerah Istimewa Yogyakarta (DIY) terhadap

Pembatasan Sosial Berskala Besar (PSBB) jilid dua di pulau Jawa dan Bali. Data yang digunakan dalam penelitian ini adalah data *tweet* dari tanggal 26 Januari sampai 8 Februari 2021. Banyaknya jumlah data yang terkumpul sebanyak 960 ulasan. Dengan menggunakan ekstraksi N-Gram, didapatkan hasil akurasi 83,04% untuk *unigram,* 80,36% untuk *bigram*, dan 82,14% untuk *trigram*.

Wijaya dan Panjaitan (2024) melakukan penelitian tentang analisis sentimen ulasan aplikasi *instagram* di *google playstore*. Data yang digunakan sebanyak 199 data ulasan. Dengan pendekatan N-Gram diperoleh akurasi 70% untuk *unigram*, akurasi 65% untuk *bigram*, dan akurasi 65% untuk *trigram*.

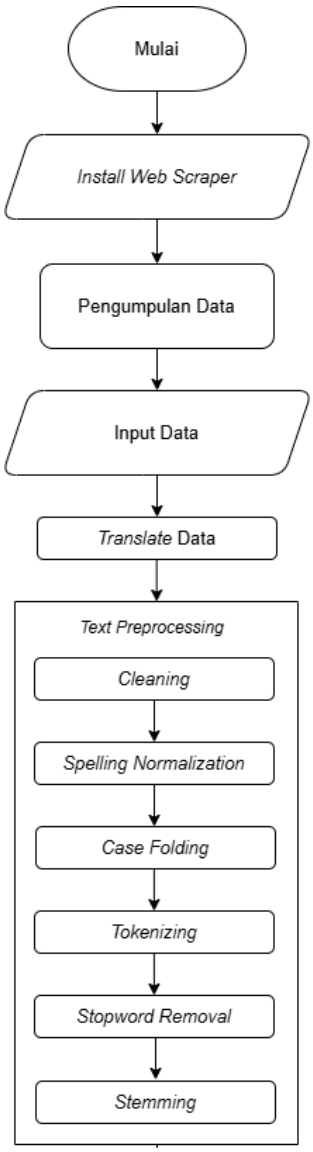
1.6 Metodologi Penelitian

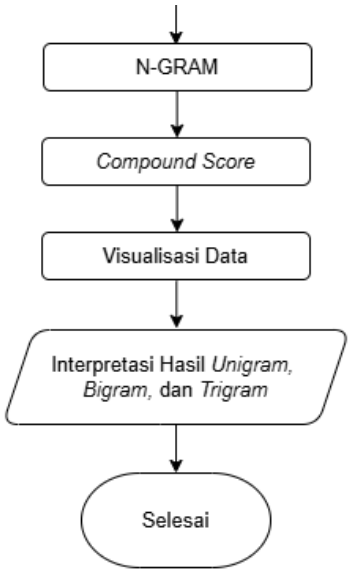
Penelitian ini diawali dengan studi literatur yaitu dengan dilakukan pencarian referensi teori yang sesuai dengan masalah yang diteliti. Data penelitian ini adalah data sekunder yang didapat dari *website Traveloka* menggunakan teknik *Web Scraping* menggunakan *browser Chrome* dengan bantuan *tools web scraper* yang telah diunduh. Data tersebut adalah ulasan atau komentar pengunjung Hotel Maestro Pontianak.

Dalam memulai proses Metode penelitian diperlukan langkah-langkah sebagai berikut.

1. *Scraping* data, yaitu suatu proses mengambil atau mengumpulkan data dari *website Traveloka*. Data yang diambil dapat berupa teks. Proses ini dilakukan menggunakan *browser Chrome* dengan bantuan *tools web scraper* yang telah diunduh.
2. *Text preprocessing*, yaitu suatu proses pengubahan bentuk sebuah data menjadi lebih terstruktur dalam proses data *mining*. Tahapan *text preprocessing* dimulai dari *cleaning, spelling normalization, case folding, tokenizing, stopword removal,* dan *stemming*.
3. Ekstraksi fitur N-Gram mencakup *unigram, bigram*, dan *trigram*. N-Gram adalah proses memecah teks menjadi deretan *string* atau urutan kata sebanyak N elemen.

4. Pemodelan *Lexicon Based*, yaitu suatu proses data setelah di ekstraksi fitur N-Gram diklasifikasikan dengan metode *Lexicon Based*, kemudian data diklasifikasi menjadi kategori positif atau negatif menggunakan *library* VADER *Sentiment*.
5. Langkah selanjutnya adalah visualisasi hasil pemodelan *Lexicon Based.* Kata yang sering muncul divisualisasikan dalam bentuk *word cloud* untuk mengetahui kata yang sering muncul dari *unigram, bigram,* dan *trigram.* Kemudian menampilkan hasil sentimen dalam bentuk *bar chart* dari *unigram, bigram*, dan *trigram*. Alur dalam penelitian ini dapat dilihat pada Gambar 1.1





Gambar 1.1 *Flowchart* Penelitian