BAB III

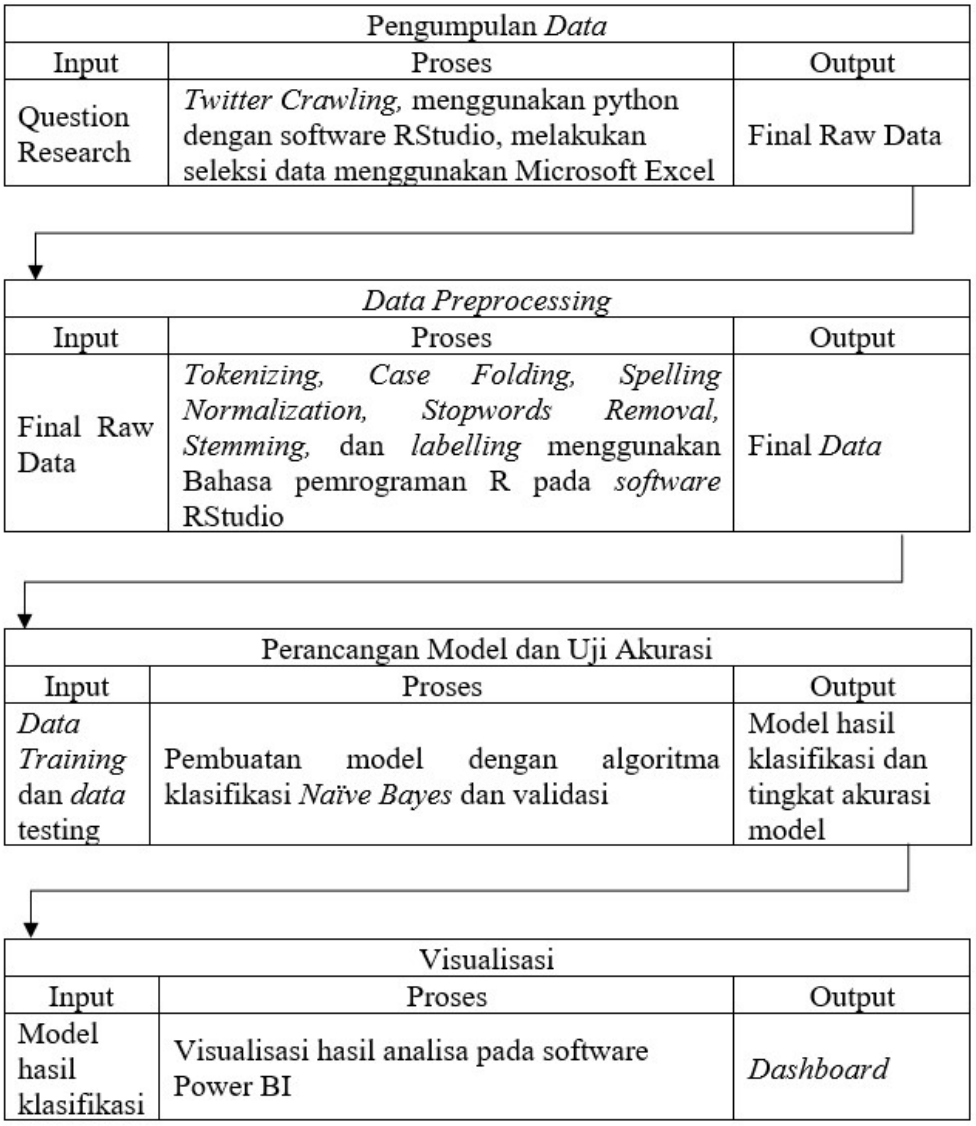
METODOLOGI PENELITIAN

3.1. Objek Penelitian

Landasan dalam menjadikan Raja Ampat sebagai objek penelitian adalah melihat sentimen wisatawan guna menjadi salah satu acuan bagi pihak terkait dalam pengembangan pariwisata dalam berbagai aspek menggunakan metode algoritma *Naïve Bayes* sebagai penghitungan kuantitatif. Penelitian ini akan mengambil hasil penemuan peluang kemunculan kata negatif atau positif dalam suatu kalimat. Lalu dipresentasikan dalam bentuk *dashboard Power BI*.

3.2. Alur Penelitian

Penelitian ini terdiri dari beberapa tahap dari pengumpulan data, data *pre-processing*, *labelling,* perancangan model, serta analisa dan pembuatan visualisasi dalam *dashboard,* untuk mempermudah penjelasan, dapat dilihat pada gambar 3.1 alur dari penelitian ini.



Gambar 3.1. Alur penelitian

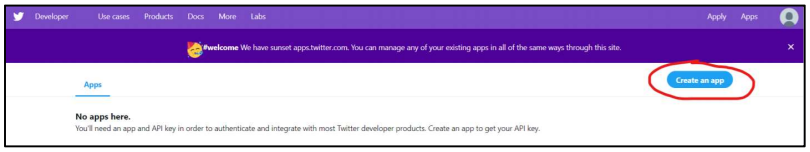
Gambar 3.1 menunjukan tahapan-tahapan yang dijalankan dalam penelitian ini. Berikut ini adalah penjabaran dari masing-masing tahapan penelitian yang dilakukan:

1. Proses pengambilan data berdasarkan *research question* dengan mencari *tweets* dengan tagar #rajaampat, dan *tweets* yang mengandung kata raja ampat. *Twitter crawling* dilakukan menggunakan *Tools Python* dan hasil *crawling* berupa dokumen format csv. Lalu hasil data tersebut dibuka menggunakan *Microsoft Excel* dan dilakukan seleksi data, seperti menghapus data duplikat, *tweets* yang hanya mengandung #rajaampat atau kata rajaampat saja.
2. *Data preprocessing* dilakukan setelah mendapatkan *raw data* yang masih harus dibersihkan, pada tahap ini dilakukan *Tokenizing, Case folding, Spelling Normalization, Stopwords Removal, Stemming* dan *data labelling.* Setelah dilakukan seluruh rangkaian maka didapatkan data yang bersih dan siap dilakukan analisa.
3. Setelah di dapatkan *data training* dan *data testing* dari proses sebelumnya, dilakukan perancangan model untuk mengolah *data training* tersebut. pada tahap ini digunakan Bahasa pemrograman R dan menggunakan algoritma *Naïve Bayes* sebagai metode untuk menghasilkan sentimen berdasarkan *data training,* lalu dilakukan validasi menggunakan *data testing.* Hasil dari tahap ini ialah model hasil klasifikasi dan presentase akurasi model.
4. Dari data yang dihasilkan pada tahap ini dilakukan visualisasi berupa *dashboard* menggunakan *software Power BI*. Isi dari *dashboard* tersebut ialah presentase pebandingan sentimen postif, negatif dan netral, dan grafik *tweets* berkala, jumlah *tweets*, rata-rata *tweets*, jumlah *likes*, rata-rata *likes* per *tweets*, dan *slicer* negatif, positif, netral.

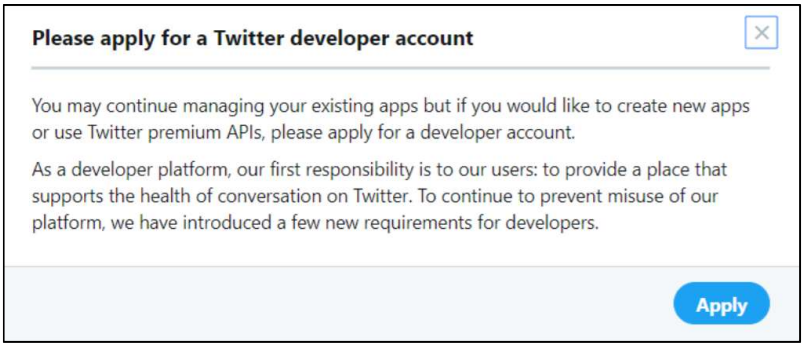
3.2.1. Pengumpulan data

Penelitian ini menggunakan *library tweepy* pada Bahasa pemrograman *Python*. *Tweets* dikumpulkan berdasarkan kata kunci tagar "#rajaampat, raja ampat" dengan memanfaatkan *streaming API Twitter* yang disediakan oleh *tweepy.* Data merupakan *tweets* dari tanggal 1 Januari 2019 sampai 5 Mei 2020. Berikut adalah tahapan secara rinci bagaimana pengumpulan data dari *Twitter*:

Register Hak Akses otentifikasi pada *Twitter developer* pada situs https://*developer*.*Twitter*.com/en/apps. Masuk menggunakan akun *Twitter* yang telah dimiliki. Lalu akan muncul gambar 3.2, klik pada tombol yang dilingkari merah. Selanjutnya klik tulisan *apply* pada gambar 3.3.

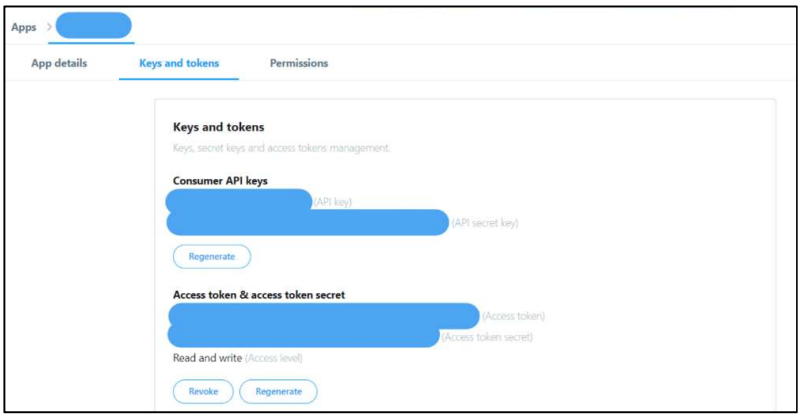


Gambar 3.2. Tampilan awal setelah masuk *Twitter developer*



Gambar 3.3. Mendaftar pembuatan aplikasi

Selanjutnya hanya tinggal mengikuti arahan dari *Twitter* dan menunggu balasan bahwa pendaftaran kita diterima oleh *Twitter*, karena pendaftaran ini diperiksa langsung oleh karyawan *Twitter* demi melihat maksud dan tujuan penelitian maka akan memakan waktu yang lebih lama, sekitar 1 hari kemudian. Setelah itu, kita dapat membuat aplikasi yang terlihat seperti gambar 3.4.



Gambar 3.4. Terdapat token untuk pengambilan data

Data *crawling* dan Menyimpan data *tweets* dalam dokumen dengan format csv seperti yang terlihat pada gambar 3.5.

```
pip install twitterscraper
twitterscraper "#rajaampat" --output raja.csv --limit 100000 --begindate 2019-01-01 --enddate 2020-05-05 –csv
twitterscraper "Raja Ampat" --output ampat.csv --limit 100000 --begindate 2019-01-01 --enddate 2020-05-05 –csv
```

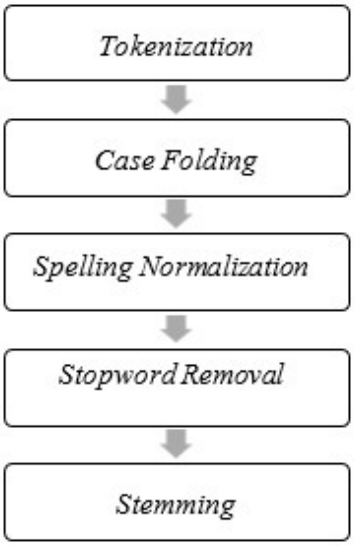
Gambar 3.5. *Script* untuk mengambil *tweets*

Setelah didapatkan data mentah hasil *crawling data,* maka selanjutnya akan dilakukan seleksi data menggunakan *Microsoft Excel*, pada proses ini

dilakukan penghapusan data yang duplikat menggunakan fitur *delete duplicate*, serta data yang hanya mengandung #rajaampat ataupun kata rajaampat, secara manual.

3.2.2. *Data preprocessing*

Gambar 3.6. menjelaskan tahapan pada *data pre-processing* yang akan dilakukan:



Gambar 3.6. proses *pre-processing* data

(Ramachandran & Parvathi, 2019)

1. *Tokenization*

   Untuk memproses teks, penting untuk menemukan batasan kata. Batas-batas diidentifikasi menggunakan spasi dan tanda baca. Proses pemisahan kalimat menjadi bagian-bagian yang bermakna dan mengidentifikasi individu dalam kalimat tersebut

disebut Tokenisasi. Keuntungan utama dari *tokenization* khususnya *Twitter* adalah pemisahan dari *URL* dan *Hashtag* yang ada di *tweets*. Identifikasi tagar sangat membantu dalam pemrosesan lebih lanjut dalam banyak aplikasi seperti deteksi tren, penambangan opini, deteksi acara, dan banyak lagi (Ramachandran & Parvathi, 2019)

2. *Case folding*

   Pada setiap tweets terdapat kemungkinan bahwa akan ada huruf kapital dan kecil, untuk memudahkan dalam penyeragaman dan pembacaan setiap kata, maka pada tahap ini bertujuan untuk membuat semua teks menjadi huruf kecil (Mujilahwati et al., 2016)

3. *Spelling Normalization*

   Pada tahap ini dilakukan penyetaraan ejaan kata atau bahasa gaul akan diubah menjadi kata baku (Mujilahwati et al., 2016).

4. *Stopword Removal*

   Stopwords adalah kata-kata yang kurang membantu dalam analisis lebih lanjut dari *tweets*. Kata-kata ini dihapus dari *tweets* sebelum diproses lebih lanjut (Ramachandran & Parvathi, 2019). Dengan dihilangkannya *stopwords* maka jumlah kata yang dikirim untuk analisis menjadi jauh lebih sedikit (Ramachandran & Parvathi, 2019).

5. *Stemming*

   Stemming adalah tahap yang bertujuan untuk menemukan kata dasar dari kata-kata yang berimbuhan. Pada tahapan ini, setiap kata akan diubah ke kata dasarnya. Seperti kata ‘menyukai’ disederhanakan menjadi suka.

3.2.3. Perancangan model

Dengan metode *Naive Bayes*, data akan diprediksi sentimennya sesuai dengan *classifier* yang sudah dilatih sebelumnya. Analisis sentimen dengan metode *Naïve Bayes* akan dilakukan dengan melakukan pelatihan pada *classifier* dengan menggunakan data pelatihan. Data pelatihan yang akan digunakan harus diberikan label sentimen terlebih dahulu. Setelah *classifier* berhasil dilatih dengan data pelatihan, maka peneliti akan menggunakan data pengujian untuk memprediksi label kelas pada data pengujian tersebut, dan menentukan tingkat akurasi dari hasil analisis sentimen ini. Untuk melakukan analisis sentimen dengan metode *Naïve Bayes*, maka akan digunakan *function classifier Naive Bayes* yang terdapat pada *RStudio*.

3.2.4. Pembuatan *Dashboard*

Setelah itu, maka akan dilakukan proses analisa dan visualisasi data dalam aplikasi *desktop Power BI* berupa *dashboard* yang mana memberikan informasi hasil dari analisa yang sudah dilakukan mengenai analisa sentimen objek wisata Raja Ampat, antara lain, perbandingan

antara sentimen positif, negatif, dan netral, dan grafik *tweets* berkala (tahun, bulan, minggu, hari), jumlah *tweets*, rata-rata *tweets*, jumlah *likes*, rata-rata *likes* per *tweets*, dan *slicer* negatif, positif, netral.

3.2.5. *Problem Solving*

Pada penelitian ini, algoritma yang digunakan untuk menjadi *problem solving* ialah Algoritma klasifikasi *naïve bayes*. *Naive Bayes* merupakan metode pengklasifikasian probabilistik sederhana. Metode ini akan menghitung probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Salah satu alasan penulis memilih untuk menggunakan metode ini adalah *Naive Bayes* menganggap semua atribut pada setiap kategori tidak memiliki ketergantungan satu sama lain (independen) sesuai dengan data yang digunakan pada penelitian ini, selain itu, memiliki kelebihan yang dapat menangani data dengan jumlah yang besar. Juga, keuntungan penggunaan *Naive Bayes* adalah hanya memerlukan sejumlah kecil data latih untuk menentukan parameter *mean* dan *varians* dari variabel yang diperlukan untuk klasifikasi. *Naive Bayes* merupakan metode *supervised document classification* yang berarti membutuhkan *data training* sebelum melakukan proses klasifikasi (Devita, Herwanto, & Wibawa, 2018) dan (Ria et al., 2018).

Selain *Naive Bayes*, terdapat algoritma untuk melakukan klasifikasi, salah satunya adalah *K-Nearest Neighbor*(KNN), berikut perbandingan

antara metode *Naïve Bayes* dan metode *K-Nearest Neighbor* yang ditunjukan pada table 3.1.

Tabel 3.1. Perbandingan *Naïve Bayes* dengan K-Nearest Neighbor

Sumber: (Devita et al., 2018)

Naïve Bayes	*K-Nearest Neighbor*
Mempertimbangkan hasil dari parameter sebelumnya	Tidak mempertimbangkan berdasarkan kemungkinan sebelumnya
Merupakan *linear classifier*, yang mana dalam memproses data besar lebih cepat	Bukan *linear classifier*, lebih lambat dalam proses data skala besar.
Naïve Bayes membuat 2 parameter, untuk disesuaikan pada perataan. *Hyperparameter* adalah parameter sebelumnya yang disetel pada set pelatihan untuk mengoptimalkannya.	Karena hanya memiliki 1 parameter yaitu "k" atau biasa disebut *number of neighbor*. Sehingga tidak mempertimbangkan berdasarkan kemungkinan parameter sebelumnya.
Metode ini tidak terpengaruh oleh dimensi dan data yang besar	KNN memiliki dapat terpengaruh hasilnya dengan dimensi dan data yang besar
Supervised learning algorithm	*Supervised learning algorithm*

Dengan membandingkan 2 algoritma diatas maka diambilah *Naïve Bayes* sebagai algoritma yang akan digunakan dalam penelitian ini, dengan keunggulan yang sudah dijelaskan diatas sesuai dengan kebutuhan analisa, ditunjang dengan spesifikasi yang dijelaskan dalam tabel 3.1 guna membantu proses penelitian.

3.3. Variabel Penelitian

Dalam penelitian ini, variabel penelitian yang digunakan diantaranya:

3.3.1. Variabel independen

X1 = Kata di dalam *tweets* yang mengandung tagar "#rajaampat, dan kata raja ampat"

3.3.2. Variabel Dependen

Y1 = Positif

Y2 = Negatif

Y3 = Netral

3.4. Teknik Pengumpulan Data

Teknik pengumpulan data penelitian ini menggunakan *library tweepy* pada Bahasa pemrograman *Python*. *Tweets* dikumpulkan berdasarkan kata kunci tagar "#rajaampat, raja ampat" dengan memanfaatkan *streaming API Twitter* yang disediakan oleh *tweepy*. Tagar tersebut dipilih karena berhubungan dengan topik yang akan di analisa pada penelitian ini yaitu objek pariwisata Raja Ampat.

3.5. Teknik Pengambilan Sampel

Pada penelitian ini digunakan teknik *random Sampling*. *Random Sampling* digunakan oleh peneliti apabila populasi diasumsikan *homogen*. Sehingga sampel dapat diambil secara acak. *Random Sampling* adalah teknik pengambilan sampel

dimana tiap indvidu dalam populasi dapat peluang yang sama untuk terpilih. Dalam penarikan sampel ini, peneliti memilih sampe secara acak tanpa bias agar diadapatkan hasil se objektif mungkin.

Teknik *Sampling* secara acak dapat dilakukan dengan beberapa cara. Antara lain, *Sampling* acak sederhana yaitu penentuan sampel dengan cara melakukan undian terhadap populasi. *Sampling* acak beraturan yaitu dalam hal ini peneliti mengambil sampel dari nomor-nomor subjek dengan jarak yang sama yang telah ditentukan sebelumnya dan Sampel acak dengan bilangan random. Pada penelitian ini penulis menggunakan teknik random *Sampling* acak sederhana.

3.6. Teknik Analisis Data

Tools yang digunakan pada penelitian ini yaitu Bahasa pemrograman R, karena R memang diunggulkan untuk melakukan analisis yang bersifat data teks atau *text mining* seperti memecah paragraf menjadi frasa dan kata-kata, lebih unggul dalam *data science* karena latar belakang statistik, Lebih banyak digunakan dalam dunia pendidikan. Selain R, *Tools* lain yang digunakan ialah *Software Power BI* untuk melakukan proses pembuatan *dashboard* untuk penyajian visualisasi.