Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

Perpustakaan Univeristas Multimedia Nusantara

BAB III
METODOLOGI PENELITIAN

3.1. Objek Penelitian

Objek penelitian adalah hal yang menjadi sasaran penelitian (Kamus Bahasa Indonersia; 1989: 622). Menurut (Supranto 2000:21) objek penelitian adalah himpunan elemen yang dapat berupa orang, organisasi atau barang yang diteliti. Kemudian dipertegas (Anto Dayan 1986: 21), objek penelitian adalah pokok persoalan yang hendak diteliti untuk mendapatkan data secara lebih terarah. Adapun objek penelitian dalam analisa ini adalah opini atau sentimen masyarakat terhadap sebuah program televisi yang berjudul Indonesia Lawyers Club (ILC) dan Mata Najwa yang diambil di media sosial Twitter.

ILC merupakan program acara *talkshow* yang mengangkat topik yang sedang hangat atau baru terjadi seputar kriminal, kebijakan pemerintah, dan fenomena-fenomena yang terjadi di lingkungan masyarakat Indonesia. Dalam setiap debatnya yang mengangkat tema yang cukup hangat, program ILC menghasilkan banyak opini publik. ILC memiliki sebuah akun media sosial di Twitter dengan username @ILCtv1. Biasanya ILC memberikan informasi seputar tayangan programnya melalui cuitan pada akun Twitternya. Sebagai program acara yang cukup aktif di Twitter, tak jarang ILC mendapatkan kritik dan pujian dari masyarakat. Banyak opini-opini yang dilemparkan kepada ILC di Twitter.

Mata Najwa merupakan program acara *talkshow* yang mengangkat tema yang hampir mirip dengan ILC. Mata Najwa memiliki akun twitter resmi @MataNajwa. Pada akun twitternya, Mata Najwa memberikan informasi terkait tema *talkshow* yang dibawakan dan juga memberikan interaksi terhadap masyarakat di Twitter.

3.2. Data

Menurut Kamus Besar Bahasa Indonesia, data merupakan keterangan atau bahan nyata yang dapat dijadikan dasar kajian (analisis atau kesimpulan). Data adalah istilah majemuk yang berisi fakta atau bagian dari fakta yang memiliki arti yang dihubungkan dengan kenyataan, simbol, gambar, angka, huruf, ataupun simbol yang menunjukkan suatu ide, objek, kondisi atau situasi (John J. Longkutoy (2012).

Data yang digunakan dalam analisa ini adalah yang diperoleh dari *twitter* yang biasa disebut *twitter* data berisi beberapa variabel yang dapat digunakan untuk analisa. Variabel-variabel yang terdapat pada *twitter* data adalah *screen_name, username, user_id, tweet_id tweet_url, timestamp, timestamp_epochs, text, links, hashtags, has_media, img_urls, video_url, likes, retweets, replies, is_replied, is_reply_to, parent_tweet_id, dan reply_to_users*.

Variabel yang digunakan dalam penelitian ini adalah *screen_name, text, timestamp*, sentimen, *accuracy*, *precision* dan *recall*. Variabel *text* berisi opini masyarakat, variabel *timestamp* berisi tanggal *tweet* diunggah dan variabel sentimen berisi positif atau negatif dari hasil analisa atau *labelling* secara manual.

3.2.1. Variabel Independen

Variabel independen (variabel bebas) adalah variabel yang mempengaruhi atau yang menjadi penyebab terjadinya perubahan atau munculnya variabel terikat (Sugiyono, 2009).

Variabel independen dalam penelitian ini adalah *screen_name*, *text* dan *timestamp* karena variabel ini yang menjadi faktor perubahan variabel sentimen. Variabel *text* berisi opini-opini masyarakat yang diproses sebagai dataset sedangkan *timestamp* berupa waktu sebuat *tweet* diunggah

3.2.2. Variabel dependen

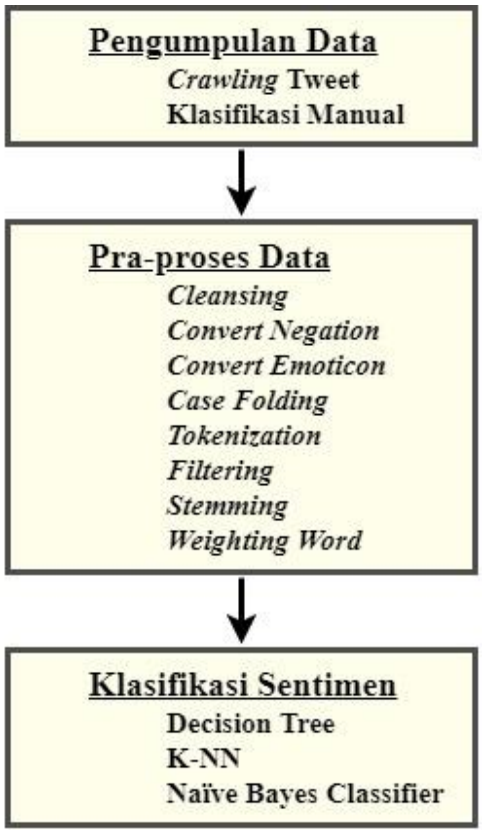
Variabel dependen (variabel terikat) adalah variabel yang dipengaruhi atau yang menjadi akibat, karena adanya variabel bebas (Sugiyono, 2009). Dalam penelitian ini variabel terikatnya adalah sentimen. Variabel sentimen bertipe binominal dimana memiliki dua buah label positif dan negatif.

Tabel 3. 1. Hasil dari nilai aktual dan prediksi.

	Prediksi Positif	Prediksi Negatif
Aktual Positif	True Positif	False Positif
Aktual Negatif	False Negatif	True Negatif

3.3. Alur Penelitian

Alur penelitian pada Gambar 3. 1 ini mengacu pada penelitian yang dibuat oleh Achmad Bayhaqy, 2018. Alur penelitian terbagi menjadi 3 tahap besar yaitu pengumpulan data, pra-proses data, dan klasifikasi sentimen.



Gambar 3. 1. Alur Penelitian.

3.3.1. Pengumpulan Data

Tahap pertama dalam penelitian ini adalah pengumpulan data. Pengumpulan data adalah proses mengumpulkan informasi yang digunakan

dalam penelitian. Dalam tahap pengumpulan data ini dibagi menjadi dua bagian yaitu *crawling tweet* dan klasifikasi manual.

a. Crawling Tweet

Crawling tweet adalah proses penarikan twitter data menggunakan aplikasi *data mining.* Proses *crawling* ini biasanya dilakukan dengan menggunakan *application interface* (API) yang sudah disediakan oleh Twitter, namun dalam penelitian ini tidak menggunakan API melainkan hanya menggunakan *library* python yang telah dibuat oleh Ahmet Taspinar pada tahun 2019. *Library* python Ahmet Taspinar diambil di sebuah aplikasi berbasis website yang menyimpan cukup banyak *repository* tidak berbayar https://github.com/taspinar/twitterscraper/.

Perbedaan menggunakan menggunakan API dengan *library* python Ahmet Taspinar adalah jika menggunakan API maka diperlukan biaya yang cukup besar untuk *crawling* data yang sudah lewat 7 hari sedangkan jika menggunakan *library* python tidak membutuhkan biaya.

Kelebihan *library* python Ahmet Taspinar yang digunakan dalam penelitian ini adalah dapat mengambil data yang lampau yang cukup besar, memiliki filter tanggal, bahasa, dan isi dari *tweet* sedangkan kekurangannya membutuhkan koneksi internet yang stabil, jika terputus harus mengulang dari awal.

b. Klasifikasi Manual

Klasifikasi manual ini dilakukan untuk memberi nilai positif dan negatif. Klasifikasi manual ini dilakukan dengan bantuan satu orang yaitu Calvin untuk menentukan 30 *tweet* perbulan selama tahun 2018 dan 2019 yang merupakan opini, setelah itu mencari tiga orang responden untuk membantu melabelkan data yaitu Cornelius Anipar, Ridho Al Qodari, Calvin. Proses *labelling* ini menggunakan data dari data twitter yang sudah di*crawling*.

3.3.2. Pra-proses Data

Tahap kedua dalam penelitian ini adalah tahap pra-proses data. Data yang didapatkan dari *crawling* adalah data mentah yang masih penuh. Pada tahap inilah data dibersihkan sebelum masuk tahap klasifikasi. Berikut ini bagian-bagian pada tahap pra proses data.

a. *Cleansing*

Cleansing adalah tahap bagian untuk menghapus duplikasi data, URL dan karakter simbol, angka dan tanda baca yang tidak diperlukan dari teks seperti tanda seru, tanda tanya, tanda petik dan lain sebagainya. Contoh pada tahap bagian *cleansing*:

Sebelum:

> Sebuah Program televisi ‘Kriminal’, tidak dicocokan untuk anak! https://kriminalprogramtvberita.com

Setelah:

> Sebuah Program televisi Kriminal tidak dicocokan untuk anak

b. *Case Folding*

Case folding adalah tahap bagian untuk membuat semua besar kecilnya huruf menjadi sama, yaitu antara 2, huruf kecil atau huruf besar. Pada penelitian ini semua huruf diubah menjadi kecil untuk mempermudah proses berikutnya. Contoh *case folding*:

Sebelum:

> Sebuah Program televisi Kriminal tidak dicocokan untuk anak

Sesudah:

> sebuah program televisi kriminal tidak dicocokan untuk anak kecil

c. *Tokenization*

Tokenisasi adalah tahap bagian untuk memecah atau memotong suatu kalimat atau dokumen menjadi beberapa bagian atau kata-kata yang disebut token. Jenis token ada tiga yaitu unigram, bigram, dan trigram (Wahyu, 2017). Dalam penelitian ini jenis token yang

digunakan adalah unigram, yaitu pemecahan dokumen menjadi per satu kata. Contoh *tokenization*:

Sebelum:

> sebuah program televisi kriminal tidak dicocokan untuk anak

Sesudah:

> sebuah, program, televisi, kriminal, tidak, dicocokan, untuk, anak

d. *Filtering*

Filterisasi dilakukan untuk menghilangkan kata-kata yang sering muncul, namun tidak dibutuhkan atau tidak memiliki makna. Kata-kata yang muncul dalam jumlah besar dan dianggap tidak memiliki makna ini disebut dengan *stopwords*. Pada dasarnya, daftar *stopwords* merupakan kumpulan dari kata-kata yang cukup banyak digunakan di berbagai macam bahasa.

Alasan untuk menghapus *stopwords* adalah karena penggunaannya yang terlalu umum dan memiliki makna yang dapat dianggap kurang. Dengan menghapus *stopwords*, maka penelitian dapat fokus pada kata-kata lain yang dianggap jauh lebih memiliki makna dan penting. Contoh kata-kata yang dapat dikategorikan sebagai *stopwords*: ada, di, kalau, pada, yaitu, sebuah, saja, itu, untuk, hari, berbagai, aku, mereka. Contoh dari penghapusan *stopwords*:

Sebelum:

> sebuah, program, televisi, kriminal, tidak, dicocokan, untuk, anak

Sesudah:

> program, televisi, kriminal, tidak, dicocokan, anak

e. *Stemming*

Stemming merupakan tahap bagian untuk menjadikan semua kata yang memiliki imbuhan atau akhiran menjadi kata dasar sesuai dengan aturan bahasa Indonesia yang benar. *Stemming* dilakukan dengan menghapus setiap imbuhan dan akhiran pada kata prefiks atau suffix. Contoh imbuhan seperti di-, meng-, ber-, ter-, per-, men-, se-, ke-, dan lain sebagainya. Sedangkan akhiran seperi –kan, -an, -i, -nya. Contoh *stemming*:

Sebelum:

> program, televisi, kriminal, tidak, dicocokan, anak

Sesudah:

> program, televisi, kriminal, tidak, cocok, anak

f. *Weighting Word*

Weighting word merupakan tahap akhir untuk memberi skor atau nilai pada frekuensi kemunculan kata dalam sebuah dokumen atau teks. Salah satu metode untuk *weighting word* adalah *term frequency-inverse document frequency* (TF-IDF).

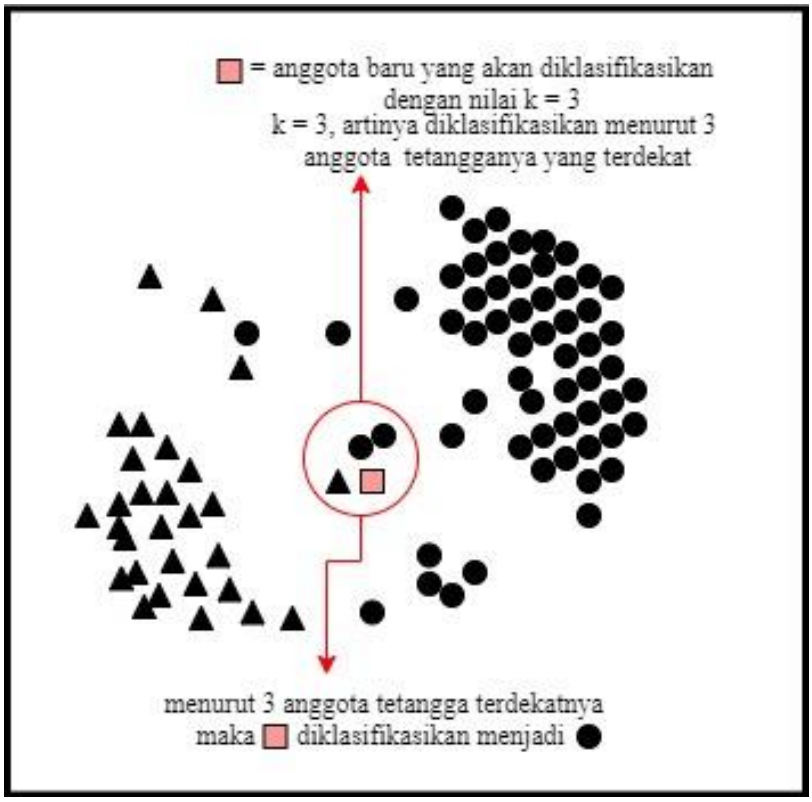
3.3.3. Klasifikasi Sentimen

Klasifikasi sentimen adalah tahap di mana semua data yang telah melewati pra-proses, diproses dengan algoritma klasifikasi. Dalam penelitian ini menggunakan tiga algoritma klasifikasi yaitu *Decision Tree*, *K-Nearest Neighbor*, dan *Naïve Bayes Classifier*.

a. K-Nearest Neighbor

Penggunaan K-NN dapat dilihat pada ilustrasi Gambar 3. 2 dimana ada dua kelas yaitu lingkaran dan segitiga. Namun ada data baru yang belum diketahui kelasnya yang ditandai dengan bentuk persegi berwarna merah. Untuk mengetahui kelas dari persegi, maka dibutuhkan sebuah aturan untuk menentukan kelasnya. Dalam contoh ini maka ditentukan k = 3. Nilai k = 3 artinya persegi di klasifikasikan kelasnya berdasarkan 3 anggota tetangganya yang terdekat. Berdasarkan nilai k = 3, terdapat 2 kelas lingkaran dan 1 kelas segitiga.

Sehingga nilai lingkaran lebih besar dari segitiga dan data persegi dapat diklasifikasikan menjadi kelas lingkaran.



Gambar 3. 2. Metode K-NN dengan memilih tetangga terdekat.

b. Naïve Bayes Classifier

Sebagai metode yang dikenal menggunakan probabilitas atau peluang bersyarat, *Naïve Bayes Classifier* ini dirumuskan sebagai

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Rumus 3. 1. Rumus *Naïve Bayes Classifier*.

X adalah bukti atau data sedangkan *H* adalah hipotesis. *P(H|X)* adalah probabilitas yang menyatakan bahwa hipotesis *H* benar untuk bukti *X*. *P(X|H)* adalah probabilitas untuk bukti *X* benar untuk hipotesis *H*. *P(H)* adalah probabilitas yang sebelumnya menyatakan bahwa hipotesis *H* benar untuk setiap objek data tidak peduli dengan nilai-nilai atributnya, sedangkan *P(X)* adalah probabilitas yang sebelumnya untuk objek data yang *X*.

c. *Decision Tree*

Algoritma ID3 yang merupakan salah satu algoritma dari metode *decision tree* yang digambarkan pada Gambar 3. 3, menggunakan sebuah fungsi rekursif atau fungsi yang dapat memanggil dirinya sendiri yang dijabarkan dengan langkah-langkah berikut:

function ID3(Himpun_Sampel, Atribut_Target, Himpun_Atribut)

1. Buat sebuah simpul *root*.
2. ***if*** semua sampel sama dengan kelas *i*, ***then return*** pohon satu simpul *Root* dengan label *i*
3. ***if*** Himpun_Atribut tidak berisi atau kosong ***then return*** pohon satu simpul *root* dengan label = nilai atribut target yang paling umum atau yang paling sering muncul ***else***

A <-- atribut berupa *best classifier* / dengan informasi gain yang besar atau terbesar

Atribut keputusan untuk untuk *root* <-- *A*

for v_i, (setiap nilai yang terdapat dalam *A*)

Tambahkan beberapa cabang dibawah *Root* sesuai dengan nilai v_i. Buat variabel contohnya seperti $Sample_{vi}$ untuk menjadi himpunan bagian atau subset dari Himpun_Sampel yang mempunyai nilai v_i pada atribut *A*

if Sampel v_1 kosong

then dibawah cabang ini tambahkan suatu simpul daun atau simpul yang tidak memiliki anak dibawahnya dengan label = nilai atribut target yang paling sering muncul atau yang umum.

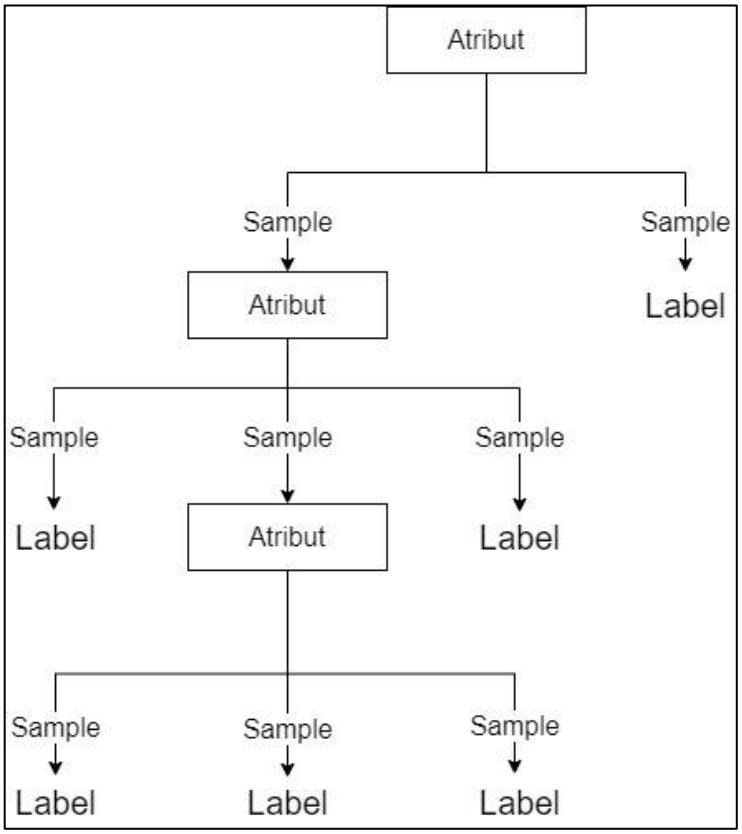
else di bawah cabang yang ini tambahkan sebuah atau beberapa subtree (pohon bagian) dengan memanggil fungsi ID3 ($Sample_{vi}$, Atribut_Target, Atribut-{A})

end

end

end

4. ***return*** *Root* untuk mengakhiri



Gambar 3. 3. Pohon keputusan yang dihasilkan fungsi ID3.