BAB III

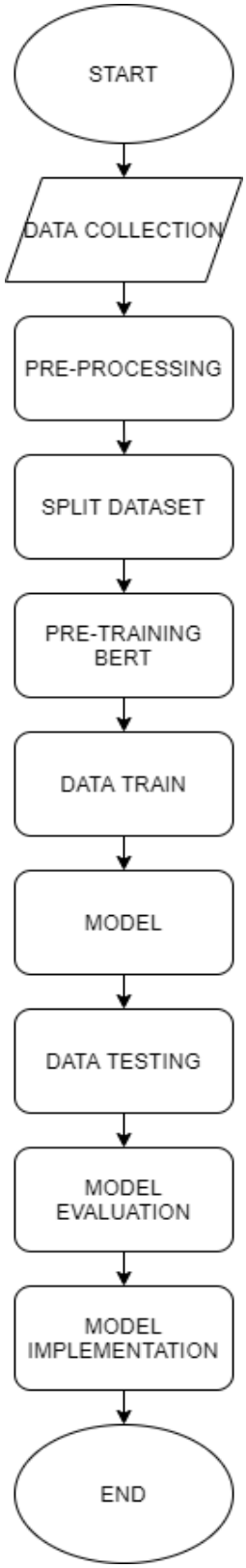
METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian ini adalah pemodelan analisis sentimen berbasis pembelajaran mesin *deep learning* dalam topik NLP (*Natural Language Processing*) menggunakan arsitektur BERT (Bidirectional Encoder Representations from Transformer) untuk mengklasifikasi sentimen ulasan aplikasi Gojek pada platform Play Store menjadi 3 label yaitu positif, netral dan negatif. Model akan dibangun dengan beberapa variasi rasio dan *hyperparameter* yang akan dijelaskan pada bab selanjutnya, dimana penelitian ini akan melakukan seleksi model sesuai dengan hasil evaluasi performa dengan nilai paling baik. Model yang terpilih akan digunakan untuk memprediksi kalimat ulasan dari data baru, yaitu teks yang didapatkan dari pengguna aplikasi Gojek dengan peserta survei yaitu mahasiswa UMN angkatan 2015 dan ulasan data baru Play Store . Hal ini ditujukan untuk menguji model dapat bekerja dengan baik pada ulasan dengan data baru.

3.2 Alur Penelitian

Alur penelitian yang digunakan dalam penelitian ini adalah sebagai berikut:

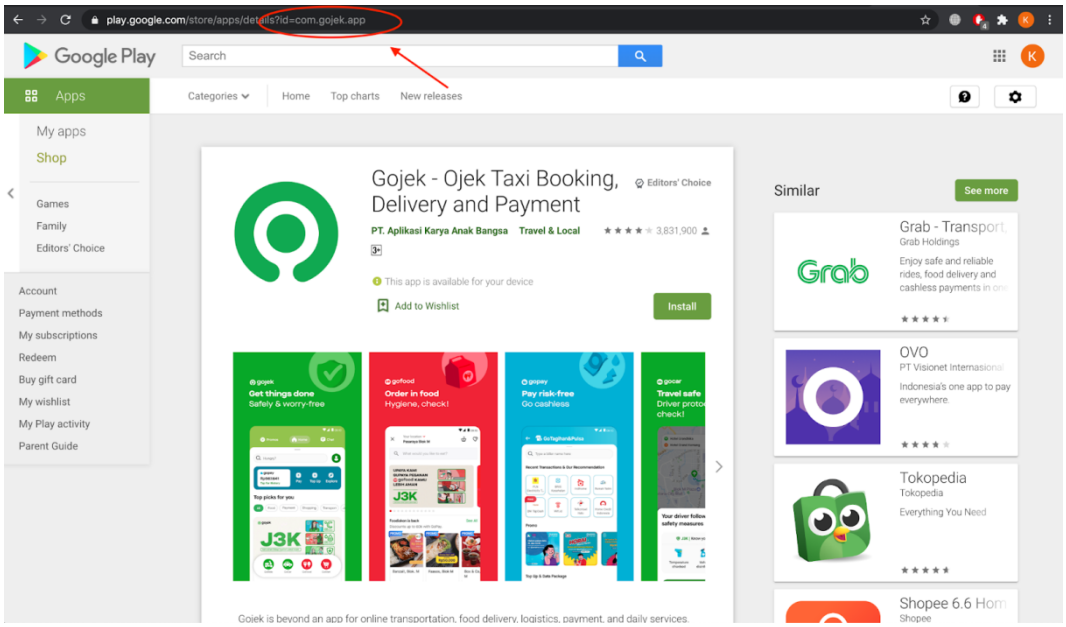


Gambar 3. 1 Alur Penelitian

Pada gambar 3.1 dapat dilihat bahwa alur penelitian tersebut yang akan diimplementasikan dalam penelitian ini akan dijelaskan pada penjelasan berikut:

3.2.1 Pengumpulan Data

Pengumpulan data dilakukan dengan metode *scraping* data di *platform* Google Play Store menggunakan *python script* dengan *library google play scraper*[1]. *Scraping* dilakukan untuk mendapatkan ulasan pengguna Gojek yang tertulis pada *platform* Google Play Store. Ulasan yang didapatkan akan difilter dengan parameter tanggal, antara 1 Januari 2021 hingga 8 Mei 2021. Berikut cara melakukan *scraping* data ulasan Gojek di Google Play Store:



Gambar 3. 2 Tampilan Google Play Store

[1] (https://github.com/JoMingyu/google-play-scraper) merupakan *library google play scraper*. Setelah itu, hasil ulasan Gojek tersebut akan disimpan dalam format csv.

Pada gambar 3.2, kata kunci yang digunakan adalah ‘com.gojek.app’ merujuk pada *query id* aplikasi Gojek di Play Store.

```
pip install google-play-scraper
```

Gambar 3. 3 Instalasi *Google Play Scrapper*

```
from google_play_scraper import Sort, reviews_all

result_id = reviews_all(
    'com.gojek.app',
    sleep_milliseconds=0, # defaults to 0
    lang='id', # defaults to 'en'
    country='id' # defaults to 'us'
)
```

Gambar 3. 4 Python Script bahasa Indonesia

```
result_en = reviews_all(
    'com.gojek.app',
    sleep_milliseconds=0, # defaults to 0
    lang='en', # defaults to 'en'
    country='id', # defaults to 'us'
)
```

Gambar 3. 5 Python Script bahasa Inggris

Gambar 3.3 merupakan *library* yang digunakan pada proses *scraping*. Selanjutnya, Gambar 3.4 dan gambar 3.5 menunjukkan *python script* yang digunakan untuk *scraping* pada Play Store, dimana bahasa yang digunakan adalah bahasa Indonesia dan bahasa Inggris.

Data baru Google Play Store dan data survei digunakan untuk melakukan validasi akhir model yang dikumpulkan mengikuti aturan *Roscoe* untuk menentukan *sampling* yaitu menggunakan teknik *proportionate simple random sampling* yang dimana ukuran sampel yang

memadai untuk digunakan adalah 30 sampel[33].Hasil data baru Google Play Store diambil dari tanggal 9 Mei 2021- 16 Mei 2021 , sedangkan hasil survei didapatkan berdasarkan penyebaran kuesioner terhadap 30 responden yang mencakup 8 fakultas. Pengambilan survei menggunakan metode *Stratified proportional simple random sampling* pada setiap program studi angkatan 2015 mahasiswa UMN dengan rumus.

$$\frac{Total\ mahasiswa\ program\ studi}{Total\ mahasiswa\ UMN\ angkatan\ 2015} x30$$

Rumus 2 Random Sampling pada Survei Mahasiswa UMN

Pada kuesioner, responden diminta untuk memberikan komentar (*review*) mengenai pengalaman mereka dalam menggunakan aplikasi Gojek dan melakukan *labeling* (positif atau negatif atau netral) terhadap komentar tersebut. Selanjutnya, kedua *dataset* ini akan digunakan sebagai pengujian akhir dari model yang dipilih untuk melihat performa model. Tabel 3.1 merupakan tabel pembagian sampel mahasiswa berdasarkan program studi.

Tabel 3. 1 Pembagian Sampel Mahasiswa berdasarkan Program Studi

Program Studi	Jumlah Mahasiswa	Jumlah Sampel
Akuntansi	78	2
Manajemen	126	3
Desain Komunikasi Visual	261	5
Film dan Televisi	275	5
Sistem Informasi	88	2

Teknik Informatika	124	2
Teknik Komputer	15	1
Ilmu Komunikasi	489	10
Total	**1456**	**30**

Perhitungan tersebut didapatkan dengan cara total mahasiswa per program studi dibagikan total keseluruhan mahasiswa UMN angkatan 2015 yang nantinya dikalikan dengan jumlah sampel. Jadi, dalam menentukan jumlah sampel yang diambil akan dijelaskan dengan contoh sebagai berikut. Program studi Sistem Informasi didapatkan jumlah mahasiswa per program studi sebanyak 88 lalu dibagikan dengan jumlah keseluruhan mahasiswa UMN angkatan 2015 yaitu 1456 didapatkan 0,0604. Lalu, dari hasil tersebut dikalikan jumlah sampel yaitu 30 sehingga angka yang didapatkan adalah 1,813 yang dimana angka tersebut dibulatkan menjadi 2.

3.2.2 *Data Preprocessing*

Pada tahapan data *preprocessing*, terdapat 4 tahapan yang dilakukan. Tahapan ini dilakukan agar ulasan-ulasan yang ada dapat diolah secara lebih lanjut:

Tahapan pertama adalah *case folding* teks. Dalam tahapan pada tabel 3.2, semua huruf yang ada pada ulasan diubah menjadi huruf kecil (tidak kapital). Sehingga memudahkan untuk pembacaan dan penyeragaman.

Tabel 3. 2 Tabel *Case Folding*

No	Content (Sebelum Casefolding)	Content (Sesudah Casefolding)
1	Aman dan cepat.....	aman dan cepat.....
2	Insentif prakerja tgl 19 Mei 2021 kenapa belum cair ewalet yang lain sudah cair tolong bantuannya Gojek	insentif prakerja tgl 19 mei 2021 kenapa belum cair ewalet yang lain sudah cair tolong bantuannya gojek
3	Thanks! Gopay saya yang hilang akhirnya sudah masuk. Saya harap, kelalaian seperti ini tidak terjadi lagi. Tingkatkan lagi keakuratan sistem, pelayanan dan keamanan, terutama untuk sistem pembayaran gopay, saya harap tidak ada lagi transaksi yang missing/ tdk tercatat oleh sistem ataupun tertunda.	thanks! gopay saya yang hilang akhirnya sudah masuk. saya harap, kelalaian seperti ini tidak terjadi lagi. tingkatkan lagi keakuratan sistem, pelayanan dan keamanan, terutama untuk sistem pembayaran gopay, saya harap tidak ada lagi transaksi yang missing/ tdk tercatat oleh sistem ataupun tertunda.
No	Content (Sebelum Casefolding)	Content (Sesudah Casefolding)
4	Pelayanan go car hari ini buruk bgt, pesan dimana malah kita costumer yg suruh jalan kaki samperin. Saya sudah chat bawa baby (kebayang ga lg covid gini) dan juga disuruh bayar parkir dmn dia berada. Tolong dong bisa direview driver nya yg sprt ini. Soalnya di app gojeknya ga bisa complain mengenai driver yg gini.	pelayanan go car hari ini buruk bgt, pesan dimana malah kita costumer yg suruh jalan kaki samperin. saya sudah chat bawa baby (kebayang ga lg covid gini) dan juga disuruh bayar parkir dmn dia berada. tolong dong bisa direview driver nya yg sprt ini. soalnya di app gojeknya ga bisa complain mengenai driver yg gini.

Tahapan kedua adalah *data cleansing*. Pada gambar 3.6 merupakan gambar sebelum *data cleansing* dan pada gambar 3.7 merupakan gambar sesudah *data cleansing*. Ketika dilakukan penarikan data, maka data yang didapat tidak hanya berupa ulasannya saja, melainkan terdapat komponen-komponen lain juga. Oleh karena itu *data cleansing* perlu dilakukan untuk

menghapus *reviewId*, *userName*, *userImage*, *score*, *thumbUpCount*, *reviewCreatedVersion*, *at*, *replyContent*, dan *repliedAt*.

	reviewId	userName	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	sentiment
0	gp:AOqpTOHoOFwMuWB7fL4gckG_8UrOGJP9zW8Y7NpiijB...	Sidik Jailani	https://play-lh.googleusercontent.com/a/AATXAJ...	Aman dan cepat.....	5	0	4.16.1	2021-05-07 13:26:43	NaN	NaN	positive
1	gp:AOqpTOHQbQj5ukkLgaJedOC3xJtpSgcGka_EIKJaDID...	A. S. W.	https://play-lh.googleusercontent.com/a-/AOh14...	Insentif prakerja tgl 19 Mei 2021 kenapa belum...	4	0	4.18.2	2021-05-07 12:04:34	NaN	NaN	neutral
2	gp:AOqpTOFCk4aNwHB6Jy6aWzXkIlwmwrBXfORUBX6TRVW...	Salsabila Laily Rahma	https://play-lh.googleusercontent.com/a-/AOh14...	Thanks! Gopay saya yang hilang akhirnya sudah ...	1	1	4.19.2	2021-05-07 11:32:58	Hai Salsabila, mohon maaf ya. Jika kamu masih ...	2021-05-07 3:43:57	positive
3	gp:AOqpTOE9RUxaQg12applB6cJGPIQfYUChsR1X_nmdIC...	Stefani Irene	https://play-lh.googleusercontent.com/a-/AOh14...	Pelayanan go car hari ini buruk bgt, pesan dim...	1	0	4.18.2	2021-05-07 8:20:27	NaN	NaN	negative
4	gp:AOqpTOHh0x5SwnpsHBBMCsCxdJcWj9iMW9USpBCT0uK...	Yoyok eko Prasetyo	https://play-lh.googleusercontent.com/a-/AOh14...	Kecewa.. tranfer uang ke rek dana status.. ngg...	1	0	4.18.2	2021-05-07 8:13:50	NaN	NaN	negative

Gambar 3. 6 Sebelum *Data Cleansing*

	content	sentiment
0	Aman dan cepat.....	positive
1	Insentif prakerja tgl 19 Mei 2021 kenapa belum...	neutral
2	Thanks! Gopay saya yang hilang akhirnya sudah ...	positive
3	Pelayanan go car hari ini buruk bgt, pesan dim...	negative
4	Kecewa.. tranfer uang ke rek dana status.. ngg...	negative

Gambar 3. 7 Sesudah *Data Cleansing*

Pada tahapan ketiga dilakukan *trim white space*. Tahapan ini dilakukan untuk menghapus spasi yang lebih dari satu, sehingga menghapus atau meminimalisir kesalahan pengertian atau data tidak terdeteksi, pada proses data training dan data testing.

Tahapan keempat atau yang terakhir, Pada gambar 3.8 merupakan gambar sebelum penghapusan *emoticon* dan pada gambar 3.9 merupakan gambar sesudah penghapusan *emoticon*. dilakukan penghapusan *emoticon*. Artian dalam *emoticon* yang sangat luas dan tergantung pada preferensi

masing-masing pribadi, memberikan kesempatan besar untuk kesalahan pengertian. Oleh sebab itu, agar mendapatkan ulasan yang apa adanya, maka *emoticon* perlu untuk dihapuskan.

No	Content
1	Sepertinya Gojek udah mau gulung tikar dehh😭. Mau order gomart kok cm ada 1 toko doang cuma alfamart😭. Ongkir dinaikin banyak yg complain, ongkir Diturunin jg bnyk yg complain kl apk ribet. Anda sehat...
2	Ini kenapa gopay nya gk bisa di hubungkan ke playstore? Error terus dah😩
3	Aplikasinya gblk jlk bngt g bisa apa apa😱
4	Gw kesel sumpah seharian gw gak dpt driver buat pesen makan,kecewa aja sih😔

Gambar 3. 8 Tabel sebelum penghapusan Emoticon

	C	D
No	Content	
1	sepertinya gojek udah mau gulung tikar dehh mau order gomart kok cm ada 1 toko doang cuma alfamart ongkir dinaikin banyak yg complain ongkir diturunin jg bnyk yg complain kl apk ribet anda sehat	
2	ini kenapa gopay nya gk bisa di hubungkan ke playstore? error terus dah	
3	aplikasinya gblk jlk bngt g bisa apa apa	
4	gw kesel sumpah seharian gw gak dpt driver buat pesen makan kecewa aja sih	

Gambar 3. 9 Tabel sesudah penghapusan Emoticon

Setelah keempat tahapan ini selesai, maka didapatkan ulasan-ulasan murni yang hanya mengandung tulisan saja. Alhasil meminimalisir kesalahan baca data pada proses data *training* dan data *testing*.

3.2.3 *Split Dataset*

Dataset Gojek yang diambil pada tanggal 1 Januari 2021 sampai 8 Mei 2021 memiliki total sebanyak 13583 data. *Dataset* akan dibagi menjadi data *training* dan data *testing*. Dari keseluruhan data terdapat 3 skenario rasio yaitu, 60% dan 40%, 70% dan 30%, serta 80% dan 20%. Pembagian ini berdasarkan penelitian rasio dari penelitian prediksi nilai tukar mata uang asing menggunakan *extreme learning machine*[34]. Setelah itu data training akan dibagi menjadi data *training* dan juga data

validasi dengan rasio masing-masing 80% dan 20% dari jumlah data *training*. Setelah dilakukan pembagian pada ke 3 data tersebut (data *training*, data *validation* dan data *testing*) di simpan dalam *format* tsv.

Pembagian data *testing*, *training*, dan *validation* menggunakan metode *proportionate stratified random sampling*. *Proportionate Stratified random sampling* adalah dimana populasi dibagi menjadi strata (atau subkelompok) dan sampel acak diambil dari setiap subkelompok. Sebuah subkelompok adalah sekumpulan item alami. Subkelompok mungkin didasarkan pada ukuran perusahaan, jenis kelamin, atau pekerjaan (untuk beberapa nama). Pengambilan sampel sering digunakan jika terdapat banyak variasi dalam suatu populasi. Tujuannya adalah untuk memastikan bahwa setiap strata terwakili secara memadai[35].

3.2.4 *Pre-Trained* BERT

1. IndoBERT *Base*

Model *pre-trained* BERT yang digunakan pada penelitian ini adalah IndoBERT yang merupakan arsitektur yang khusus dilatih menggunakan data *corpus* bahasa Indonesia[9]. Data set yang digunakan mencapai 4 milyar kata, baik bahasa informal maupun formal dengan 12 korpus bahasa Indonesia yang berbeda. *Dataset* ini kemudian dilatih dengan arsitektur BERT standard yang memiliki 12 *transformers layers*, 78 *hidden layers* dan 12 *attentions heads* yang menghasilkan 124,5 juta parameter.

2. Hyperparameter tuning

Menurut penelitian dari BERT, ada tiga parameter yang dapat disesuaikan untuk mengoptimalkan performa pada saat fine-tuning, di antaranya: *batch size*, *learning rate*, dan jumlah *epochs*[11]. Maka, berdasarkan rekomendasi tersebut akan dilakukan pencarian nilai terbaik dari ketiga parameter tersebut.

Dikarenakan kombinasi yang cukup banyak dan membutuhkan waktu komputasi yang cukup panjang, maka proses ini dibantu menggunakan modul *trainer* yang tersedia pada *library "transformer*" dari *huggingface*. Model IndoBERT yang sudah tersedia pada *huggingface* akan di unduh sebagai model *default* yang selanjutnya akan digunakan untuk *transfer learning* atau *fine-tuning*.

Berdasarkan hasil pencarian *hyperparameter tuning* ini melalui *default* model IndoBERT, nilai-nilai parameter tersebut yang akan diambil dan digunakan untuk tahap pelatihan model. Adapun daftar alternatif nilai parameter yang digunakan [11] adalah sebagai berikut:

- *Batch size : 8*
- *Learning rate : 3e-6*
- *Epoch : 3,4*

Pada penelitian ini, *hyperparameter tuning* akan disesuaikan dengan spesifikasi dari google collab yang digunakan yaitu *batch size* sebesar 8 , *learning rate* sebesar 3e-6 dan 2 variasi *epoch* yaitu 3 dan 4.

3.2.5 *Data Training*

Data *training* yang digunakan merupakan file tsv yang didapatkan dari proses pembagian data sebelumnya yaitu *train data* dan *validation data*. Data *training* di *load* sesuai *hyperparameter tuning* yang telah ditetapkan sebelumnya yaitu *batch size* sebesar 8, jumlah *worker* sebanyak 16, maksimal dari panjang *sequence* adalah 512 dan data *training* akan di acak. Selama proses pelatihan, akan dilakukan pemantauan terhadap *loss* dan akurasi dari data *training* dan data validasi pada setiap *epoch*. Proses pelatihan akan dijalankan menggunakan Google Collab, *IDE* berbasis *Jupyter Notebook* untuk pemrograman *Python* yang menggunakan sumber daya komputasi yang disediakan oleh Google Cloud dengan spesifikasi: RAM sebesar 26 GB, GPU NVIDIA Tesla P100 dengan VRAM sebesar 16 GB. Data latih menggunakan *Adam* sebagai *Optimizer* dan menyesuaikan pada pengaturan default dari model BERT dengan menggunakan *learning rate* 3e-6. Pada tahapan sebelumnya, didapatkan hasil pencarian parameter yang optimal. Hasil pencarian ini berguna untuk menyesuaikan penyetelan parameter.

3.2.6 Model

Model yang didapatkan setelah proses pelatihan akan disimpan dalam format *pickle* yang selanjutnya akan dibandingkan pada proses *data testing* untuk mendapatkan model terbaik, dimana model terbaik akan digunakan sebagai model akhir yang memprediksi ulasan yang didapatkan dari survei dan *dataset* Gojek yang baru.

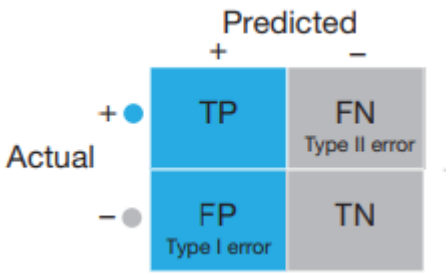
3.2.7 *Data Testing*

Data t*esting* yang telah disiapkan akan di *load* sesuai *hyperparameter tuning* yang telah ditentukan seperti pada proses menyiapkan data *training*. Data *testing* ini kemudian akan digunakan pada proses selanjutnya yaitu untuk melakukan evaluasi model dan menetapkan performa dari setiap model.

3.2.8 *Model Evaluation*

Evaluasi model akan dilakukan dengan menggunakan *confusion matrix* untuk mengukur performa model yang didapatkan dari proses *fine-tuning*. Hasil yang akan dihitung adalah *precision*, *recall*, *accuracy* dan *f1 measure*. *Confusion Matrix* berukuran n x n berkaitan dengan pengklasifikasian yang menunjukkan prediksi klasifikasi dan aktual klasifikasi, yang mana n adalah jumlah kelas yang berbeda.

Confusion matrix atau *classification matrix* dikalkulasi menggunakan *TP (True Positive), FP (False Positive), FN (False Negative),* dan *TN (True Negative)*, seperti yang ditunjukkan pada Gambar 3.10



Gambar 3. 10 *Confusion Matrix*

Sumber : [36]

Suatu prediksi termasuk *TP* jika yang diprediksi benar sesuai dengan yang terjadi yaitu bernilai benar, *FP* atau kesalahan tipe jika yang diprediksi adalah salah tetapi tidak sesuai dengan kenyataan yaitu bernilai benar, *FN* atau kesalahan tipe 2 jika suatu prediksi menyatakan benar tetapi nyatanya bernilai salah, dan *TN* jika prediksi menyatakan nilai salah dan keadaan yang terjadi juga bernilai salah.

Melalui empat tipe tersebut, terdapat [36]:

a. *Precision* yang digunakan untuk mengukur frekuensi jawaban/prediksi yang tepat dari suatu kenyataan dan menunjukan kualitas berhasil melakukan prediksi dari sistem yang diterapkan. *Precision* dapat dirumuskan sebagai berikut:

$$Precision = \frac{TP}{TP + FP}$$

Rumus 3 Menghitung *Precision*

b. *Recall* digunakan untuk mengukur frekuensi berapa kali suatu kategori yang ditetapkan di deteksi dan menunjukan kualitas untuk menunjukan setiap prediksi. *Recall* dapat dirumuskan sebagai berikut:

$$Recall = \frac{TP}{TP + FN}$$

Rumus 4 Menghitung *Recall*

c. *Accuracy* adalah pembagian dari semua hasil prediksi yang benar. *Accuracy* dapat dirumuskan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Rumus 5 Menghitung *Accuracy*

d. *F1-Score* mengkalkulasi harmoni antara *precision* dan *recall*. *F1-Score* dapat dirumuskan sebagai berikut:

$$F1\ Score = 2\frac{precision\ x\ recall}{precision + recall}$$

Rumus 6 Menghitung *F1-Score*

Pada tahapan evaluasi model, digunakan kurva *loss* dan kurva akurasi. Kurva *loss* digunakan untuk mengukur tingkatan optimal model dalam menggunakan parameter (acuan) yang didapat untuk proses pembelajaran. Kemudian, kurva akurasi berguna untuk melihat kinerja yang dihasilkan model. Berdasarkan evaluasi tersebut, maka model yang dipersiapkan akan lebih maksimal pada saat masuk ke tahap *model implementation*.

3.2.9 *Model Implementation*

Model terbaik dipilih melalui penilaian performa menggunakan *f-1 measure* dan juga *accuracy*. Setelah itu, model terbaik akan diimplementasi untuk memprediksi hasil survei mahasiswa UMN 2015 mengenai pengalaman mereka sendiri terhadap pelayanan Gojek dan juga data baru ulasan Play Store.