OUTLINE SKRIPSI
Revisi 0

Penerapan Metode Feature Selection Pada Algoritma Naive Bayes Untuk Analisis Sentimen Pengguna APLIKASI Pembajakan Konten Digital

Disusun Oleh :

Jeremy Jason 2301852050

Stephen Oanjiro Andy 2301860974

Jessen Lailovy Kunawan 2301882400

Program: Sistem Informasi

KAMPUS: ALAM SUTERA

TANGERANG

7

HALAMAN PENGESAHAN OUTLINE SKRIPSI

Data Mahasiswa

NIM	Nama	Program	No HP
		Sistem Informasi	
		Sistem Informasi	
2301882400	**Jessen Mailovy Gunawan**	Sistem Informasi	

Kampus *	**:**	Alam Sutera		
Jalur Skripsi *	**:**	● Non Class	● Non-Class (Internship 3+1)	● Non-Class Artikel Ilmiah
Tipe Skripsi		● Pre-Thesis	● Thesis	
Topik	**:**	**Business Intelligence**		
Judul (Indonesia)	**:**	**Penerapan Metode Feature Selection Pada Algoritma Naive Bayes Untuk Analisis Sentimen Pengguna Twitter Terhadap Pembajakan Konten Digital**		
Judul (Inggris)	**:**			
Usulan Pembimbing	**:**	1. D3408 - Mediana Aryuni		
		2. D2971 - Eka Miranda		
		3. <<kode Dosen – Nama Dosen>>		
		<<Sengaja dikosongkan>>		
Usulan Pembimbing Akuntansi / Manajemen *) **)	**:**	1. <<kode Dosen – Nama Dosen>>		
		2. <<kode Dosen – Nama Dosen>>		
		3. <<kode Dosen – Nama Dosen>>		
		<<Sengaja dikosongkan>>		

HALAMAN PENILAIAN REVIEWER

Topik Final (Diisi oleh jurusan) :

Hasil Review Outline Skripsi (Diisi oleh Reviewer)

Tanggal Review	:	
Keputusan	:	Diterima / Revisi
Paraf Reviewer	:	

OUTLINE SKRIPSI

1. LATAR BELAKANG

Akibat majunya bidang teknologi di beberapa puluh tahun lalu, hadir juga banyak macam alat dan metode komunikasi yakni internet yang mempermudah masyarakat secara umum mengakses segalasektor industri *entertainment* baik dalam media musik, film, buku, atau permainan untuk mengadopsi model bisnis masing-masing agar mudah diakses secara digital melalui internet untuk meraup popularitas dan keuntungan yang lebih besar.

Seiring sejarah digitalisasi media entertainment, hal yang terus hadir dan dihadapi para produser konten adalah pembajakan. Pembajakan dalam konteks ini ialah aksi seseorang atau sebuah kelompok dalam mngi pembajakan, dan bahkan memiliki efek samping mempersulit akses konten bagi pengguna yang membeli produk secara legal.

Permasalahan yang diobservasi dari aktivitas dan penanggulangan pembajakan mengangkat beberapa pertanyaan bagi penulis, yaitu penyebab atas hadirnya keinginan untuk melakukan pembajakan dari k melindungi produknya dari pembajakan tanpa mengurangi kenyamanan akses penggunanya.

Sebagai upaya untuk menjawab pertanyaan-pertanyaan berikut, peneliti menggunakan salah satu platform komunikasi terbesar, yaitu Twitter untuk mengumpulkan opini para pengkonsumsi konten. Pemilihan platform Twitter didasari ata kelebihan berikut, selain user base yang besar, terdapat banyak dokumentasi yang telah dilakukan untuk membuktikan keefektifan sentiment analysis, layaknya sentimen analisis Twitter untuk Covid-19 oleh Manguri *et al*. pada tahun 2020, sentimen analisis Twitter pada *critical event* oleh Ruz *et al.* pada tahun 2020. Metode pengumpulan dataset untuk *tweet* juga memiliki dokumentasi dan pengujian dari banyak pihak, selebihnya penelitian sentimen analisis juga didukung oleh Twitter sendiri dalam bentuk bangunan API *software* Twitter yang memudahkan pengumpulan dan analisa data.

Dalam penelitian ini, algoritma Naive Bayes Classifier digunakan untuk mengidentifikasi dan mengkalkulasi hasil sentimen dari *tweets* para pengguna Twitter

atas topik pembajakan konten digital. Selebihnya, teknik *feature selection* akan diterapkan pada tahapan *preprocessing* untuk diaplikasikan pada algoritma Naive Bayes, menurut beberapa penelitian terdahulu layaknya Ernawati *et al.* pada tahun 2018 untuk analisis sentimen perusahaan fashion online atau Bashir *et al.* pada tahun 2019 pada prediksi penyakit jantung, *feature selection* memiliki dampak yang signifikan terhadap algoritma Naive Bayes dalam meningkatkan akurasi hasil prediksi.

Mereferensikan penelitian terdahulu, selebih menjawab beberapa pertanyaan atas sentimen pengguna Twitter tpkan teknik *feature selection* dalam penelitian sentimen analisis.

2. TUJUAN DAN MANFAAT

Tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

- Menjawab beberapa pertanyaan atas sentimen pengguna twitter terhadap pembajakan konten digital.
- Menguji keakuratan prediksi sentimen algoritma Naive Bayes berdasarkan *dataset* yang digunakan.
- Menguji keakuratan data Twitter yang digungan yang tidak menggunakan *feature selection.*

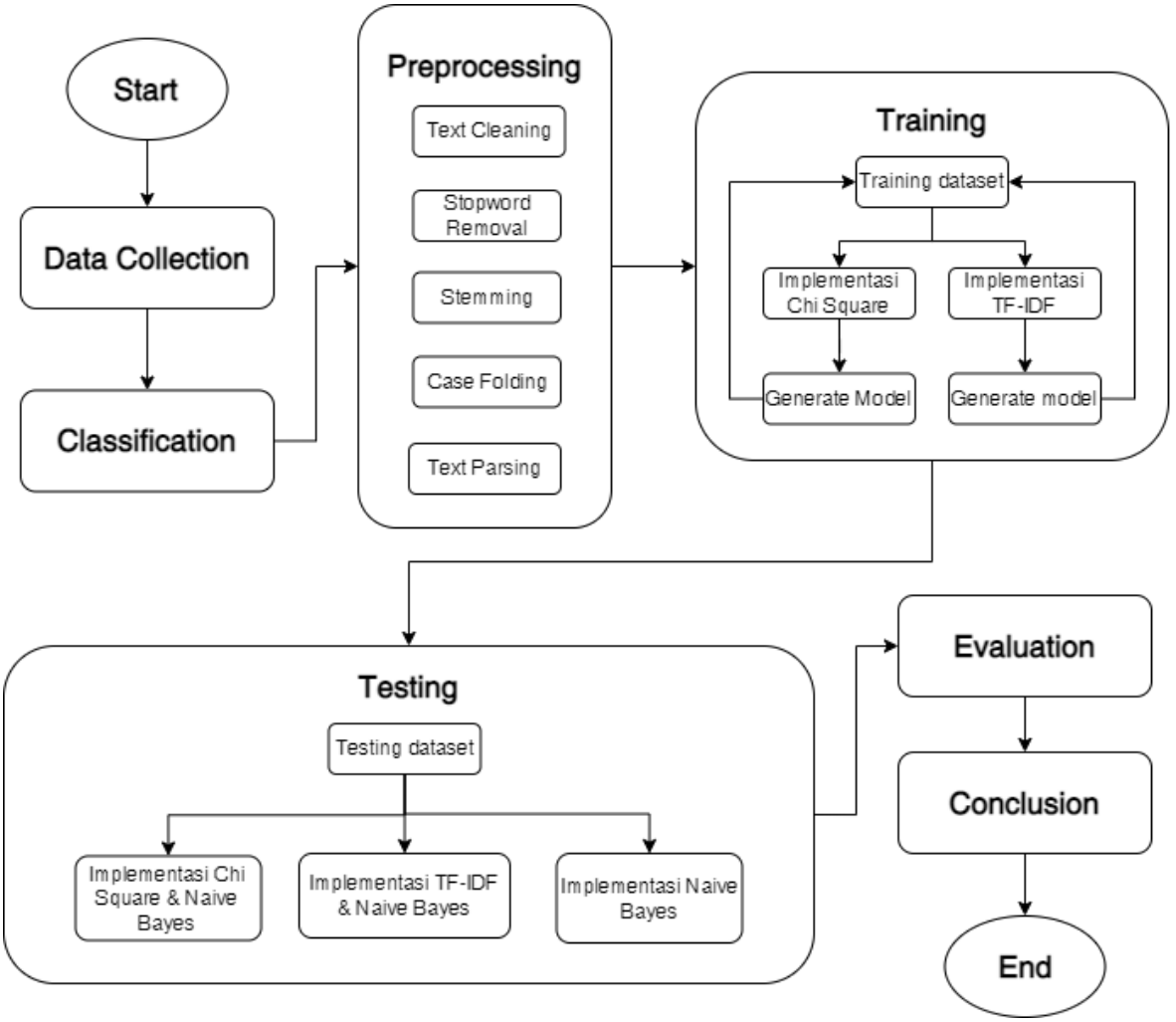
Manfaat yang diharapkan dapat diperoleh dari penelitian ini adalah sebagai berikut:

- Memberi kontribusi dalam bentuk pengujian algoritma Naive Bayes untuk sentimen analisis.
- Mendapat *insight* atas fenomena pembajakan konten digital yang hadir di era digitalisasi.
- Memberikan cara penanggulangan pembajakan yang sesuai dengan kebutuhan para pengkonsumsi konten.
- Memberikan akurasi keakuratan data Twitter untuk penggunaannya dalam penelitian sentimen anali

3. RUANG LINGKUP

Ruang lingkup penelitian adalah para penggumpulan sampel data untuk analisis sentimen akan diambil dari postingber 2021 (1 tahun) menggunakan bantuan dari *software data crawler* untuk Twitter.

4. METODOLOGI



Gambar 1. Skema metodologi penelitian sentimen analisis menggunakan metode *feature selection* dan algoritma Naive Bayes.

4.1 Data Collecting

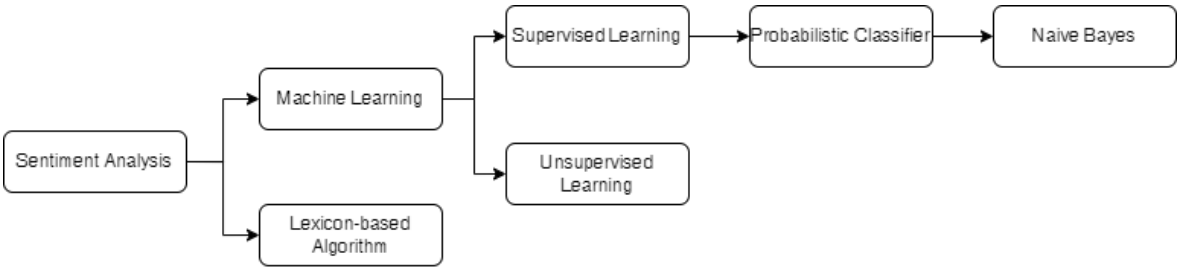
Tahap pertama dalam penelitian ini adalah mengumpulkan data *tweet* menggunakan Twitter scraper untuk setiap *tweet* dengan kata kunci yang berhubungan dengan "pembajakan" dan "piracy".

No	Teks

1	Normalisasi pembajakan konten digital nganter pesan yg buruk untuk ditiru generasi yG datang.
2	Pembajakan dpt mempreservasi game agar tdk ilang stlh berhenti dijual
3	PembAjaKan film diperlukan utk konten yang tidak tersedia di region tertentu.

Tabel 1. Contoh hasil koleksi data.

4.2 Klasifikasi Naive Bayes



Gambar 2. Skema Hierarki Naive Bayes (Manguri *et al*, 2020).

Dalam tahap klasifikasi atau *labeling*, kalimat yang telah melalui tahap *preprocessing* diberikan kategori yang menggambarkan perasaan atau sentimen para subjek terhadap topik "pembajakan" dan "*piracy*". Kategori sentimen yang ditetapkan dalam penelitian topik ini adalah "positif" atau "ne mengajar mesin mengidentifikasi kalimat berdasarkan label yang telah diberikan pada masing-masing kata dan menghasilkan prediksi akurat atas sentimen kalimat tersebut.

4.3 *Text Processing/Preprocessing*

4.3.1 *Text Cleaning*

Tahap memproses data yang terkumpul menjadi teks yang lebih mudah diidentifikasi, seperti menghapus data yang berulang, menghapus url, baris baru, tanda baca, dan simbol dalam kalimat.

4.3.2 *Stop-word removal*

Tahap menyeleksi kata-kata yanga karena kata-kata tersebut membawa nilai yang kecil untuk hasil analisis.

4.3.3 *Stemming*

Tahap menghilangkan infleksi kata ke bentuk dasar kata. Sebagai contoh kata "mempreservasi" setelah dilakukan proses stemming akan bertransformasi menjadi "preservasi".

4.3.4 *Case Folding*

Tahap mengubah semua bentuk huruf menjadi huruf kecil. Sebagai contoh kata "PembAjaKan" setelah dilakukan proses *case folding* akan bertransformasi menjadi "pembajakan".

4.3.5 *Text Parsing/Tokenization*

Tahap memproses dan membealimat menjadi string kata individu yang dapat dihubungkan dengan sebuah nilai dan dapat dihitung frekuensi kemunculannya dalam seluruh proses analisis.

No	Teks
1	normalisasi bajak konten digital antar pesan buruk tiru generasi datang
2	bajak dapat preservasi game hilang berhenti jual
3	bajakan film perlu konten tidak tersedia daerah

Tabel 2. Hasil data sebelum melalui *tokenization*.

No	Teks
1	[normalisasi, bajak, konten, digital, antar, pesan, buruk, tiru, generasi, datang]
2	[bajak, preservasi, game, hilang, berhenti, jual]
3	[bajak, film, perlu, konten, tersedia, daerah]

Tabel 3. Hasil data *tokenization*.

4.4 Feature Selection

Dalam melakukan tahap *preprocessing*, disertakan juga *feature selection*, *feature selection* dapat mengurangi beberapa nilai atau data yangngi beberapa fitur yang tidak diperlukan.

Caranya dilakukan lewat beberapa metode: *Filter Method, Wrapper Method, Embedded Method*. Pada penelitian ini kami menggunakan 2 algoritma, yaitu algoritma Chi-Square dan algoritma *Term Frequency Inverse Document Frequency* (TF-IDF).

Chi-Square merupakan metode pengujian yang banyak digunakan saat ini untuk mengevaluasi korelasi dengan menggunakan pendekatan statistika. Algoritma Chi-Square bekerja dengan menguji independensi sebuah term dan kategorinya sehingga algoritma ini dapat menghilangkan fitur -fitur pengganggu dan tidak relevan. Formula dari metode pengujian ini adalah sebagai berikut :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Gambar 3. Formula algoritma chi-square.

Pada pembobotan TF-IDF, bobot akan semakin besar jika frekuensi kemunculan kata semakin tinggi, tetapi bobot akan berkurang jika kata tersebi dokumen dengan rumus:

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

tf_{ij} = total number of occurences of i in j
df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

Gambar 4. Formula algoritma TF-IDF.

4.5 Training

Hasil dataset yang telah melalui tahap pre-processing akan diuji menggunakan *teknik feature selection* lewat 2 algoritma, yakni chi-square dan TF-IDF. Tahap pelatihan ini akan dilakukan sebanyak 3 kali dan di dalam prosesnya beberapa feature yang tidak relevan akan diseleksi oleh algoritma dan dipotong untuk meningkatkan akurasi algoritma prediksi.

4.6 Testing

Hasil data training *feature selection* chi-square dan TF-IDF dikombinasikan dengan algoritma Naive Bayes untuk memulai proses sentimen analisis, .selebihnya hasil data yang telah melalui preprocessing namun tidak melewati proses *feature selection* diikutsertakan untuk menguji perbedaan hasil akurasi algoritma.

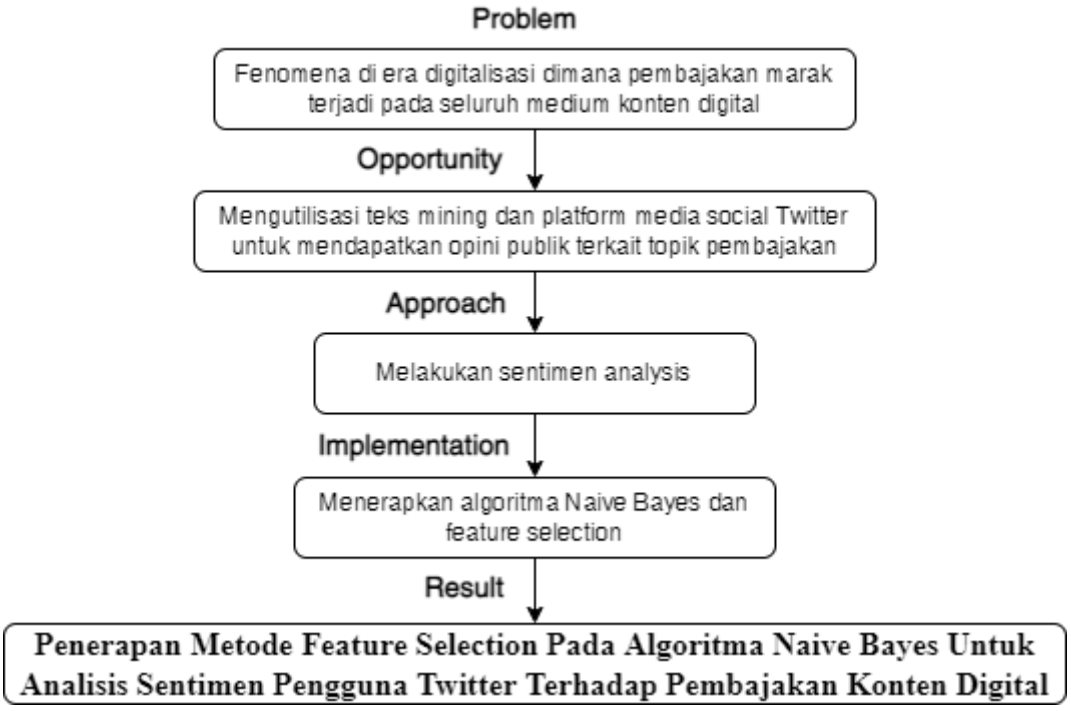
4.7 Evaluation

Evaluasi hasil sentimen untuk mengetahui performa dari pengelompokkan menggunakan Textblob untuk analisis sentimen. Perhitungan kinerkan menggunakan *confusion matrix* dengan menghitung akurasi, presisi, *recall*, dan F1 *score*. Hasil evaluasi ini akan menjadi gambaran seberapa akurat dari vader dalam mengelompokkan sentimen pada *tweet*.

4.8 Conclusion

Melalui hasil evaluasi yang diperoleh, kami sebagai peneliti menarik kesimpulan terkait kinerja dari metode Feature Selection dan Naïve Bayes dalam melakukan ekstraksi pengelompokan sentimen pada tweet.

5. KERANGKA PIKIR



Gambar 5. Skema masalah penelitian.

Proses penelitian ini ditujukan untuk menganalisis ruang lingkup aplikasi Twitter dengan topik yang berhubungan dengan pembajakan konten digital. Pengetahuan yang didapatkan atas hasil analisis tersebut akan digunakan untuk menguji keakuratan metode bayes dalam memprediksi analisis sentimk sentimen positif atau negatif, diikuti dengan seberapa kuat kandungan sentimen tersebut untuk diterapkan pada algoritma Naive Bayes. Untuk meningkatkan akurasi prediksi algoritma Naive Bayes pada analisis sentimen Twitter, peneliti juga ikut mengimplementasi *feature selection* yang diketahui memiliki dampak positif terhadap hasil akurasi, dampak ini akan diukur signifikansinya dengan hasil yang tidak menggunakan implementasi teknik *feature selection* untuk memberi estimasi atas perubahan hasil akurasi yang dapat diekspektasikan atas penggunaan teknik tersebut.

6. OBJEK PENELITIAN

Objek penelitian yang ingin dibahas lebih lanjut dalam penelitian ini adalah sentimen para pengguna Twitter terkait topik pembajakan. Menggunakan metode bayes, peneliti

akan menganalisis *tweet* akun penang untuk mendukung aksi pembajakan konten digital serta mengetahui metode apa yang dapat dilakukan pembuat atau *publisher* untuk melindungi kontennya dari pembajakan tanpa mengganggu kemudahan atau kenyaman aksesibilitas konten bagi para pengguna legal.

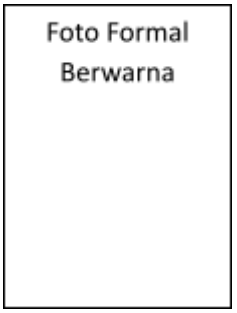
7. PERAN DAN TANGGUNG JAWAB MAHASISWA

Tabel 4. Peran dan Tanggung Jawab Mahasiswa

No	NIM	Nama	Program Studi	Alokasi Waktu	Uraian Tugas	Output
1	2301852051	Jeremy Jason	Sistem Informasi	10 Hari	Menentukan latar belakang, tujuan, manfaat dan metodologi penelitian	Latar belakang, tujuan, manfaat & metodologi penelitian
2	2301860974	Stephen Tanjiro Suwandy	Sistem Informasi	10 Hari	Menyusun tujuan, manfaat, dan metodologi penelitian	Tujuan, manfaat, & metodologi penelitian
3	2301882400	Jessen Mailovy Gunawan	Sistem Informasi	10 Hari	Menyusun kerangka latar belakang, menentukan tujuan dan manfaat penelitian, menentukan ruang lingkup penelitian,	Latar belakang, tujuan & manfaat penelitian, ruang lingkup, metodologi, kerangka pikir, &

					metode penelitian, kerangka pikir, dan objek penelitian.	objek penelitian.

DATA DIRI MAHASISWA



NIM	**:**	**2301852051**
Nama Lengkap	**:**	**Jeremy Jason**
Program	**:**	**Sistem Informasi**
Peminatan (Jika Ada)	**:**	**Business Intelligence**
IPK	**:**	
No HP	**:**	
Email	**:**	
Alamat Domisili		
Poin SAT Saat ini	**:**	
Jumlah Jam Community Service saat ini	**:**	

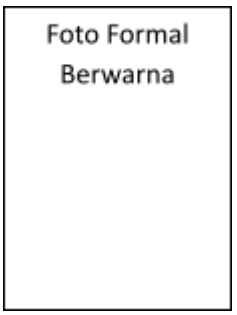
Riwayat Organisasi (UKM / HMJ)

Periode Waktu	Posisi / Jabatan	Nama UKM / HMJ
2019 - Sekarang	**Member**	**UKM Badminton BINUS**

Riwayat Pekerjaan :

Periode Waktu	Posisi / Jabatan	Nama Perusahaan
2022 - 2023	**Business Analyst Intern**	**PT. Hexaon Business Mitrasindo**

DATA DIRI MAHASISWA



NIM	**:**	**2301860974**
Nama Lengkap	**:**	**Stephen Tanjiro Suwandy**
Program	**:**	**Sistem Informasi**
Peminatan (Jika Ada)	**:**	**Business Intelligence**
IPK	**:**	**3.34**
No HP	**:**	**081342102680**
Email	**:**	**stephen.suwandy@binus.ac.id**
Alamat Domisili	**:**	**Villa Taman Bandara A7/17, RT01/RW08, Dadap, Kosambi, Tangerang, 15211**
Poin SAT Saat ini	**:**	**131**
Jumlah Jam Community Service saat ini	**:**	**40**

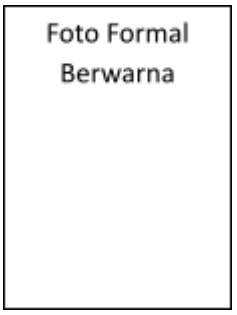
Riwayat Organisasi (UKM / HMJ)

Periode Waktu	Posisi / Jabatan	Nama UKM / HMJ
2019 - Sekarang	**Member**	**UKM Badminton BINUS**

Riwayat Pekerjaan :

Periode Waktu	Posisi / Jabatan	Nama Perusahaan
2022 - 2023	**Functional Consultant Intern**	**PT. Sterling Tulus Cemerlang**

DATA DIRI MAHASISWA



NIM	**:**	**2301882400**
Nama Lengkap	**:**	**Jessen Mailovy Gunawan**
Program	**:**	**Sistem Informasi**
Peminatan (Jika Ada)	**:**	**Business intelligence**
IPK	**:**	**3.59**
No HP	**:**	**08111502310**
Email	**:**	**jessen.gunawan@binus.ac.id**
Alamat Domisili	**:**	**Perumahan Permata Kranggan, Cipayung, Jakarta 17433, Indonesia.**
Poin SAT Saat ini	**:**	**147**
Jumlah Jam Community Service saat ini	**:**	**40**

Riwayat Organisasi (UKM / HMJ)

Periode Waktu	Posisi / Jabatan	Nama UKM / HMJ

Riwayat Pekerjaan :

Periode Waktu	Posisi / Jabatan	Nama Perusahaan