BAB IV
ANALISA DAN PERANCANGAN

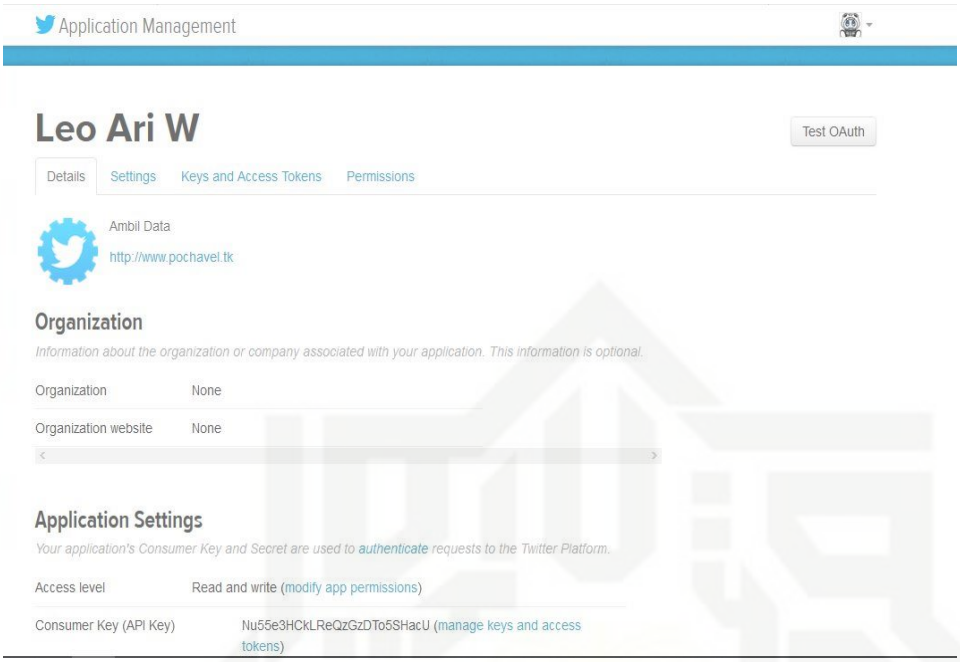
4.1 Analisa

Pada Bab ini menjelaskan tahapan – tahapan yang dilakukan pada sistem yaitu bagaimana proses pengumpulan data, analisa proses NER dan penerapan metode k-NN serta contoh penyelesaian masalah dalam pengenalan entitas yang akan dijelaskan sebagai berikut.

4.1.1 Pengumpulan Data

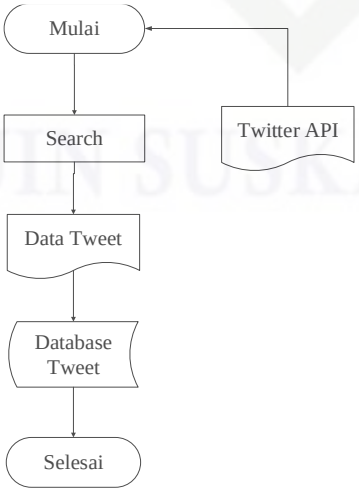
Penelitian ini menggunakan data berupa *tweet* dari para pengguna twitter. *Tweet* yang digunakan pada penelitian ini berjumlah 1000 *tweet* yang merupakan *tweet* iklan dengan rincian 900 data latih dan 100 data uji. Untuk mendapatkannya maka digunakan *Application Programming Interface* (API) Twitter yang diberikan oleh pihak Twitter bagi para pengembang teknologi informasi. API twitter ini nantinya akan disiapkan dalam *script* PHP agar memperoleh data yang akan digunakan dalam penelitian ini.

Untuk mendapatkan API twitter, langkah pertama yang dilakukan adalah dengan mendaftarkan aplikasi di situs apps.twitter.com. Proses pendaftaran dapat dilakukan jika sudah memiliki akun twitter yang sudah tertaut nomor ponsel dan email pendaftaran yang sudah dikonfirmasi pemilik akun. Twitter kemudian memberikan form pendaftaran aplikasi yang wajib diisi. Gambar 4.1 memperlihatkan *Aplication Management* pada situs apps.twitter.com untuk pendaftaran Twitter Apps.



Gambar 4.1 Tampilan Halaman Aplikasi Management

Proses pengunduhan data *tweet* dilakukan dengan *crawling* Twitter dengan teknik *search*. Teknik ini menggunakan satu atau beberapa kata kunci (*keyword)* untuk mengumpulkan data. Pengunduhan data dilakukan dengan membangun program dari *script* PHP dengan menyisipkan informasi aplikasi dari API Twitter dalam *script* tersebut. Program akan mengambil data *tweet* yang berasal dari basis data Twitter melalui API Twitter. Berikut Gambar 4.2 *flowchart* proses pengunduhan data.



Gambar 4.2 Proses Pengunduhan Data

Selain API Twitter, digunakan pula *twitter aoutentification* yang berfungsi untuk mengakses ke API Twitter. *OAuth* adalah sebuah *authorization framework* yang memungkinkan aplikasi pihak ketiga untuk mendapatkan akses terbatas secara aman dan ringkas. Dengan *OAuth*, untuk melakukan *request* ke API Twitter, setiap aplikasi harus terlebih dahulu mendapatkan *OAuth* akses token. Akses token ini yang kemudian digunakan ketika menuliskan kode program.

Data *tweet* yang telah diunduh menggunakan API Twitter akan disimpan kedalam basis data. Selanjutnya dilakukan seleksi data dari *tweet* yang telah disimpan tadi untuk menghindari adanya *tweet* yang ganda dan *tweet* yang tidak berisi informasi iklan. Adapun *tweet* yang digunakan adalah *tweet* iklan dari akun Twitter promosi iklan. Berikut nama-nama akun yang akan digunakan adalah sebagai berikut.

1. @Iklan_masadepan
2. @FJB_Bandoeng
3. @postingiklan
4. @tempat_promosi
5. @InfoJual_Beli
6. @jualbeli_klaten
7. @Iklan_ProdukTOP
8. @iklan123
9. @fjubel
10. @iklan_jualan

Pada Tabel 4.1 diperlihatkan beberapa contoh data yang berhasil di unduh.

Tabel 4.1 Hasil pengunduhan data *tweet*

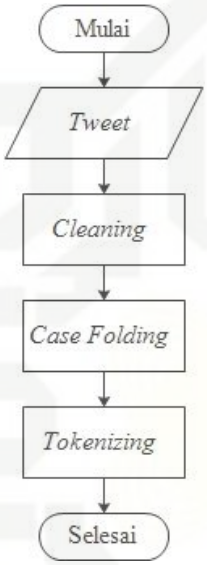
No	Id	user_id	Text
1	808144389621055000	1447874514	Rencana Beli Mobil Honda? DP & Angsuran Ringan, PROSES Gak Ribet? Follow @ahmadhonda / 081282218485 #BelanjaSeru. https://t.co/KwFgjitrot

No	Id	user_id	Text
2	808153166109929000	1447874514	Distributor Sepatu Murah. Follow IG @sultanshoes | WA: 08119200044 https://t.co/YQ2MtDIZSG
3	800696578319257000	1447874514	Supplier baju dari Pabriknya. 50 ribuan aja Sista. Follow Instagram @rainbowboutiq | PIN:268E2739. https://t.co/NtLP7CFXDG
4	818033272215207030	1447874514	Supplyer/Gudang Jaket kaos distro/sport termurah di bandung @denigshop BBM: 2A1EC8D2 /08987037959 / IG: denigshop http://t.co/znFPSKAiW4
5	567546536167759000	730048591	Baju Anak Branded Murah Berkualitas. Mulai Harga 35 Ribuan. Follow IG @digy_store | WA: 081239745471. https://t.co/uW9uHkY5zU
6	150171977871654400	394592985	Grosir Liquid Rokok Elektrik Flavors 35 ribuan, Ejuice 50 ribuan, V2/Blaze 25 ribuan, Noboru 50 ribuan @AksesorisCenter 08997171905 http://t.co/xFTajofhX2
7	808144389621054500	1447874514	Jam tangan BabyG, 125 ribuan @blues_shop 0896562223456 pin:27D6ACEE http://t.co/usUVhEsj"

No	Id	user_id	Text
8	808153166109924200	1447874514	Grosir Modem GSM ZTE MF190 140 ribuan, Huawei E3276 LTE 240 ribuan, WiFi, Perdana Internet | @AksesorisCenter 08997171905 http://t.co/T1LADMkRmA
9	800696578319223020	730048591	Celana Dalam Wanita Branded Harga Grosir. Mulai Harga 10 Ribuan. Follow IG @brasisterofficial WA:+6281316355321. https://t.co/lZ7tndImvP
10	818033272215207020	394592985	Sepatu Adidas Superstar Cuma 175 Ribuan Jogger Pants Cuma 100 Ribuan. Follow IG @an_nishop | BBM: D34CF489. https://t.co/NEaI9tlUil
11	567546536167359000	730048591	@Twinky_Shop Sepatu reebok 200 ribuan info bbm 7648010D @PromoIklanjual @akunpromoID http://t.co/N3xcW3OS6q
12	567546536157759000	394592985	SCARF Cantik Mulai Harga 25 Ribuan. Follow IG @kalsproject LINE @xsp8496i https://t.co/SMaITTFYfF

4.1.2 *Preprocessing*

Pada penelitian ini proses *preprocessing* dilakukan dengan membangun *script* PHP. Proses ini bertujuan untuk membersihkan data dari hal-hal yang tidak diperlukan dan menyederhanakan dokumen *tweet* untuk proses *Named Entity Recognition* (NER). Berikut Gambar 4.3 *flowchart preprocessing* yang dilakukan pada penelitian ini.



Gambar 4.3 *Flowchart Preprocessing*

1. *Cleaning*

Setelah data *tweet* berhasil diunduh maka selanjutnya dilakukan proses pembersihan (*cleaning*) terhadap sejumlah karakter seperti (!@$%^&*()-={}|\/.,;’) link, dan hastag. Hal ini dilakukan untuk penyederhanaan data dan menghindari data duplikat. Penghilangan data duplikat terkendala karena adanya spam yang terdapat pada twitter. *Tweet-tweet* spam merupakan data duplikat yang tidak terdeteksi karena link yang dilampirkan berbeda. Untuk itu dalam penelitian ini proses pembersihan dilakukan pada proses pengunduhan data agar lebih efektif dalam menentukan dataset. Berikut Tabel 4.2 hasil proses *cleaning* pada dokumen *tweet.*

Tabel 4.2 Proses *Cleaning*

Dokumen 1	Rencana Beli Mobil Honda DP Angsuran Ringan PROSES Gak Ribet Follow @ahmadhonda 081282218485 BelanjaSeru

Dokumen 2	Distributor Sepatu Murah Follow IG @sultanshoes WA 08119200044
Dokumen 3	Supplier baju dari Pabriknya 50 ribuan aja Sista Follow Instagram @rainbowboutiq PIN 268E2739
Dokumen 4	Supplyer Gudang Jaket kaos distro sport termurah di bandung @denigshop BBM 2A1EC8D2 08987037959 IG denigshop
Dokumen 5	Baju Anak Branded Murah Berkualitas Mulai Harga 35 Ribuan Follow IG @digy_store WA 081239745471
Dokumen 6	Grosir Liquid Rokok Elektrik Flavors 35 ribuan Ejuice 50 ribuan V2 Blaze 25 ribuan Noboru 50 ribuan @AksesorisCenter 08997171905
Dokumen 7	Jam tangan BabyG 125 ribuan @blues_shop 0896562223456 27D6ACEE
Dokumen 8	Grosir Modem GSM ZTE MF190 140 ribuan Huawei E3276 LTE 240 ribuan WiFi Perdana Internet @AksesorisCenter 08997171905
Dokumen 9	Celana Dalam Wanita Branded Harga Grosir Mulai Harga 10 Ribuan Follow IG @brasisterofficial WA +6281316355321
Dokumen 10	Sepatu Adidas Superstar Cuma 175 Ribuan Jogger Pants Cuma 100 Ribuan Follow IG @an_nishop BBM D34CF489
Dokumen 11	@Twinky_Shop Sepatu reebok 200 ribuan info bbm 7648010D @PromoIklanjual @akunpromoID
Dokumen 12	SCARF Cantik Mulai Harga 25 Ribuan Follow IG @kalsproject LINE @xsp8496i

2. *Case Folding*

Case folding adalah mengubah semua huruf kapital dalam dokumen menjadi huruf kecil. Proses *case folding* diperlukan untuk mengatasi kata atau *term* ganda hanya karena penulisan katanya yang tidak sama. Oleh karena itu, untuk meratakan semua kata maka diubah semua huruf kapital kebentuk huruf kecil. Berikut Tabel 4.3 hasil proses *case folding* pada dokumen *tweet*.

Tabel 4.3 Proses *case folding*

Dokumen 1	rencana beli mobil honda dp angsuran ringan proses gak ribet follow @ahmadhonda 081282218485 belanjaseru
Dokumen 2	distributor sepatu murah follow ig @sultanshoes wa 08119200044
Dokumen 3	supplier baju dari pabriknya 50 ribuan aja sista follow instagram @rainbowboutiq pin 268e2739
Dokumen 4	supplyer gudang jaket kaos distro sport termurah di bandung @denigshop bbm 2a1ec8d2 08987037959 ig denigshop
Dokumen 5	baju anak branded murah berkualitas mulai harga 35 ribuan follow ig @digystore wa 081239745471
Dokumen 6	grosir liquid rokok elektrik flavors 35 ribuan ejuice 50 ribuan v2 blaze 25 ribuan noboru 50 ribuan @aksesoriscenter 08997171905
Dokumen 7	jam tangan babyg 125 ribuan @bluesshop 0896562223456 27d6acee
Dokumen 8	grosir modem gsm zte mf190 140 ribuan huawei e3276 lte 240 ribuan wifi perdana internet @aksesoriscenter 08997171905
Dokumen 9	celana dalam wanita branded harga grosir mulai harga 10 ribuan follow ig @brasisterofficial wa +6281316355321
Dokumen 10	sepatu adidas superstar cuma 175 ribuan jogger pants cuma 100 ribuan follow ig @annishop bbm d34cf489
Dokumen 11	@twinkyshop sepatu reebok 200 ribuan info bbm 7648010d @promoiklanjual @akunpromoid
Dokumen 12	SCARF Cantik Mulai Harga 25 Ribuan Follow IG @kalsproject LINE @xsp8496i

3. *Tokenizing*

Tokenizing adalah proses pemisahan kalimat menjadi kata perkata atau berupa potongan kata tunggal. Berikut Tabel 4.4 hasil proses *tokenizing* pada dokumen *tweet.*

Tabel 4.4 Proses *Tokenizing*

Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Dokumen 10	Dokumen 11	Dokumen 12
rencana	distributor	Supplier	supplyer	baju	grosir	Jam	grosir	celana	sepatu	@twinky_shop	SCARF
beli	sepatu	Baju	gudang	anak	liquid	Tangan	modem	wanita	adidas	sepatu	Cantik
mobil	murah	Pabrik	jaket	branded	rokok	Baby	gsm	branded	superstar	reebok	Mulai
honda	follow	50	kaos	murah	elektrik	125	Zte	harga	cuma	200	Harga
dp	ig	Ribuan	distro	berkualitas	flavors	Ribuan	mf190	grosir	175	ribuan	25
angsuran	@sultanshoes	aja	sport	mulai	35	@blues_shop	140	mulai	ribuan	info	Ribuan
ringan	wa	sista	termurah	harga	ribuan	896562E	ribuan	harga	jogger	bbm	Follow
proses	8119200044	follow	bandung	35	ejuice	Pin	huawei	10	pants	7648010d	IG
follow	@kabirashop	instagram	@denigshop	ribuan	50	27d6acee	e3276	ribuan	cuma	@promoikljanjual	@kalsproject
@ahmadhonda	wa	@rainbowboutiq	bbm	follow	ribuan		Lte	follow	100	@akunpromoid	LINE
81282218485	081220656965	pin	2a1ec8d2	ig	v2		240	Ig	ribuan		@xsp8496i
belanjaseru		268e2739	8987037959	@digy_store	blaze		ribuan	@brasisterofficial	follow		
			Ig	wa	25		wifi	Wa	ig		

Dokumen 1	Dokumen 2	Dokumen 3	Dokumen 4	Dokumen 5	Dokumen 6	Dokumen 7	Dokumen 8	Dokumen 9	Dokumen 10	Dokumen 11	Dokumen 12
			denigshop	81239745471	ribuan		perdana	+6281316355321	@an_nishop		
					noboru		internet		bbm		
					50		@aksesoris center		d34cf489		
					ribuan		8997171905				
					@aksesoris center						
					8997171905						

4.1.3 POS *Tagging*

Setelah proses *preprocessing* kemudian dilakukan proses pemberian POS *Tagging* pada kata untuk menandai kelas kata pada dokumen *tweet*. POS *Tagging* dilakukan dengan memberi tagset sebagaimana Tabel 2.1. Pemberian POS *Tagging*. Pada proses POS Tagging kata yang sudah di tokenisasi dilakukan pengecekan kata pada kamus kateglo dan untuk kata yang tidak terdapat dikamus akan diberi tag (XX). Adapun jenis kelas yang digunakan adalah *Nomina* (N), *Numerelia* (NUM), *Preposisi* (PRE), *Pronomina* (PRO), *Verba* (V), *Adverbial* (ADV), *Konjungsi* (K), *Interjeksi* (I), *Adjektiva* (ADJ) dan Kata asing (FW). Berikut Gambar 4.4 *flowchart* alur proses POS *Tagging*.

Gambar 4.4 *Flowchart* POS *Tagging*

Berikut penjelasan dari Gambar 4.4 *flowchart* POS *Tagging*.

1. Lakukan *preprocessing* pada data *tweet* yang telah diunduh
2. Proses POS *Tagging* dilakukan dengan pengecekan kata pada kamus dan dilihat kelas kata pada kamus tersebut.

3. Apabila kelas katanya ada pada kamus beri tanda kata sesuai dengan tag dikamus dan apabila tidak ada kata dikamus maka beri tanda XX

Tabel 4.5 hasil proses POS *Tagging* pada dokumen *tweet.*

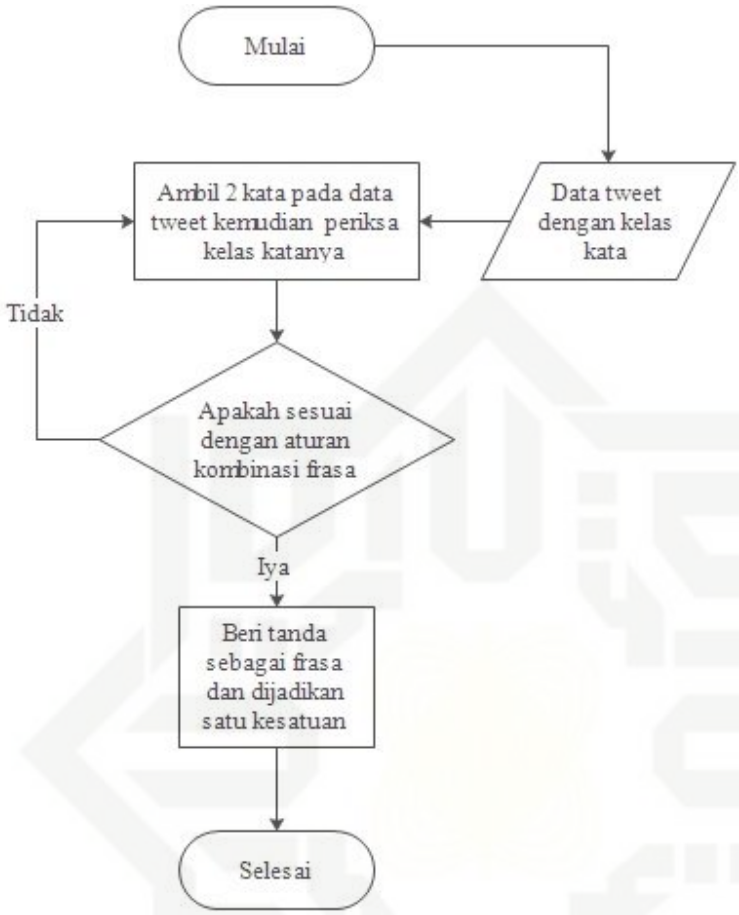
Tabel 4.5 Proses POS *Tagging*

Dokumen 1	rencana|N beli|V mobil|N honda|N dp|XX angsuran|N ringan|ADJ proses|N follow|FW @ahmadhonda165|XX 081282218485|NUM belanjaseru|XX
Dokumen 2	distributor|N sepatu|N murah|ADJ follow|FW ig|XX @sultanshoes|XX wa|XX 08119200044|NUM
Dokumen 3	supplier|XX baju|N pabrik50|NUM ribuan|N aja|N sista|XX follow|FW instagram|XX @rainbowboutiq|XX pin|N 268e2739|NUM
Dokumen 4	supplyer|XX gudang|N jaket|N kaos|N distro|XX sport|FW termurah|ADJ bandung|N @denigshop|XX bbm|N 2a1ec8d2|NUM 08987037959|NUM ig|XX denigshop|XX
Dokumen 5	baju|N anak|N branded|XX murah|ADJ berkualitas|V mulai|V harga|N 35|NUM ribuan|N follow|FW ig|XX @digy_store|XX wa|XX 081239745471|NUM
Dokumen 6	grosir|N liquid|FW rokok|N elektrik|N flavors|XX 35|NUM ribuan|N ejuice|XX 50|NUM ribuan|N v2|NUM blaze|XX 25|NUM ribuan|N noboru|XX 50|NUM ribuan|N @aksesoriscenter|XX 08997171905|NUM
Dokumen 7	jam|N tangan|N babyg|XX 125|NUM ribuan|N @blues_shop|XX 0896562223456|NUM pin|N 27d6acee|NUM
Dokumen 8	grosir|N modem|N gsm|XX zte|XX mf190|NUM 140|NUM ribuan|N huawei|XX e3276|NUM lte|XX 240|NUM ribuan|N wifi|XX perdana|N internet|FW @aksesoriscenter|XX 08997171905|NUM

Dokumen 9	celana|N wanita|N branded|XX harga|N grosir|N mulai|V harga|N 10|NUM ribuan|N follow|FW ig|XX @brasisterofficial|XX wa|XX +6281316355321|NUM
Dokumen 10	sepatu|N adidas|XX superstar|N cuma|ADV 175|NUM ribuan|N jogger|XX pants|FW cuma|ADV 100|NUM ribuan|N follow|FW ig|XX @an_nishop|XX bbm|N d34cf489|NUM
Dokumen 11	@twinky_shop|XX sepatu|N reebok|XX 200|NUM ribuan|N info|N bbm|N 7648010d|NUM @promoiklanjual|XX @akunpromoid|XX
Dokumen 12	scarf|XX cantik|ADJ mulai|V harga|N 25|NUM ribuan|N follow|FW ig|XX @kalsproject|XX line|XX @xsp8496i|NUM

4.1.4 *Chunking*

Chunking adalah metode untuk mengidentifikasi frasa dalam teks. Langkah ini dilakukan untuk mendeteksi adanya dua kata atau lebih yang menjadi frasa. Proses chunking dilakukan setelah kata dilakukan proses POS Tagger, kata yang telah diberi kelas kata dilakukan proses mendeteksi kata menjadi frasa. Pendekatan *chunking* yang digunakan pada penelitian ini berbasis aturan. Proses pendeteksian chunking dilakukan dengan dengan pengecekan kata yang telah diberi kelas kata dengan aturan yang telah ditentukan dan kombinasi kelas kata akan dijadikan frasa. Jenis frasa merujuk Tabel 2.2 terdapat 5 jenis frasa yaitu *Frasa Verbal* (FV), *Frasa Nomina* (FN), *Frasa Adverbial* (FADV) dan *Frasa Pronominal* (FPRO) dengan 16 aturan frasa. Berikut Gambar 4.5 *flowchart* alur proses *chunking*



Gambar 4.5 *Flowchart* Proses *Chunking*

Berikut penjelasan dari Gambar 4.5 *flowchart* proses *chunking.*

1. Setelah dilakukan proses POS *Tangging* diperoleh kata pada data *tweet* yang telah diberi kelas katanya
2. Kata yang sudah diberi kelas kata di lakukan pengecekan 2 kata apabila memenuhi syarat dilakukan pemberian *Tag* pada 2 kata tersebut dengan satu kesatuan dan di beri tanda *Tag* frasa (*chunking*) dan apabila tidak ada dilakukan pengecekan kata berikutnya.

Tabel 4.6 hasil proses *chunking* pada dokumen *tweet*

Tabel 4.6 Proses *Chunking* pada Dokumen Tweet.

Dokumen 1	rencana|FN beli|FV mobil honda|FN dp|XX angsuran ringan|FN proses|FN follow|FW @ahmadhonda165|XX 081282218485|NUM belanjaseru|XX

Dokumen 2	distributor sepatu|FN murah|FADJ follow|FW ig|XX @sultanshoes|XX wa|XX 08119200044|NUM
Dokumen 3	supplier|XX baju pabrik|FN 50 ribuan|FN aja|FN sista|XX follow|FW instagram|XX @rainbowboutiq|XX pin|FN 268e2739|NUM
Dokumen 4	supplyer|XX gudang jaket|FN kaos|FN distro|XX sport|FW termurah|FADJ bandung|FN @denigshop|XX bbm|FN 2a1ec8d2|XX 08987037959|NUM ig|XX denigshop|XX
Dokumen 5	baju anak|FN branded|XX murah|FADJ berkualitas|FV mulai harga|FV 35 ribuan|FN follow|FW ig|XX @digy_store|XX wa|XX 081239745471|NUM
Dokumen 6	grosir|FN liquid|FW rokok elektrik|FN flavors|XX 35 ribuan|FN ejuice|XX 50 ribuan|FN v2|XX blaze|XX 25 ribuan|FN noboru|XX 50 ribuan|FN @aksesoriscenter|XX 08997171905|NUM
Dokumen 7	jam tangan|FN babyg|XX 125 ribuan|FN @blues_shop|XX 0896562223456 pin|FN 27d6acee|XX
Dokumen 8	grosir modem|FN gsm|XX zte|XX mf190|XX 140 ribuan|FN huawei|XX e3276|NUM lte|XX 24 ribuan|FN wifi|XX perdana|FN internet|FW @aksesoriscenter|XX 08997171905|NUM
Dokumen 9	celana wanita|FN branded|XX harga grosir|FN mulai harga|FV 10 ribuan|FN follow|FW ig|XX @brasisterofficial|XX wa|XX +6281316355321|NUM
Dokumen 10	sepatu|FN adidas|XX superstar cuma|FN 175 ribuan|FN jogger|XX pants|FW cuma|FADV 100 ribuan|FN follow|FW ig|XX @an_nishop|XX bbm|FN d34cf489|XX
Dokumen 11	@twinky_shop|XX sepatu|FN reebok|XX 200 ribuan|FN info bbm|FN 7648010d|NUM @promoiklanjual|XX @akunpromoid|XX

Dokumen 12	scarf|XX cantik|FADJ mulai harga|FV 25 ribuan|FN follow|FW ig|XX @kalsproject|XX line|XX @xsp8496i|NUM

4.1.5 *Tag* BIO

BIO (*Beginning, Inside, Outside*) adalah format pemberian tag yang umum untuk menandai beberapa token yang merupakan bagian frasa, pada format BIO dimana B-*begin* dan I-*inside* menunjukkan token milik Entitas Bernama dan "O-*outside*" digunakan untuk semua token. Setelah proses menentukan frasa, BIO bertugas untuk memberikan tag yang masih dalam satu frasa, B-*begin* merupakan tag untuk awal kata dan I-*inside* merupakan tag untuk kata selanjutnya yang masih termasuk kedalam frasa tersebut. Contoh pemberian tag BIO sebagai berikut.

1. distributor sepatu|FN = distributor|*B*-FN sepatu|*I*-FN
2. baju pabrik|FN = baju|B-FN pabrik|*I*-FN
3. 50 ribuan|FN = 50|*B*-FN ribuan|*I*-FN

4.1.6 Pembobotan TF-IDF

Tahapan pembobotan TF-IDF adalah proses mengubah kata menjadi bentuk vektor. Pada proses pembobotan TF-IDF dilakukan perhitungan bobot pada kata, tag dan frasa. Tahapan – tahapan yang dilakukan pada proses pembobotan adalah menghitung *document frequency* (df), *term frequency* (tf), *inverse document frequency* (idf). Setelah selesai menghitung df, tf, idf selanjutnya mengalikan tf dengan idf sebagai bobot dari kata. Untuk lebih jelasnya dapat dilihat pada Tabel 4.7 berikut.

Tabel 4.7 Pembobotan TF-IDF Kata

Kata	TF												df	Log(n/df)	IDF									
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12			D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
rencana	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
beli	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
mobil	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
honda	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
dp	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
angsuran	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
ringan	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
proses	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
follow	1	1	1	0	1	0	0	0	1	1	0	1	6	0,30103	0,30103	0,30103	0,30103	0	0,301	0	0	0	0,3	0,30103	0	0,3
ahmadhonda	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
81282218485	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0
belanjaseru	1	0	0	0	0	0	0	0	0	0	0	0	1	1,079181	1,07918	0	0	0	0	0	0	0	0	0	0	0

Tabel 4.8 Pembobotan TF IDF Kelas Kata

Kelas Kata	TF												DF	Log(n/df)	$w_{dt} = tf_{dt} * IDF_t$									
	XI	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11			X1	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
Nomina	2	5	2	4	5	4	7	4	5	6	5	4	51	-0,6283889	-1,256777	-3,1419447	-1,2567779	-2,5135557	-3,1419447	-2,5135557	-4,3987225	-2,513555	-3,1419447	-3,7703336	-3,1419447	-2,513555
Verb	1	1	0	0	0	2	0	0	0	1	0	0	4	0,4771215	0,477121	0,4771215	0	0	0	0,9542425	0	0	0	0,4771215	0	0
Adverbia	0	0	0	0	0	0	0	0	0	0	2	0	2	0,7781515	0	0	0	0	0	0	0	0	0	0	1,5563025	0
Adjectiva	1	1	1	0	1	1	0	0	0	0	0	0	4	0,4771215	0,477121	0,4771215	0,4771215	0	0,4771215	0,4771215	0	0	0	0	0	0
FW	1	1	1	1	1	1	1	0	1	1	2	0	10	0,0791812	0,079181	0,0791812	0,0791812	0,0791812	0,0791812	0,0791812	0,0791812	0	0,0791812	0,0791812	0,1583624	0
XX	4	3	3	4	5	4	6	1	6	4	4	4	48	-0,60206	-2,40824	-1,80618	-1,80618	-7,2247199	-9,0308999	-2,40824	-3,6123599	-0,60206	-3,6123599	-2,40824	-2,40824	-2,40824
Numerelia	2	1	1	2	2	2	5	3	5	2	3	2	28	-0,3679768	-0,735953	-0,3679768	-0,3679768	-0,7359536	-0,7359536	-0,7359536	-1,8398839	-1,103934	-1,8398839	-0,7359536	-1,1039304	-0,735953

Tabel 4.9 Pembobotan TF IDF Frasa

Kelas Kata	TF												DF	Log(n/df)	$w_{dt} = tf_{dt} * IDF_t$									
	XI	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11			X1	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
FN	2	4	1	4	4	2	5	3	4	3	4	3	39	-0,5118834	-1,0237667	-2,0475334	-0,5118834	-2,0475334	-2,0475334	-1,0237667	-2,5594168	-1,535650 1	-2,0475334	-1,5356501	-2,0475334	-1,535650 1
FV	2	1	0	0	0	2	0	0	0	1	0	0	6	0,30103	0,60206	0,30103	0	0	0	0,6020599	0	0	0	0,30103	0	0
FADJ	1	0	1	0	1	1	0	0	0	0	0	0	4	0,4771212 5	0,4771213	0	0,4771212 5	0	0,4771212 5	0,4771212 5	0	0	0	0	0	0
FW	1	1	1	1	1	1	1	0	1	1	2	0	11	0,0377885 6	0,0377886	0,0377885 6	0,0377885 6	0,0377885 6	0,0377885 6	0,0377885 6	0,0377885 6	0	0,0377885 6	0,0377885 6	0,0755771 2	0
XX	4	3	3	4	5	4	6	1	6	4	4	4	48	-0,60206	-2,40824	-1,80618	-1,80618	-2,40824	-3,0103	-2,40824	-3,6123599	-0,60206	-3,6123599	-2,40824	-2,40824	-2,40824
Num	1	1	1	1	2	1	1	1	1	1	1	1	13	-0,0347621	-0,0347621	-0,0347621	-0,0347621	-0,0347621	-0,0695242	-0,0347621	-0,0347621	-0,03 4762 1	-0,0347621	-0,0347621	-0,0347621	-0,03 4762 1
FADV	0	0	0	0	0	0	0	0	0	0	2	0	2	0,7781512 5	0	0	0	0	0	0	0	0	0	0	1,5563025	0

-Fitur-

currentWord : Leksikal token yang diproses

currentTag : POS *tag* dari token yang diproses

Bef1Word : Leksikal token sebelum token yang diproses

Bef1Tag : POS *tag* dari token sebelum token yang diproses

Bef1Frasa : Frasa dari token sebelum token yang diproses

Bef2Word : Leksikal token dengan gap 2 dari token yang diproses

Bef2Tag : POS *tag* dari token dengan gap 2 dari token yang diproses

Bef2Frasa : Frasa dari token dengan gap 2 dari token yang diproses

4.1.7 Penerapan k-NN

Berdasarkan tahap *text mining* yang telah dijabarkan sebelumnya, maka pada bagian ini dijelaskan teknik-teknik yang akan digunakan dalam klasifikasi *Name Entity Recognition* pada *tweet* iklan. Proses perhitungan kuadrat jarak *query* menggunakan *Cosine Similarity*, berdasarkan pada persamaan 2.4 adalah sebagai berikut hitung hasil perkalian skalar antara dokumen 1 Token 1-6 dengan 4 dokumen lainnya perkalian berdasarkan token. Hasilnya perkalian dari setiap token dengan dokumen 1 Token 1-6 dijumlahkan

currentWord : X1T1*D2T1 = 1,079181 * 1,079181 = 1,164632162

: X1T1*D2T2 = 1,079181 * 1,079181 = 1,164632162

Fitur	CurrentWord	Bef1Word	Bef2word	CurrentTag	Bef1Tag	Bef2Tag	CurrentFrasa	Bef1Frasa	Bef2Frasa	Jumlah
D12*D1T1	1,164632162	0	0	7,566556676	0	0	4,93095187	0	0	13,66214
D12*D1T2	1,164632162	1,164632162	0	-1,149022474	7,566556676	0	4,93095187	4,930952	0	18,6087
D12*D1T3	1,164632162	1,164632162	1,164632162	7,566556676	-1,149022474	7,566556676	4,93095187	4,930952	4,930952	32,27084
D12*D1T4	1,164632162	1,164632162	1,164632162	7,566556676	7,566556676	-1,149022474	4,93095187	4,930952	4,930952	32,27084
D12*D1T5	1,1646	1,1646	1,1646	4,3497	7,5665	7,5665	4,34971	4,9309	4,9309	37,1883

Fitur	CurrentWord	Bef1Word	Bef2word	CurrentTag	Bef1Tag	Bef2Tag	CurrentFrasa	Bef1Frasa	Bef2Frasa	Jumlah
D12*D1 T6	1,1646	1,1646	1,1646	7,5665	4,34971	7,5665	4,9309	4,3497	4,9309	37,1883
D12*D1 T7	1,1646	1,1646	1,1646	-1,1490	7,5665	4,3497	4,9309	4,9309	4,3497	28,4727
D12*D1 T8	1,1646	1,164632	1,1646	7,566556676	-1,1490	7,5665	4,9309	4,9309	4,9309	32,2708
D12*D1 T9	0,2526	1,1646	1,1646	-0,1906	7,5665	-1,1490	-0,091	4,9309	4,9309	18,5796
D12*D1 T10	1,1646	0,2526	1,1646	4,3497	-0,1906	7,5665	4,3497	-0,091	4,9309	23,4971
D12*D1 T11	1,1646	1,1646	0,2526	0,8861	4,3497	-0,1906	0,0837	4,349715	-0,091	11,9695
D12*D1 T12	1,1646	1,1646	1,1646	4,3497	0,8861	4,3497	4,3497	0,08371	4,34971	21,8626

Hasil perkalian setiap token lebih detail terdapat pada Lampiran B.

Berdasarkan pada persamaan 2.6 hitung panjang setiap token dengan cara kuadratkan bobot setiap term dalam setiap token, jumlahkan nilai kuadrat tersebut dan kemudian diakarkan.

Fitur	CurrentWord	Bef1Word	Bef2word	CurrentTag	Bef1Tag	Bef2Tag	CurrentFrasa	Bef1Frasa	Bef2Frasa	jumlah	Akar dari jumlah
D1T1	1,164632162	0	0	9,87181619	0	0	4,1923932	0	0	15,2288416	3,90241484
D1T2	1,164632162	1,164632162	0	0,22764469	9,87181619	0	4,1923932	4,1923932	0	20,8135116	4,56218277
D1T3	1,164632162	1,164632162	1,164632162	9,87181619	0,22764469	9,87181619	4,1923932	4,1923932	4,1923932	36,0423532	6,00352839
D1T4	1,164632162	1,164632162	1,164632162	9,87181619	9,87181619	0,22764469	4,1923932	4,1923932	4,1923932	36,0423532	6,00352839
D1T5	1,164632162	1,164632162	1,164632162	3,2622861	9,87181619	9,87181619	3,2622861	4,1923932	4,1923932	38,1468875	6,17631666
D1T6	1,164632162	1,164632162	1,164632162	9,87181619	3,2622861	9,87181619	4,1923932	3,2622861	4,1923932	38,1468875	6,17631666
D1T7	1,164632162	1,164632162	1,164632162	0,22764469	9,87181619	3,2622861	4,1923932	4,1923932	3,2622861	28,502716	5,33879349
D1T8	1,164632162	1,164632162	1,164632162	9,87181619	0,22764469	9,87181619	4,1923932	4,1923932	4,1923932	36,0423532	6,00352839
D1T9	0,054794947	1,164632162	1,164632162	0,00626967	9,87181619	0,22764469	0,00142798	4,1923932	4,1923932	20,8760042	4,56902661
D1T10	1,164632162	0,054794947	1,164632162	3,2622861	0,00626967	9,87181619	3,2622861	0,001428	4,1923932	22,9805385	4,79380209
D1T11	1,164632162	1,164632162	0,054794947	0,13540691	3,2622861	0,00626967	0,0012084	3,2622861	0,001428	9,05294443	3,00881113
D1T12	1,164632162	1,164632162	1,164632162	3,2622861	0,13540691	3,2622861	3,2622861	0,0012084	3,2622861	16,6796562	4,08407348

Cos(D12T1, D1T1) = 13,66214/(D1T1* D2T1)

= 13,66214/(3,572656*3,90241484)

= 0,92308

Berikut hasil cosine similarity pada D12 dengan dokumen latih

Cos(D12T1, D1T1)	Rencana	0,92308	Other
Cos(D12T1, D1T2)	Beli	1,026934	Other
Cos(D12T1, D1T3)	Mobil	1,380405	Produk
Cos(D12T1, D1T4)	Honda	1,380405	Produk
Cos(D12T1, D1T5)	Dp	1,587563	Other
Cos(D12T1, D1T6)	Angsuran	1,587563	Other
Cos(D12T1, D1T7)	Ringan	1,406178	Other
Cos(D12T1, D1T8)	Proses	1,417286	Other
Cos(D12T1, D1T9)	Follow	1,072179	Other
Cos(D12T1, D1T10)	@ahmadhonda165	1,292375	Penjual
Cos(D12T1, D1T11)	81282218485	1,048903	Kontak
Cos(D12T1, D1T12)	Belanjaseru	1,411442	Other

Keterangan :

D1 = data uji

D2 = data latih

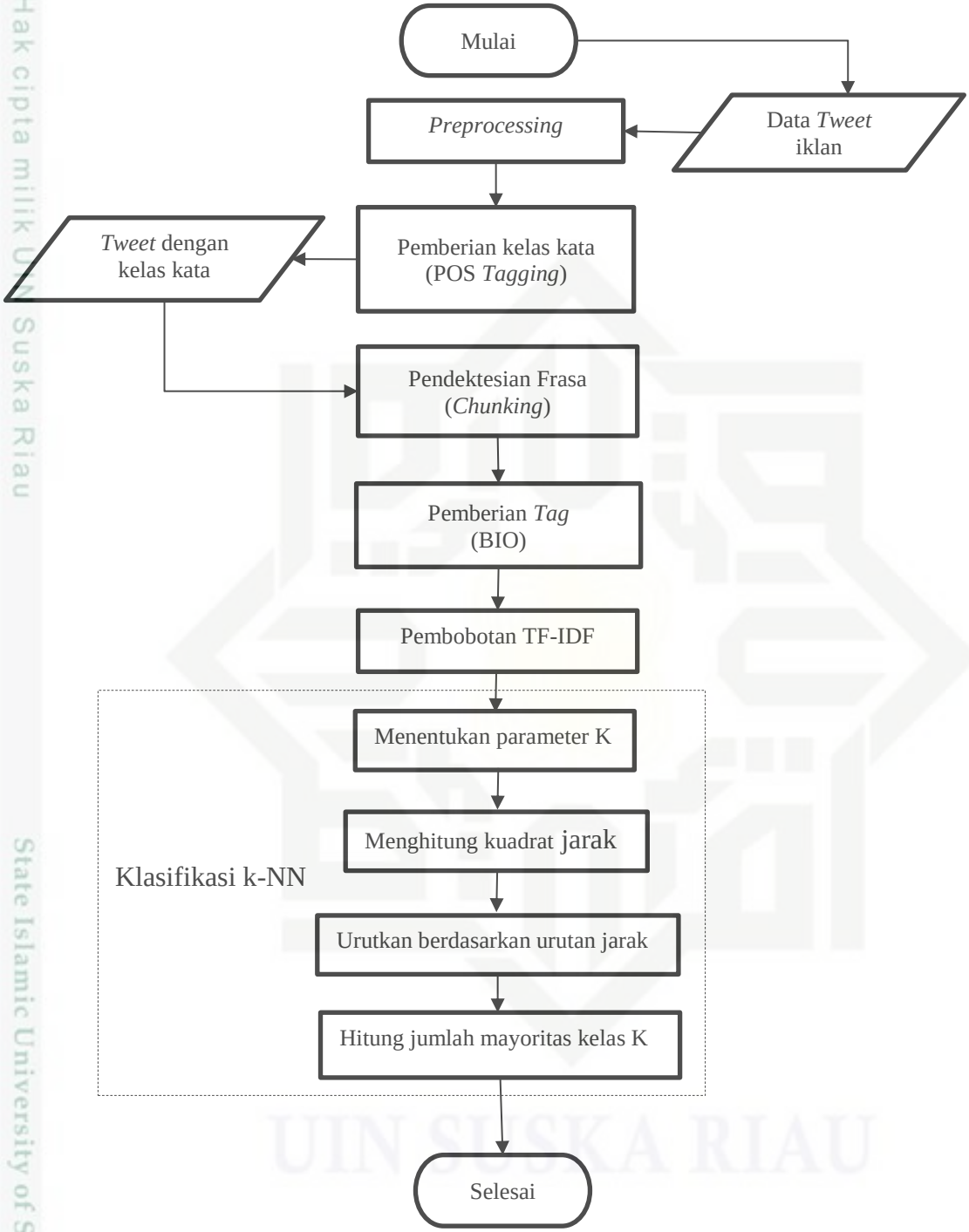
T_n = Token ke-n, dimana n = posisi kata

4.2 Perancangan

Tahap perancangan adalah tahapan membuat rancangan dalam proses *Named Entity Recognition* pada *tweet* iklan menggunakan metode *k-Nearest Neighbor*. Perancangan yang dibuat meliputi rancangan analisa yang dibuat menggunakan *flowchart* dan perancangan basis data berupa struktur tabel yang digunakan pada penelitian ini.

4.2.1 *Flowchart*

Flowchart merupakan penggambaran secara grafik dari langkah-langkah dan urut-urutan prosedur dari suatu program. *Flowchart* membantu untuk memecahkan masalah kedalam segmen-segmen yang lebih kecil. *Flowchart* dapat menunjukkan langkah kerja yang terjadi pada suatu sistem. Adapun *Flowchart* langkah kerja proses *Named Entity Recognition* (NER) dengan menggunakan k-NN dapat dilihat pada Gambar 4.6 berikut.



Gambar 4.6 Flowchart Langkah Kerja NER dengan Menggunakan k-NN

4.2.2 Perancangan Basis Data

Perancangan basis data menjelaskan deskripsi tabel yang dirancang pada basis data sistem. Berikut ini tabel – tabel yang terdapat di basis data.

Tabel 4.10 Rancangan Tabel Tweet

No	Nama *Field*	Tipe Data	Ukuran	Keterangan
1	Id	Varchar	50	*primary key*
2	user_id	Varchar	500	-
3	Teks	Varchar	500	-

Tabel 4.11 Rancangan Tabel Tagset

No	Nama *Field*	Tipe Data	Ukuran	Keterangan
1	Kata	varchar	50	*primary key*
2	Tag	Varchar	10	-