BAB III
METODE PENELITIAN

A. Subjek dan Objek Penelitian

Subjek penelitian dari penelitian ini adalah pesan SMS yang diterima oleh pengguna pada ponsel mereka. Objek penelitian dari penelitian ini adalah sistem dari aplikasi web yang akan dibangun. Sistem dari aplikasi tersebut akan mengklasifikasi pesan SMS, apakah pesan tersebut spam atau tidak spam.

B. Alat dan Bahan Penelitian

1. Alat Penelitian

Alat penelitian berupa perangkat keras seperti laptop, yang akan diimplementasikan model machine learning, dan juga perangkat lunak. Alat – alat penelitianya sebagai berikut :

a. Perangkat Keras

- Laptop (Processor Intel Core i5 / RAM 8GB)
- Mouse dan Keyboard

b. Perangkat Lunak

- Jupyter Notebook
- Anaconda Navigator
- Visual Studio Code
- MySQL Database
- XAMPP
- Flask Framework

2. Bahan Penelitian

Bahan penelitian yang akan digunakan pada penelitian ini berupa dataset pesan teks yang memiliki jumlah data 1140 data pesan. Dataset tersebut diperoleh dari penelitian peneliti lain [9]. Pada dataset memiliki dua kolom, kolom teks dan label.

C. Metodologi Penelitian

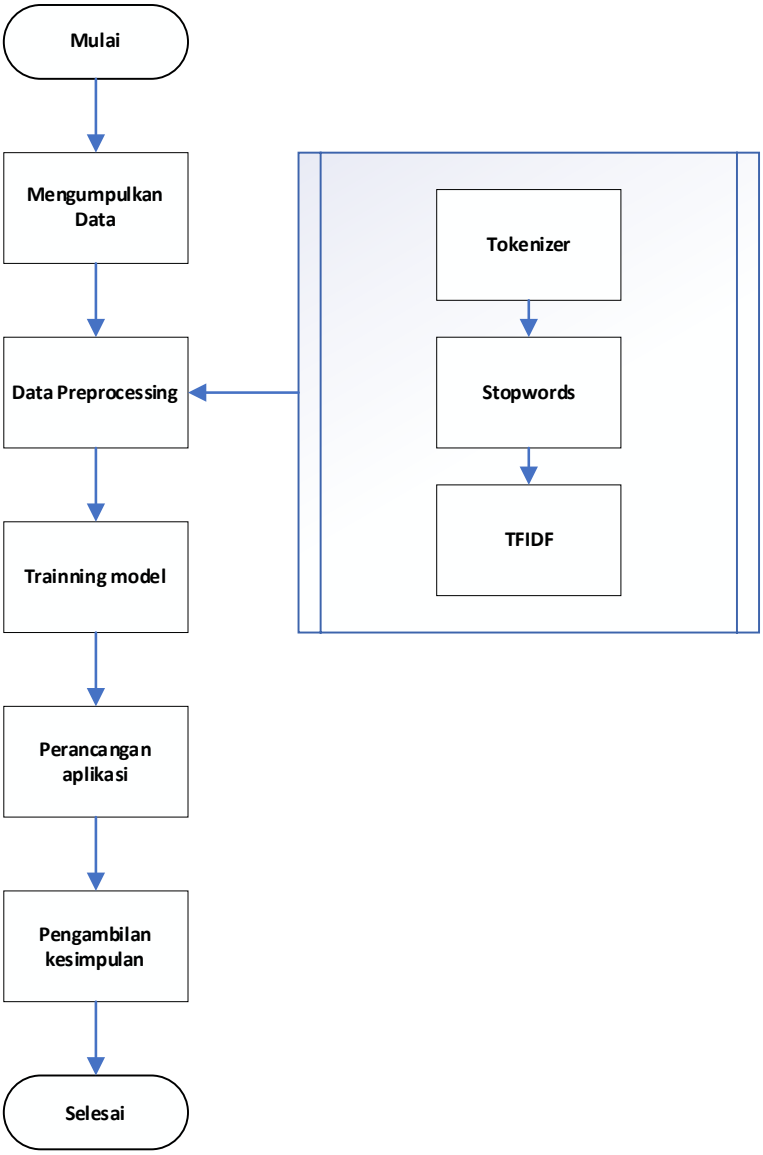
Metode yang digunakan dalam penelitian ini adalah metode kuantitatif. Menurut V. Wiratna Sujarweni [29] penelitian kuantitatif adalah jenis penelitian yang menghasilkan penemuan-penemuan yang dapat dicapai (diperoleh) dengan menggunakan prosedur-prosedur statistik atau cara lain dari kuantifikasi (pengukuran). Berdasarkan dari pengertian diatas, dapat disimpulkan bahwa penelitian metode kuantitatif adalah jenis penelitian yang cara melakukannya adalah dengan menggunakan pengukuran atau prosedur statistik.

Karakteristik penelitian kuantitatif menurut Kasiram [30] adalah sebagai berikut :

a. Menggunakan pola berpikir deduktif (rasional – empiris atau top-down), yang berusaha memahami suatu fenomena dengan cara menggunakan konsep-konsep yang umum untuk menjelaskan fenomena-fenomena yang bersifat khusus.
b. Logika yang dipakai adalah logika positivistik dan menghindari hal-hal yang bersifat subjektif.
c. Proses penelitian mengikuti prosedur yang telah direncanakan.
d. Melibatkan penghitungan angka atau kuantifikasi data.
e. Analisis data dilakukan setelah semua data terkumpul.
f. Dalam analisis data, peneliti dituntut memahami teknik-teknik statistik.
g. Hasil penelitian berupa generalisasi dan prediksi, lepas dari konteks waktu dan situasi.

D. Diagram Alir Penelitian

Tahapan proses yang akan dilakukan dalam penelitian ini digambarkan dalam diagram alir pada gambar berikut :



Gambar 3.1 Diagram alir penelitian

Tahapan awal penelitian dimulai dari pengumpulan dataset yang didapatkan melalui penelitian peneliti lain [9]. Setelah data dihimpun, tahapan selanjutnya adalah melakukan *preprocessing* data. Metode yang digunakan pada tahapan ini adalah menggunakan *tokenizer* atau pemenggalan kata, lalu *removal stopwords* yaitu menghapus kata – kata yang sering keluar dalam bahasa Indonesia, dan terakhir adalah pemberian bobot nilai menggunkan TFIDF. Setelah dataset berhasil melalui tahapan preprocessing, dataset tersebut akan di-*trainning* menggunakan algorimta Logistic Regression. Model machine learning yang sudah di-*trainning* selanjutnya diintegrasikan ke dalam aplikasi web.

E. Teknik Pengumpulan Data

Pengumpulan data adalah langkah penting dalam melakukan sebuah penelitian, karena tujuan utama dari penelitian data adalah mendapatkan data yang akan diteliti. Tanpa mengetahui teknik dalam mengumpulkan data, peneliti tidak akan mendapatkan data yang diperlukan untuk penelitian. Teknik pengumpulan data yang digunakan dalam penelitian ini adalah Teknik Dokumentasi. Teknik dokumentasi adalah teknik pengumpulan data dengan menghimpun dan menganalisis dokumen-dokumen, baik dokumen tertulis, gambar maupun elektronik, Sukmadinata [31].

F. Sumber dan Jenis Data

Jenis data yang digunakan pada penelitian ini adalah data sekunder. Menurut Sugiyono [32] mendefinisikan data sekunder adalah sebagai berikut: "Sumber Sekunder adalah sumber data yang diperoleh dengan cara membaca, mempelajari dan memahami melalui media lain yang bersumber dari literatur, buku-buku, serta dokumen".

Sumber data diperoleh dari hasil penelitian peneliti lain. Untuk kasus ini, peneliti memperoleh data untuk dataset SMS dari peneliti Rahmi, F. dan Wibisono, Y [9]. Jumlah keseluruhan sms pada dataset adalah 1140 pesan. Contoh dari isi dataset SMS adalah sebagai berikut :

Keterangan :
0 = Tidak SPAM
1 = SPAM

Tabel 3.1 Contoh dataset SMS

Teks	Label
[PROMO] Beli paket Flash mulai 1GB di MY TELKOMSEL APP dpt EXTRA kuota 2GB 4G LTE dan EXTRA nelpon hingga 100mnt/1hr. Buruan, cek di tsel.me/mytsel1 S&K	1
BONUS PULSA 50% dari Indosat Ooredoo, MAU?? Ketik *345# dan	1

Tabel 3.1 Contoh dataset SMS

Teks	Label
call dari HP kamu sebelum 28 April 2016	
BONUS PULSA 50rb cuma dgn isi ulang 25rb!! Hny Indosat yg bisa begini. Penawaran berlaku s/d 11 April 2016. Bonus 1x utk digunakan ke sesama Indosat Ooredoo	1
Aktifkan iRing Coboy Jr - Terhebat. Tekan *808*7#. Info: 100&111 Ada hits terbaru dari NOAH - Jika Engkau. Aktifkan iRing nya di HP kamu. Ketik MG NOAH02 kirim ke 808 Info: 100&111 Berkah iRing Rp 1000 dr Yuni Shara - Akhirnya. Aktifkan iRing nya, Tkn *808*1*2*3# lalu Ok/Call. Raih THR PuluhanJt!. Berhenti: Unreg ke 808	1
Anda akan membeli Paket Gampang Internetan Rp.1250 utk Chatting Sepuasnya (BBM, WhatsApp dan Line)/1 hari.Jika setuju balas FLASH<spasi>YA kirim ke 3636 utk melanjutkan.Tunggu SMS konfirmasi sblm pemakaian.Info detail tariff lihat di Tsel.me/internet	1
Abis maghrib yaa	0
Ada diruangan nya tadi	0
Ayam aja Pak, terus satu paket itu sama isinya ap aja?	0
Pak mau tanya kalau nilai remed lebih kecil itu diambil nilai nya yang remed atau tetap yang sebelum nya pak ?	0
Sama ingetin ya, ini kan web nya mau di publish, kita kan ikut develop tuh, nah di web nya ada nama kita atau jurusan ga?	0

G. Langkah – Langkah Logistic Regression pada Kasus Sms Spam

Dataset yang digunakan adalah dataset SMS yang memiliki 1440 pesan. Contoh dataset seperti di bawah ini.

	Teks	label
0	[PROMO] Beli paket Flash mulai 1GB di MY TELKO...	1
1	2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A...	1
2	2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ...	1
3	2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ...	1
4	4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an...	1
...	...	...
1138	Yooo sama2, oke nanti aku umumin di grup kelas	0
1139	sebelumnya ga ad nulis kerudung. Kirain warn...	0
1140	Mba mau kirim 300 ya	0
1141	nama1 beaok bwrangkat pagi...mau cas atay tra...	0
1142	No bri atas nama kamu mana	0

Gambar 3.2 Dataset SMS

Berikut penulis sajikan langkah – langkah untuk pengklasifikasian sms spam menggunakan algoritma *Logistic Regression*.

1. Ekstraksi Teks menggunakan *TF-IDF*

   TF-IDF adalah algoritma yang umum digunakan untuk seleksi fitur yang berupa teks[33]. *TF* memiliki arti sebagai seberapa banyak sebuah kata muncul dalam suatu dokumen, sedangkan *IDF* seberapa jarang suatu kata muncul dalam dokumen [34]. Berikut formulanya untuk *TF* dan *IDF*:

$$TF(t,d) = \frac{f_{t,d}}{(jumlah\ total\ kata\ dalam\ dokumen)} \tag{3.1}$$

   Keterangan :

   $f_{t,d}$ = Jumlah frekuensi kemunculan suatu kata pada dokumen

$$IDF(t) = log\frac{1+n}{1+df(t)} + 1 \tag{3.2}$$

Keterangan :

n = Jumlah dokumen

$df(t)$ = Jumlah dokumen yang mengandung kata tertentu

Pada kasus ini, penulis menggunakan bantuan *TfidfVectorizer* pada Jupyter Notebook untuk mengimplementasikan *TFIDF* pada dataset yang ada. Hasil dari ekstraksi teks adalah sebagai berikut :

```
(0, 2313)     0.6219833648003625
(0, 1429)     0.659510277335598
(0, 3861)     0.42211714961648017
(1, 4249)     0.27066478750901896
(1, 2277)     0.1286935232695319
(1, 1942)     0.12245265280946672
(1, 1138)     0.2552636419070825
(1, 1099)     0.27066478750901896
(1, 4100)     0.23586049314903626
(1, 1786)     0.12914842569733312
(1, 2473)     0.21800792289650264
(1, 2255)     0.27066478750901896
(1, 3067)     0.15156649276580716
(1, 2880)     0.22893520960082436
(1, 3712)     0.12692651170708147
(1, 2050)     0.18135038788204824
(1, 1149)     0.22307996579024286
(1, 3856)     0.1941309152408416
(1, 3956)     0.20591180862585595
(1, 3554)     0.18720563169262974
(1, 1135)     0.2026067772945662
(1, 929)      0.18515194783568611
(1, 2280)     0.4461599315804857
(2, 2863)     0.34979602517473424
(2, 3383)     0.6995920503494685
```

Gambar 3.3 Representasi contoh hasil perhitungan

Pada gambar diatas dijelaskan bahwa untuk kolom pertama adalah indeks dari kalimat, untuk kolom kedua adalah indeks *stopwords* dari kalimat tersebut. *Stopwords* adalah kata-kata yang sering muncul dan tidak memiliki arti penting [35]. Pada kolom berikutnya menampilkan bobot nilai *IDF* yang telah didapat.

2. Implementasi Algoritma *Logistic Regression*

   Setelah melakukan ekstraksi teks untuk dataset, selanjutnya mengimplementasikan *logistic regression* dengan rumus yang penulis sudah cantumkan pada Bab 2 di penelitian ini. Hasil prediksi sms dari proses kalkulasi yang dilakukan di Jupyter Notebook adalah sebagai berikut :

```
PRED: 1 - SMS : INFO RESMI Tri Care SELAMAT Nomor Anda terpilih mendapatkan Hadiah 1 Unit MOBIL Dri Tri Care Dengan PIN Pemenan
g: br25h99 info: www.gebeyar-3care.tk

PRED: 0 - SMS : Ceritanya biasa aja. Tp tetep bikin nangis

PRED: 1 - SMS : Aktifkan iRing Coboy Jr - Terhebat. Tekan *808*7#. Info: 100&111 Ada hits terbaru dari NOAH - Jika Engkau. Akti
fkan iRing nya di HP kamu. Ketik MG NOAH02 kirim ke 808 Info: 100&111 Berkah iRing Rp 1000 dr Yuni Shara - Akhirnya. Aktifkan i
Ring nya, Tkn *808*1*2*3# lalu Ok/Call. Raih THR PuluhanJt!. Berhenti: Unreg ke 808

PRED: 0 - SMS : Assalamualaikum akang teteh, jangan lupa ya hari ini ada carrier day jam 10:30, exhibition fpmipa b :) (love) H
ayuu teh ditungguuu

PRED: 0 - SMS : Rasa mau sidang bulan depaaan mel. Jahaaat pisaaan
```

Gambar 3.4 Hasil uji prediksi

H. Evaluasi Performa Model

Scoring yang digunakan untuk mengukur performa model machine learning yang penulis gunakan adalah *accuracy*. *Accuracy* adalah metrik untuk mengevaluasi model klasifikasi. *Accuracy* merupakan perhitungan pecahan dari jumlah prediksi yang benar pada model dengan total prediksi yang ada. *Accuracy* dapat digunakan untuk dataset yang memiliki *class* seimbang [36].

$$Accuracy = \frac{Jumlah\ prediksi\ benar}{Jumlah\ prediksi\ (benar\ dan\ salah)} \tag{3.3}$$

Untuk klasifikasi biner, *accuracy* bisa dihitung dalam hal positif dan negatif, seperti berikut ini :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

Keterangan :

TP = True Positive

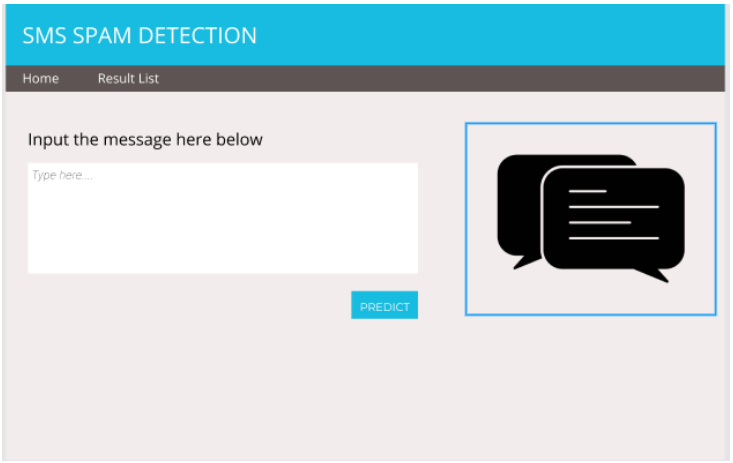
TN = True Negative

FP = False Positive

FN = False Negative

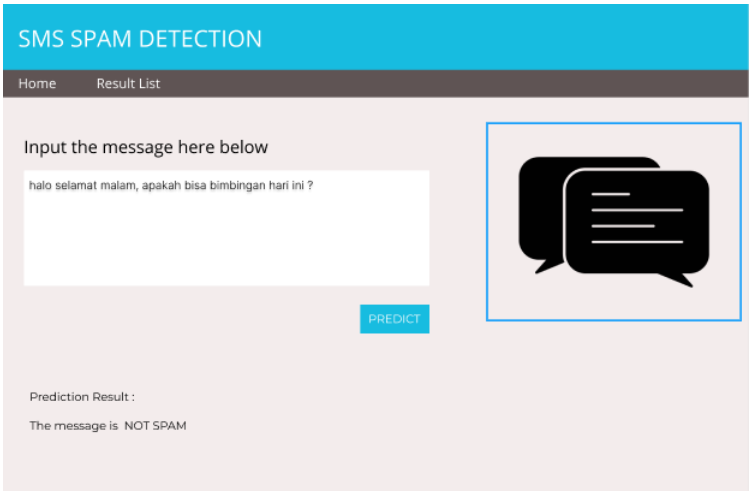
I. Mockup Aplikasi

1. Halaman Utama



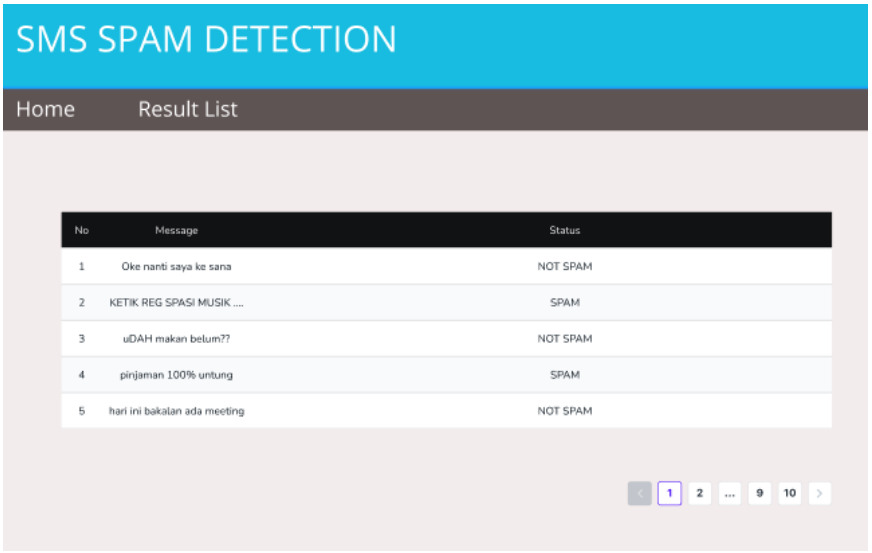
Gambar 3.5 Halaman utama *website*

Pada halaman utama website terdapat dua menu pada *navigation bar*, menu Home dan Result List. Pada menu Home menampilkan fitur utama yaitu input pesan pada textarea dan juga ada tombol PREDICT. Nantinya ketika user klik tombol tersebut, akan muncul hasil prediksi apakah pesan yang diinputkan adalah spam atau tidak spam.



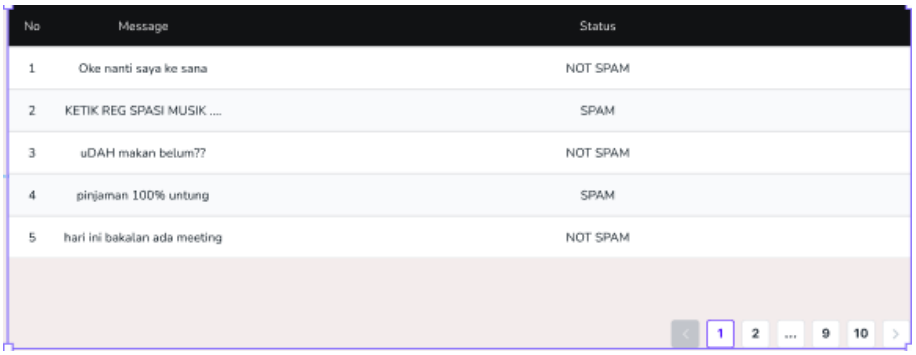
Gambar 3.6 Halaman utama *website* dengan prediction result

2. Halaman Result List

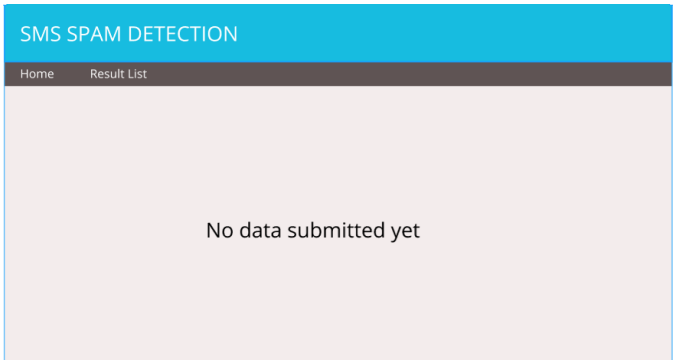


Gambar 3.7 Halaman Result List

Pada halaman Result List ini akan menampilkan hasil riwayat inputan pesan dari user beserta dengan hasil dari prediksi nya yang ditempatkan pada kolom status



Gambar 3.8 Isi konten tabel di Halaman Result List



Gambar 3.9 Halaman Result ketika belum ada pesan yang diprediksi

Pada gambar diatas menampilkan halaman Result List ketika belum ada pesan SMS yang diprediksi. Tampilan halaman akan berubah seperti pada Gambar 3.9, ketika ada pesan SMS yang sudah diprediksi oleh sistem.