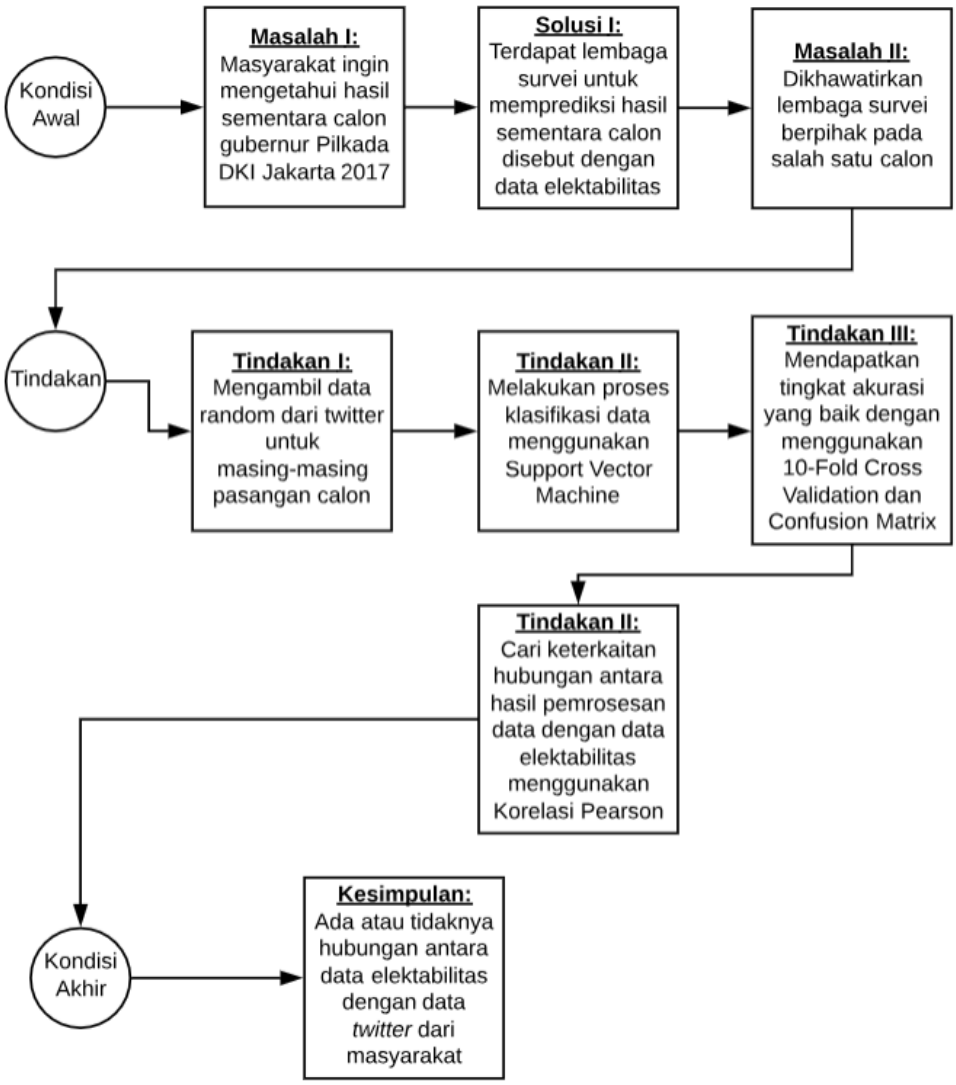
BAB III
METODE PENELITIAN

Pada bab ini berisi tentang Kerangka Pikir, Tahapan Penelitian, Pengumpulan Data, Analisis Data dan Design Interface, berikut penjelasannya:

3.1. Kerangka Pikir

Kerangka Pikir pada bagian ini merupakan penjelasan dari permasalahan yang dihadapi hingga permasalahan tersebut dapat diselesaikan dengan beberapa tahapan, dan ada hasil yang dapat dicapai untuk menyelesaikan permasalahan tersebut. Berikut penjelasan dari kerangka pikir.



Gambar 3.1. Kerangka Pikir

Berdasarkan gambar dari kerangka pikir diatas, dasar penulis melakukan penelitian ini karena adanya permasalahan yang berkaitan dengan politik pemilihan Gubernur dan Wakil Gubernur DKI Jakarta 2017. Permasalahan pertama yaitu masyarakat yang ingin mengetahui hasil sementara calon gubernur Pilkada DKI Jakarta 2017, maka lembaga survei telah menemukan solusi pertama yaitu lembaga survei dapat memprediksi hasil sementara calon gubernur yang

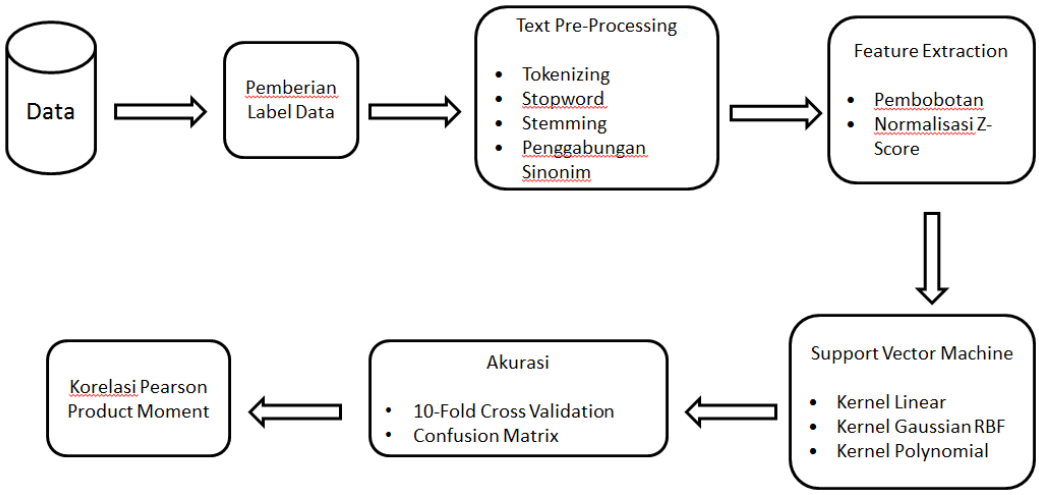
disebut dengan data elektabilitas yang melibatkan beberapa koresponden, maka dari itu muncul permasalahan kedua yaitu dikhawatirkan lembaga survei berpihak pada salah satu pasangan calon gubernur, maka adanya beberapa tindakan yang dilakukan, yaitu: tindakan pertama yang dilakukan adalah penulis mengambil data dari *twitter* untuk masing-masing pasangan calon yang kemudian data tersebut diberikan label oleh beberapa Tim dari lulusan Psikologi.

Tindakan kedua yang dilakukan adalah melakukan proses klasifikasi data menggunakan Support Vector Machine dengan menggunakan tiga kernel yaitu Kernel Linear, Kernel Gaussian RBF dan Kernel Polynomial sehingga tindakan ketiga yang dilakukan dari kernel tersebut mendapatkan tingkat akurasi yang baik dengan menggunakan perhitungan 10-*fold Cross Validation* dan Confusion Matrix. Hingga masuk dalam tindakan terakhir yaitu mencari keterkaitan hubungan antara hasil pemrosesan data dengan data elektabilitas menggunakan teknik Korelasi Pearson.

Sehingga pada akhirnya dapat ditarik kesimpulan serta solusi yaitu berdasarkan keterkaitan hubungan antara hasil pemrosesan data sentimen analisis dari *twitter* dengan data elektabilitas menggunakan teknik Korelasi Pearson ditemukan ada atau tidaknya hubungan antara kedua hal tersebut.

3.2. Tahapan Penelitian

Dibawah ini adalah tahapan-tahapan yang akan dilakukan dalam penulisan ini:



Gambar 3.2. Tahapan Penelitian

Berdasarkan gambar diatas dapat dijabarkan, untuk tahapan pertama yaitu tahapan input data, dimana untuk data diambil dari *twitter* menggunakan aplikasi *python* dengan menggunakan data sebanyak 2.000 data, kemudian tahapan pemberian label data yang dimaksudkan bahwa data yang sudah terkumpul akan diberikan label oleh sepuluh orang dari lulusan Psikologi, sehingga data yang sudah terlabel bisa dilakukan proses berikutnya.

Selanjutnya masuk dalam tahapan *text pre-processing* dalam tahapan ini terdapat empat cakupan yaitu *tokenizing* dimana berfungsi untuk memisahkan kata demi kata lalu *stopword* berfungsi untuk mencari kalimat yang relevan, apabila terdapat kalimat yang tidak relavan akan dibuang, dimana menggunakan Kamus Besar Bahasa Indonesia (KBBI) untuk mencari kata yang memiliki arti, langkah selanjutnya dalam *text pre-processing* adalah *stemming*, dimana berfungsi untuk

mencari kata dasar, dan menghilangkan beberapa kata yang memiliki imbuhan. Tahapan terakhir dalam *text pre-processing* adalah penggabungan sinonim, apabila ada kata-kata yang sama maka akan digabungkan dengan menambahkan jumlah frekuensi.

Tahapan selanjutnya adalah pembobotan, dalam tahapan ini terdapat dua bagian yaitu *feature extraction* dimana berfungsi untuk memberikan bobot menggunakan TF-IDF dan bagian kedua adalah normalisasi Z-Score, dimana berfungsi untuk mencari kedekatan antara bobot yang satu dengan yang lain.

Selanjutnya ialah masuk ke dalam algoritma *Support Vector Machine* dengan membandingkan ketiga kernel yaitu Kernel Linear, Kernel Gaussian RBF dan Kernel Polynomial dimana diharapkan pada langkah ini proses yang dilakukan sudah tepat.

Untuk tahapan penghitungan akurasi menggunakan *K-Fold Cross Validation* dan diharapkan akan memiliki akurasi tinggi, dan metode kedua untuk akurasi menggunakan Confusion Matrix, yang diharapkan kedua metode memiliki tingkat akurasi yang sama, serta adanya input data baru, untuk mengetahui data baru termasuk dalam kelompok yang mana.

Langkah terakhir adalah mencari keterkaitan hubungan menggunakan Korelasi Pearson Product Moment antara hasil dari sentimen twitter yang terbagi menjadi dua kategori (positif dan negatif) dengan data elektabilitas pasangan calon gubernur dan calon wakil gubernur dari beberapa lembaga survei.

Dibawah ini adalah langkah detail yang akan dilakukan:

3.1.1. Data pada Twitter

Data yang digunakan berasal dari *tweet* berbahasa Indonesia yang terbagi menjadi dua emosi (positif dan negatif). Dibawah ini adalah emosi yang berkaitan dengan positif :

LenteraCinta @SemuaCintaKamu

Mengalah bukan berarti kalah, mengalahlah itu utk mencapai kebahagiaan kamu & dia #mengalah #bahagia #**cinta**

1Motivasi @1MENITcom

Cinta tak akan menuntut kesempurnaan. Cinta akan menerima, memahami, pun rela berkorban. #1Menit #**cinta**

Toko Tiens Bandung @TiensBandung

Lari Meningkatkan Hormon Endorphine Ditubuh Sehingga Membuat Kita Lebih Senang #Lari #Olahraga #**Senang**

Martin Angguli @Martin_Angguli

Senengnya liat joel ketawa #joelsalvatore #ketawa #**senang**

Dibawah ini adalah emosi yang berkaitan dengan negatif :

Viral Effecto @ViralEffecto

#**marah**¸massa ngadem¸orator demo¸ Orator Demo Tolak Ahok Marah Karena

Peserta Lebih Memilih Ngadem

LenteraCinta @SemuaCintaKamu

Amarah akan menimbulkan rasa sesal, Sabar akan melahirkan rasa syukur **#marah** #sabar #syukur #sesal

Annisa Citra @citranisa21

malming tanpa #pacar belum tentu **#sedih** | malming sama pacar belum tentu #bahagia

kamal7qilla @kamal7qilla

Sehari sebelum kedatangan Raja Arab di Bogor pd tau ga #bogor sempat banjir Bandung n rumah gw di #depok sempat kebanjiran **#sedih**

dw @bang_dw

Sikap Jokowi yang membiarkan Kemendagri tidak berhentikan sementara Ahok | adalah bukti negara takut satu sosok bernama Ahok | **#takut**

Motivasi Kerja @Motivasi_Kerja_

Halangan Besar Untuk Manusia Biasa Adalah Takut Gagal.. #Gagal **#Takut** #Manusia #Inspirasi #Motivasi #Sukses

3.1.2. Text Preprocessing

Pada tahap *preprocessing*, sistem melakukan tahap *tokenizing*, *remove stopword* dan *stemming*. Sistem juga melakukan beberapa perlakuan khusus terhadap data yang digunakan karena *tweet* mengandung banyak *noise*. Sistem akan menghapus *link url*, *username*, tanda *retweet*, dan beragam *noise* lain. Sistem akan mengubah kata tidak baku atau kata yang disingkat menjadi kata yang baku. Sistem juga akan mengambil kata yang diawali tanda pagar (*hashtag*).

Langkah-langkah *tokenizing* :

1. Baca tiap baris pada *file text* sebagai satu *tweet.*
2. Ambil tiap *token* pada kalimat *tweet* dengan menggunakan spasi sebagai pemisah antara satu *token* dengan *token* lain.
3. Simpan tiap kalimat *tweet* yang terdiri dari *token* penyusun.

Langkah-langkah *remove stopword* :

1. Baca tiap *token* dan cocokkan dengan kata pada daftar *stopword.*
2. Hapus *token* jika cocok dengan kata pada daftar *stopword.*

Langkah-langkah *stemming* :

1. Baca tiap *token* dan cocokkan dengan kata pada daftar kamus kata dasar.
2. Jika *token* cocok dengan kata pada daftar kamus kata dasar, berarti *token* adalah *root word.*.
3. Jika *token* tidak cocok dengan kata pada daftar kamus kata dasar, hapus akhiran dan awalan pada *token.*

4. Cocokkan hasil langkah 3 dengan kata pada daftar kamus kata dasar, jika tidak cocok, anggap *token* sebelum dikenai langkah 3 sebagai *root word*.

Langkah-langkah hapus *noise tweet* :

1. Menghapus *url* : menghapus kalimat yang berawalan "*www*", "*http*" atau "*https*".

2. Menghapus *username* : menghapus kata yang berawalan tanda "@" misalnya @lambeturah.

3. Menghapus kata berawalan angka misalnya "30hari".

4. Memangkas huruf sama berurutan misalnya "jalannn" menjadi "jalan".

5. Menghapus angka, tanda baca, dan karakter selain huruf.

6. Menghapus *noise* lain yang ada dalam data seperti tanggal penulisan *tweet*, tanda *retweet*, kata "*view*" dan "*translation*", serta penanda waktu penulisan *tweet* misalnya "*hour*", "*hours*", "*ago*".

Langkah-langkah sinonim kata :

1. Cari sinonim kata pada daftar kata sinonim.

2. Jika ditemukan, ganti kata awal dengan kata sinonim.

3. Jika tidak ditemukan, kata awal tidak diganti.

Langkah-langkah penanganan kata negasi :

1. Temukan kata tidak, bukan, atau tanpa.

2. Gabung kata tidak, bukan, atau tanpa dengan kata di belakang misalnya "tidak" "senang" menjadi "tidaksenang".

3. Hapus kata yang telah digabung dengan kata tidak, bukan, atau tanpa.

3.1.3. Feature Extraction

Feature Extraction bertujuan untuk mengambil ciri unik atau informasi yang penting dari data yang akan diolah pada tahapan proses selanjutnya. Feature extraction juga bertujuan untuk memperkecil jumlah data dengan tahapan seperti dibawah ini:

- **Tahap Pembobotan**

Pada tahap pembobotan, sistem akan merepresentasikan *tweet* sebagai *vector* dengan nilai bobot masing-masing *term*. Perhitungan bobot *term* menggunakan metode pembobotan *tf-idf*.

Langkah-langkah pembobotan *tf-idf* :

1. Untuk setiap data *tweet*, lakukan langkah 2 – 4.
2. Hitung nilai *tf* masing-masing kata.
3. Hitung nilai *idf* masing-masing kata.
4. Hitung bobot *tweet* dengan mengalikan nilai *tf* dan *idf*.

- **Tahap Normalisasi Z-Score**

Pada tahap normalisasi metode yang digunakan dengan menggunakan Normalisasi Z-Score. Nilai bobot *term* yang dinormalkan hanya bobot *term* yang dominan saja. Bobot dominan yaitu bobot yang bernilai lebih dari *threshold* tertentu.

Langkah-langkah normalisasi *z-score* :

1. Hitung nilai *mean* pada setiap *tweet*.

2. Hitung nilai *standard deviation* pada setiap *feature*.

3. Hitung bobot baru. Bobot baru didapat dari bobot lama dikurangi rata-rata (*mean*) kemudian dibagi *standard deviation*

3.1.4. Support Vector Machine

Pada tahapan ini, sistem akan mengelompokkan *tweet* ke dalam dua *cluster* yaitu positif dan negatif. Setiap *tweet* akan dikelompokkan berdasarkan *hyperplane*.

Terdapat tiga kernel yang digunakan untuk mencari *hyperplane* terbaik yaitu Kernel Linear, Kernel Gaussian RBF dan Kernel Polynomial. Maka dari itu dari ketiga kernel tersebut dapat terlihat kernel yang terbaik berdasarkan akurasi tertinggi, sehingga nantinya akan digunakan pada tahapan Korelasi Pearson Product Moment sebagai Hasil Sentimen Analisis.

3.1.5. Perhitungan Akurasi

K-fold cross validation merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. K-fold cross validation diawali dengan membagi data sejumlah n-fold yang diinginkan. Dalam proses cross validation data akan dibagi dalam n buah partisi dengan ukuran yang sama D1, D2, D3 .. Dn

selanjutnya proses uji dan latih dilakukan sebanyak n kali. Dalam iterasi ke-i partisi Di akan menjadi data uji dan sisanya akan menjadi data latih. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model. Maka penulis dalam tahapan akurasi menggunakan 10-fold cross validation.

Cara kerja 10-fold cross validation adalah sebagai berikut:

1. Total instance dibagi menjadi 10 bagian.

2. Fold ke-1 adalah ketika bagian ke-1 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut. Perhitungan akurasi tersebut dengan menggunakan persamaan sebagai berikut:

$$\text{Akurasi} = \frac{\sum \ data\ uji\ benar\ klasifikasi}{\sum total\ data\ uji} x\ 100\%$$

3. Fold ke-2 adalah ketika bagian ke-2 menjadi data uji (testing data) dan sisanya menjadi data latih (training data). Selanjutnya, hitung akurasi berdasarkan porsi data tersebut.

4. Demikian seterusnya hingga mencapai fold ke-10. Hitung rata-rata akurasi dari 10 buah akurasi di atas. Rata-rata akurasi ini menjadi akurasi final.

Metode kedua yang digunakan adalah Confusion Matrix, dengan perhitungan akurasi (AD) dimana akurasi dihitung berdasarkan jumlah prediksi yang benar bahwa contoh bersifat negatif ditambahkan dengan jumlah prediksi yang benar bahwa contoh bersifat positif dan dibagi dengan keseluruhan jumlah prediksi. Sehingga diharapkan kedua metode perhitungan akurasi tersebut memiliki nilai tingkat akurasi yang sama.

3.1.6. Korelasi Pearson Product Moment

Korelasi Pearson Product Moment digunakan untuk mengukur uji validitas, dimana dalam penelitian ini Korelasi Pearson digunakan untuk membandingkan hasil sentimen analisis pada twitter dengan data elektabilitas pasangan calon gubernur dan calon wakil gubernur putaran kedua.

Korelasi Pearson yang digunakan terdiri dari dua sentimen yaitu positif dan negatif yang dapat dilihat dalam bentuk grafik masing-masing pasangan Calon Gubernur dan Calon Wakil Gubernur. Sehingga melalui Korelasi Pearson terlihat adanya keterkaitan antara hasil sentimen dengan data elektabilitas, dan dapat diambil kesimpulan yaitu *tweet* dari masyarakat terkait dengan pemilihan gubernur dan wakil gubernur apakah sesuai dengan data elektabilitas yang bersumber dari beberapa lembaga survei di Indonesia.

Pada data elektabilitas yang didapat ialah dari beberapa lembaga survei tiap bulan, dan tiap data per bulan jumlah lembaga survei tidak seimbang, maka data elektabilitas perbulan di rata-rata untuk akhirnya hasil dari ra-rata tersebut dapat diproses. Data elektabilitas yang didapat berupa persentase (%).

Menurut Shaver, Murdaya dan Fraley di tahun 2011 yang sudah dijelaskan pada Bab II, sentiment terdiri dari positif dan negatif, sehingga untuk analisa sentimen yang digunakan menggunakan dua emosi dasar yaitu positif dan negatif.

Supaya data dari hasil sentimen dari *twitter* dapat dibandingkan dengan data elektabilitas, maka data hasil sentimen *twitter* diubah dalam persentase, sebagai contoh total data adalah 1000, terdiri dari 600 label sentiment positif dan

400 label sentiment negatif, maka untuk mengubah dalam bentuk persentase ialah total label sentiment dibagi dengan total keseluruhan data dikali dengan 100%, sehingga didapat persentase masing-masing hasil sentiment.

3.3. Pengumpulan Data

Pada tahapan pengumpulan data memiliki dua sumber data yang dibutuhkan yaitu sumber data dari *twitter* dan sumber data elektabilitas dari beberapa lembaga survei. Berikut rinciannya:

3.3.1. Pengumpulan Data Twitter

Pada tahapan pengumpulan data twitter, data dikumpulkann berdasarkan nama calon gubernur dan calon wakil gubernur putaran kedua yang dimulai Bulan Januari hingga Bulan April, maka penulis melakukan pencarian dimulai dari bulan Januari hingga April. Kategori pencarian pertama ialah calon gubernur dan calon wakil gubernur "Basuki Tjahaja Purnama-Djarot Saiful Hidayat" dan kategori pencarian kedua ialah calon gubernur dan calon wakil gubernur "Anies Rasyid Baswedan-Sandiaga Salahuddin Uno" dengan total keseluruhan data sebanyak 2.000 data

Selanjutnya data yang diperoleh kemudian diberikan label per *tweet* oleh sepuluh orang dari lulusan Psikologi, dimana label yang diberikan disesuaikan dengan dua emosi yang sudah dipaparkan pada latar belakang positif dan negatif. Berikut Tabel yang berkaitan dengan label dan emosi.

Tabel 3.1. Keterangan Emosi

Label	Emosi
0	Tidak ada Hubungan
1	Positif
2	Negatif

Untuk mempermudah tim Psikologi memberikan label maka dapat diwakilkan dengan memasukkan angka 0-2 dengan keterangan seperti tabel diatas. Label 0 yaitu *tweet* yang tidak ada hubungan, maka data tidak dipergunakan.

3.3.2. Pengumpulan Data Elektabilitas

Pengumpulan data elektabilitas didapat dari beberapa lembaga survei Indonesia yang diambil dari bulan Januari hingga April. Data elektabilitas yang telah terkumpul nantinya akan dihitung rerata tiap bulan, supaya data yang digunakan seimbang.

Sumber yang diambil dari data elektabilitas berbagai lembaga survei yaitu lembaga survei Charta Politika, Poltracking, SMRC, Populi Center, Indikator, Lembaga Survei Indonesia dan LSI Denny JA. Data diambil dari bulan Januari 2017 hingga April 2017 dan penulis membuat rerata tiap-tiap bulan pada seluruh lembaga survei yang digunakan untuk penelitian ini.

3.4. Analisis Data

Pada tahapan ini adalah melakukan tahapan analisis dimana melihat performa atau akurasi algoritma *support vector machine* dan melihat hubungan antara hasil

sentiment analysis dengan hasil pilkada. Dan akan lebih detail dibahas pada BAB IV, Hasil dan Pembahasan, setelah seluruh data diimplementasikan.

Langkahnya adalah melalui data *dari* twitter, teks akan mengalami tahapan *preprocessing* yang terdiri dari (*tokenizing, stopword, stemming* dan penggabungan sinonim). Tujuan dari tahapan penggabungan sinonim adalah apabila terdapat kata yang berbeda namun makna sama, maka gabungkan menjadi satu kata, sehingga diharapkan tidak terdapat kata yang duplikat yang nantinya berpengaruh pada hasil analisis.

Tahap kedua yaitu tahapan pembobotan kata menggunakan TF-IDF untuk menentukan nilai frekuensi dari dokumen, maka hasil pembobotan di normalisasi menggunakan *z-score*, supaya dapat membandingkan bobot pada kata satu terhadap kata lainnya. Tahapan ketiga algoritma *Support Vector Machine* dengan menggunakan tiga kernel yaitu Kernel Linear, Kernel Gaussian RBF dan Kernel Polynomial untuk mencari *hyperplane* terbaik. Pada penghitungan akurasi menggunakan *10-Fold cross-validation* dan Confusion Matrix sehingga dapat membandingkan akurasi mana yang lebih baik dengan menggunakan ketiga kernel.

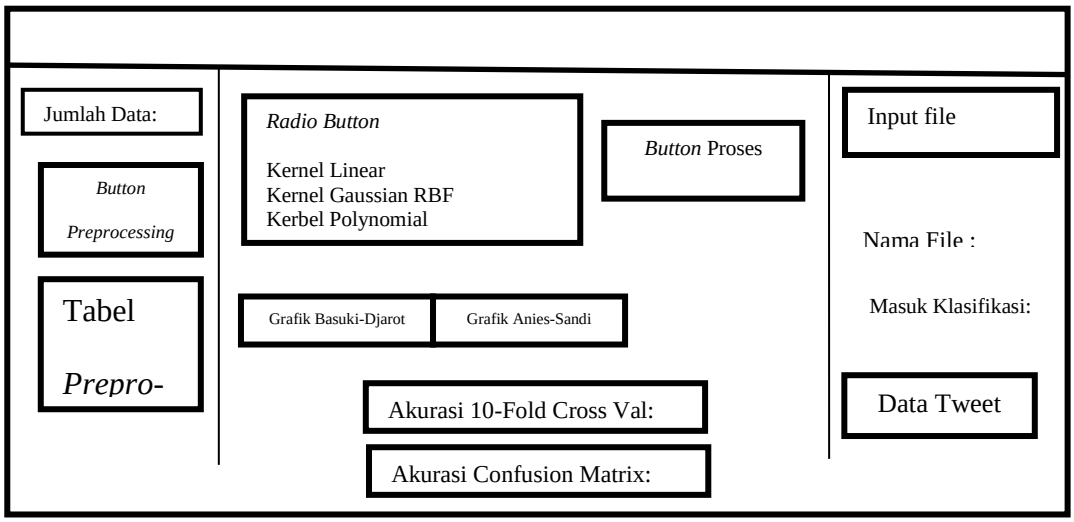
Tahapan terakhir yaitu metode Korelasi Pearson, dimana digunakan untuk mencari keterkaitan hubungan antara hasil sentimen analisis menggunakan algoritma Support Vector Machine dengan data elektabilitas pasangan calon gubernur dan calon wakil gubernur putaran kedua.

Sehingga, setelah mendapatkan data elektabilitas dari masing-masing pasangan calon gubernur dan wakil gubernur dapat mengambil kesimpulan darii

apa yang sudah penulis lakukan yaitu adakah hubungan antara data elektabilitas dengan data *twitter* dari masyarakat.

3.5. Desain Interface

Pada gambar merupakan desain interface yang akan dibuat pada sistem ini.



Gambar 3.3. Desain Interface