

**SENTIMENT ANALYSIS PEMILIHAN CALON PRESIDEN 2024
MENGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)**

SKRIPSI

Sebagai Salah Satu Syarat untuk Meraih
Gelar Sarjana Komputer

Oleh :

Michael Alfonso
32190039



Fakultas Teknologi dan Desain
Program Studi Informatika
Universitas Bunda Mulia
Tangerang
2023

**UNIVERSITAS BUNDA MULIA
FAKULTAS TEKNOLOGI DAN DESAIN
PROGRAM STUDI INFORMATIKA**

Pernyataan Kesiapan Ujian Pendadaran Skripsi

Saya Michael Alfonso, dengan ini menyatakan bahwa Skripsi yang berjudul:

**SENTIMENT ANALYSIS PEMILIHAN CALON PRESIDEN 2024
MENGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)**

merupakan hasil karya saya dan belum pernah diajukan sebagai karya ilmiah, sebagian atau seluruhnya, atas nama saya atau pihak lain

tanda tangan

Michael Alfonso
32190039

Disetujui oleh Pembimbing,

Kami setuju Skripsi tersebut diajukan untuk Ujian Pendadaran

tanda tangan

Dionisia Bhisetya Rarasati, S.Kom., M.T.I

02 Mei 2023

Disetujui oleh Ketua Program Studi,

tanda tangan

Dr. Fransiskus Adikara, S.Kom., M.M.

02 Mei 2023

P E R N Y A T A A N

Saya menyatakan bahwa Skripsi yang berjudul **SENTIMENT ANALYSIS PEMILIHAN CALON PRESIDEN 2024 MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)**, sepenuhnya karya saya sendiri. Tidak ada bagian di dalamnya yang merupakan plagiat dari karya orang lain dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, saya siap menanggung resiko / sanksi yang dijatuhkan kepada saya, atau ada klaim dari pihak lain terhadap keaslian karya saya ini.

Tangerang, 02 Mei 2023

Yang membuat pernyataan

Materai Rp 10.000,-

Ttd

Michael Alfonso

ABSTRAK

Pemilihan calon presiden dilaksanakan setiap 5 tahun dengan berbagai kandidat yang mencalonkan diri, terutama dalam media sosial *Twitter* lebih sering terjadi argumen seputar hal-hal politik yang tentunya banyak pengguna *Twitter* turut ikut berdiskusi tentang pemilihan calon presiden ini. Pengguna *Twitter* akan melakukan *tweet* untuk menyampaikan argumentasi dan diskusi terkait dengan pemilihan calon presiden ini. Oleh karena itu, penelitian ini berfokus pada sentiment analysis untuk melakukan penyimpulan respon pengguna terhadap pemilihan calon presiden serta melakukan validasi dengan mencari korelasi antara hasil survei elektabilitas dan data sentiment twitter dengan menggunakan Korelasi Pearson.

Dalam pembangun mesin sentiment, metode *10-Fold Cross Validation* digunakan untuk mencari model mesin terbaik dari suatu dataset dengan pembagian data training dan data test sebesar 90:10. Lalu data alfabet akan diubah menjadi bentuk numerik dengan menggunakan metode pembobotan TF-IDF. Selanjutnya, untuk melakukan validasi dari model terbaik menggunakan *Confusion Matrix* untuk mendapat *f1-score*. Algoritma yang digunakan untuk membuat model adalah algoritma Support Vector Machine dengan kernel Gaussian RBF (Radial Basis Function).

Berdasarkan hasil pencarian fold terbaik, ditemukan fold terbaik untuk masing-masing calon presiden yaitu fold ke-8 dengan f1-score 0,66 untuk calon Anies Baswedan dengan total 2.554 data training dan 283 data testing, fold ke-5 dengan f1-score 0,72 untuk calon Ganjar Pranowo dengan total 3.330 data training dan 370 data testing, dan fold ke-4 dengan f1-score 0,78 untuk calon Prabowo Subianto dengan total 3487 data training dan 387 data testing. Selanjutnya pada Korelasi Pearson, ditemukan koefisien untuk masing-masing calon presiden yaitu Anies Baswedan dengan koefisien sentiment positif sebesar 0,876 dan koefisien sentiment negatif sebesar -0,876. Selanjutnya untuk calon presiden Ganjar Pranowo dengan koefisien sentiment positif sebesar 0,894 dan koefisien sentiment negatif sebesar -0,894. Lalu untuk calon presiden Prabowo Subianto dengan koefisien sentiment positif sebesar 0,97 dan koefisien sentiment negatif sebesar -0,97.

Penelitian ini menghasilkan fold terbaik untuk tiap data pada masing-masing calon presiden dengan ukuran f1-score untuk mencari model terbaik dari tiap fold. Pada Korelasi Pearson, semakin tinggi sentiment positif setiap calon presiden, maka semakin tinggi juga data survei elektabilitas. Untuk penelitian selanjutnya, dapat dilakukan penelitian yang membahas *hyper tuning* parameter dan menggunakan kernel lain pada algoritma Support Vector Machine.

Kata Kunci

NLP, Pearson Correlation, Sentiment Analysis, SVM, TF-IDF

ABSTRACT

Elections for presidential candidates are held every 5 years with various candidates, especially on Twitter, arguments about political matters often occur that many Twitter users participate in discussions about the election for presidential candidate. Twitter users will tweet to convey arguments and discussions related to the election. Therefore, this study focuses on sentiment analysis to infer user responses to the presidential election and validate it by looking for a correlation between electability survey results and Twitter sentiment data using Pearson Correlation.

In sentiment analysis model, the 10-Fold Cross Validation method is used to find the best model from a dataset with a division of training data and test data with 90:10 split. Then the alphabetic data will be converted into numeric data using the TF-IDF weighting method. To validate the best model, Confusion Matrix is used to get the best f1-score. The model is using Support Vector Machine algorithm with the Gaussian RBF (Radial Basis Function) kernel.

Based on the search results for the best fold, the best fold was found for each presidential candidate, namely the 8th fold with an f1-score of 0.66 for candidate Anies Baswedan with a total of 2,554 training data and 283 testing data, the 5th fold with an f1-score of 0.72 for the Ganjar Pranowo candidate with a total of 3,330 training data and 370 testing data, and the 4th fold with an f1-score of 0.78 for the Prabowo Subianto candidate with a total of 3,487 training data and 387 testing data. Furthermore, in the Pearson Correlation, a coefficient was found for each presidential candidate, namely Anies Baswedan with a positive sentiment coefficient of 0.876 and a negative sentiment coefficient of -0.876. Furthermore, for the presidential candidate Ganjar Pranowo with a positive sentiment coefficient of 0.894 and a negative sentiment coefficient of -0.894. Then for presidential candidate Prabowo Subianto with a positive sentiment coefficient of 0.97 and a negative sentiment coefficient of -0.97.

This study produces the best fold for each data on each presidential candidate with the f1-score to find the best model for each fold. In the Pearson Correlation result, the higher positive sentiment of each presidential candidate, the higher electability survey data. For further research, research can be discuss about hyper tuning parameters and using other kernels on Support Vector Machine algorithm.

Keywords

NLP, Pearson Correlation, Sentiment Analysis, SVM, TF-IDF

PRAKATA

Puji syukur penulis sampaikan kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul **“SENTIMENT ANALYSIS PEMILIHAN CALON PRESIDEN 2024 MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM)”** tepat waktu sebagai syarat untuk mendapatkan gelar sarjana komputer. Penulis sadar penelitian ini jauh dari kata sempurna, namun dari ketidaksempurnaan itu dilengkapi oleh banyak pihak yang turut terlibat baik secara langsung maupun tidak langsung dalam menyelesaikan penelitian ini. Sebab itu, pada kesempatan ini penulis ingin mengucapkan banyak terimakasih secara khusus pada beberapa nama yang disebutkan dibawah ini:

1. Tuhan Yang Maha Esa yang memberikan kekuatan kepada penulis untuk menyelesaikan skripsi ini.
2. Bapak Doddy Surja Bajuadji, S.E., M.B.A., selaku Rektor Universitas Bunda Mulia.
3. Bapak Howard S.Giam, S.E., Ak., M.B.A., selaku Pelaksana Harian Rektor Universitas Bunda Mulia.
4. Ibu Kandi Sofia Senastri Dahlan, S.E., M.B.A., Ph.D., selaku Wakil Rektor Bidang Akademik Universitas Bunda Mulia.
5. Ibu Shanty Sudarji, S.Psi., M.Psi., Psikolog, selaku Direktor Universitas Bunda Mulia Kampus Serpong.
6. Bapak Dr.Fransiskus Adikara, S.Kom., MM., selaku Dekan Fakultas Teknologi dan Desain Universitas Bunda Mulia.

7. Ibu Henny Hartono, S.Kom., M.M., selaku Head of Akademik 1 Universitas Bunda Mulia.
8. Bapak Dr. Fransiskus Adikara, S.Kom., MM., selaku Ketua Program Studi Informatika.
9. Bapak Ignatius Adrian Mastan, S.E., S.Kom., S.A.B., M.M., M.Eng., selaku Manager Akademik 1 Universitas Bunda Mulia Kampus Serpong.
10. Ibu Dionisia Bhisetya Rarasati, S.Kom., M.T.I., selaku Dosen Pembimbing.
11. Seluruh Dosen dan Staff Universitas Bunda Mulia yang telah memberikan ilmu dan masukan terhadap skripsi ini.
12. Orang tua penulis yang mendukung penulis dalam segi finansial dalam menyelesaikan skripsi penulis.
13. Afiyah Salsabila Arief selaku teman penulis yang memberikan skripsinya sebagai referensi penulis dan melakukan diskusi dengan penulis tentang penyusunan skripsi penulis
14. Teman-teman penulis yang juga berjuang menyelesaikan skripsi dan menjadi teman diskusi penulis untuk bertukar pikiran dan informasi dalam penyelesaian skripsi penulis.

Akhir kata, penulis berharap bahwa skripsi ini dapat bermanfaat untuk pembaca baik, terutama sebagai bahan referensi pengetahuan untuk pembaca yang sedang belajar atau melakukan penelitian tentang topik yang sama dengan penulis.

Tangerang, 02 Mei 2023

Penulis

DAFTAR ISI

	Hal
ABSTRAK	i
<i>ABSTRACT</i>	ii
PRAKATA	iii
DAFTAR ISI	v
DAFTAR TABEL	viii
DAFTAR GAMBAR	ix
BAB 1 PENDAHULUAN	1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Tujuan dan Manfaat	2
1.3.1. Tujuan Penelitian	2
1.3.2. Manfaat Penelitian	3
1.4. Ruang Lingkup	4
1.5. Metodologi Penelitian	4
1.6. Sistematika Penulisan	5
BAB 2 LANDASAN TEORI	7
2.1. Landasan Teori	7
2.1.1. Twitter	7
2.1.2. <i>Text Mining</i>	7
2.1.3. <i>Sentiment Analysis</i>	8
2.1.4. Pembobotan TF-IDF	8
2.1.5. Support Vector Machine (SVM)	9
2.1.6. Python	10
2.1.7. <i>K-Fold Cross Validation</i>	10
2.1.8. <i>Confusion Matrix</i>	11
2.1.9. Pearson Correlation	12

2.2.	Penelitian Terdahulu.....	13
2.3.	Kerangka Berpikir	17
BAB 3 ANALISIS DAN PERANCANGAN		19
3.1.	Analisis Kebutuhan Fungsional dan Non Fungsional	19
3.1.1.	Analisis Kebutuhan Fungsional	19
3.1.2.	Analisis Kebutuhan Non Fungsional	19
3.2.	Pemilihan Algoritma Pemecahan	20
3.3.	Perancangan User Interface	20
3.4.	Perancangan Proses	21
3.4.1.	<i>Scrap Data</i>	24
3.4.2.	Labeling.....	25
3.4.3.	<i>Pre-processing</i>	25
3.4.4.	<i>Support Vector Machine</i>	29
3.5.	Perancangan Sarana Pendukung.....	37
3.5.1.	Kamus Kata Dasar.....	37
3.5.2.	Kamus Sinonim Kata	37
3.5.3.	Kamus Kata Tidak Baku	38
3.6.	Jadwal Pengerjaan	39
BAB 4 IMPLEMENTASI DAN EVALUASI		40
4.1.	Implementasi User Interface.....	40
4.2.	Implementasi Metode dan Algoritma.....	44
4.3.	Pengujian Sistem	49
4.3.1.	Hasil Train Model	49
4.3.2.	Hasil Test Model	65
4.3.3.	Pearson Moment Product Correlation.....	73
BAB 5 SIMPULAN DAN SARAN.....		78
5.1.	Simpulan.....	78
5.2.	Saran	79
DAFTAR PUSTAKA		80
RIYAWAT HIDUP.....		82

LAMPIRAN	83
----------------	----

DAFTAR TABEL

Tabel 2. 1 Contoh Tabel Confusion Matrix	11
Tabel 2. 2 Tabel Penelitian Terdahulu	13
Tabel 3. 1 Tabel Hasil Lowercasing	25
Tabel 3. 2 Tabel Hasil Remove Special Character and Number.....	26
Tabel 3. 3 Tabel Hasil Stemming.....	26
Tabel 3. 4 Tabel Hasil Tokenize	27
Tabel 3. 5 Tabel Hasil Remove Stopwords.....	27
Tabel 3. 6 Tabel Hasil Slangwords Filtering	28
Tabel 3. 7 Tabel Hasil Remove Synonim	28
Tabel 3. 8 Tabel Contoh Corpus	29
Tabel 3. 9 Tabel Perhitungan TF-IDF.....	30
Tabel 3. 10 Contoh Data 2 Dimensi.....	31
Tabel 3. 11 Tabel Jarak Euclidean Distance Per Titik.....	32
Tabel 3. 12 Tabel Hasil Nilai Kernel	33
Tabel 3. 13 Contoh Dataset untuk 10-Fold Cross Validation	34
Tabel 3. 14 Tabel Hasil Pembagian Data Training dan Data Testing.....	35
Tabel 3. 15 Tabel Confusion Matrix	35
Tabel 3. 16 Hasil Perhitungan f1-score.....	36
Tabel 3. 17 Tabel Sumber Kamus Kata Dasar	37
Tabel 3. 18 Tabel Sumber Sinonim.....	38
Tabel 3. 19 Tabel Sumber Kata Tidak Baku.....	38
Tabel 4. 1 Tabel Sample Data Test dari Dataset Anies Baswedan	49
Tabel 4. 2 Tabel Sample Data Test dari Dataset Ganjar Pranowo.....	54
Tabel 4. 3 Tabel Sample Data Test dari Dataset Prabowo Subianto	60
Tabel 4. 4 Tabel Datatest Twitter Anies Baswedan.....	66
Tabel 4. 5 Tabel Datatest Twitter Ganjar Pranowo.....	68
Tabel 4. 6 Tabel Datatest Twitter Prabowo Subianto	71
Tabel 4. 7 Tabel Data Survei Elektabilitas.....	73
Tabel 4. 8 Tabel Data Sentiment dan Elektabilitas Anies Baswedan	75
Tabel 4. 9 Tabel Korelasi Anies Baswedan	75
Tabel 4. 10 Tabel Data Sentiment dan Elektabilitas Ganjar Pranowo.....	76
Tabel 4. 11 Tabel Korelasi Ganjar Pranowo.....	76
Tabel 4. 12 Tabel Data Sentiment dan Elektabilitas Prabowo Subianto.....	77
Tabel 4. 13 Tabel Korelasi Prabowo Subianto.....	77

DAFTAR GAMBAR

Gambar 2. 1 Contoh Hyperplane Support Vector Machine	9
Gambar 2. 2 Ilustrasi Cara Kerja 10-Fold Cross Validation.....	11
Gambar 2. 3 Gambar Kerangka Berpikir	17
Gambar 3. 1 Rancangan User Interface Halaman Get Data.....	20
Gambar 3. 2 Rancangan User Interface Halaman Train	21
Gambar 3. 3 Rancangan User Interface Halaman Predict.....	21
Gambar 3. 4 Flowchart Get Data	22
Gambar 3. 5 Flowchart Train	23
Gambar 3. 6 Flowchart Predict	24
Gambar 3. 7 Visualisasi Data 2 Dimensi	32
Gambar 3. 8 Gantt Chart Jadwal Pengerjaan	39
Gambar 4. 1 Gambar Tampilan Get Data	40
Gambar 4. 2 Gambar Tampilan Get Data Download.....	41
Gambar 4. 3 Gambar Tampilan Train	41
Gambar 4. 4 Gambar Tampilan Train Upload File	42
Gambar 4. 5 Gambar Tampilan Predict	42
Gambar 4. 6 Gambar Tampilan Predict Input Text	43
Gambar 4. 7 Gambar Tampilan Predict Input File.....	43
Gambar 4. 8 Gambar Split Data dengan K-Fold Cross Validation.....	44
Gambar 4. 9 Gambar Pembobotan dan Train Mesin.....	45
Gambar 4. 10 Gambar Tampilan Persiapan Data.....	45
Gambar 4. 11 Gambar Implementasi Confusion Matrix.....	46
Gambar 4. 12 Gambar Pencarian Fold Terbaik berdasarkan f1-score	46
Gambar 4. 13 Gambar Plot Confusion Matrix dan Hyperplane SVM.....	47
Gambar 4. 14 Gambar Implementasi WordCloud	47
Gambar 4. 15 Gambar Implementasi Tabel DF	48
Gambar 4. 16 Gambar Save Model.....	48
Gambar 4. 17 Gambar f1-score Anies Baswedan	51
Gambar 4. 18 Grafik Score setiap Fold Anies Baswedan	51
Gambar 4. 19 Grafik Pie Jumlah Tiap Sentiment Anies Baswedan	52
Gambar 4. 20 Confusion Matrix Anies Baswedan	52
Gambar 4. 21 Gambar WordCloud Positif Anies Baswedan	53
Gambar 4. 22 Gambar WordCloud Netral Anies Baswedan	53
Gambar 4. 23 Gambar WordCloud Negatif Anies Baswedan	54
Gambar 4. 24 Gambar f1-score Ganjar Pranowo.....	56
Gambar 4. 25 Grafik Score setiap Fold Ganjar Pranowo.....	57
Gambar 4. 26 Grafik Pie Jumlah Tiap Sentiment Ganjar Pranowo	57
Gambar 4. 27 Confusion Matrix Ganjar Pranowo	58
Gambar 4. 28 Gambar WordCloud Positif Ganjar Pranowo	58
Gambar 4. 29 Gambar WordCloud Netral Ganjar Pranowo	59
Gambar 4. 30 Gambar WordCloud Negatif Ganjar Pranowo	59
Gambar 4. 31 Gambar f1-score Prabowo Subianto	62
Gambar 4. 32 Grafik Score setiap Fold Prabowo Subianto	62

Gambar 4. 33 Grafik Pie Jumlah Tiap Sentiment Prabowo Subianto	63
Gambar 4. 34 Confusion Matrix Prabowo Subianto	63
Gambar 4. 35 Gambar WordCloud Positif Prabowo Subianto	64
Gambar 4. 36 Gambar WordCloud Netral Prabowo Subianto.....	64
Gambar 4. 37 Gambar WordCloud Negatif Prabowo Subianto.....	65
Gambar 4. 38 Grafik Pie Sentiment Test Twitter Anies Baswedan.....	65
Gambar 4. 40 Gambar WordCloud Negatif Twitter Anies Baswedan.....	67
Gambar 4. 39 Gambar WordCloud Positif Twitter Anies Baswedan	67
Gambar 4. 41 Grafik Pie Sentiment Test Twitter Ganjar Pranowo	68
Gambar 4. 42 Gambar WordCloud Positif Twitter Ganjar Pranowo	70
Gambar 4. 43 Gambar WordCloud Negatif Twitter Ganjar Pranowo	70
Gambar 4. 44 Grafik Pie Sentiment Test Twitter Prabowo Subianto	71
Gambar 4. 45 Gambar WordCloud Positif Twitter Prabowo Subianto	72
Gambar 4. 46 Gambar WordCloud Netral Twitter Prabowo Subianto.....	73

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Pemilu terjadi setiap 5 tahun sekali dengan jumlah pemilih yang semakin bertambah setiap periodenya, tercatat penambahan pemilih tetap sejak tahun 2004 sampai 2019 sebanyak 28,90% [1]. Media sosial seperti Facebook, Twitter, LinkedIn, YouTube sudah menjadi bagian dari kehidupan manusia, dibuktikan dengan cara berinteraksi kebanyakan dapat dilakukan dengan sosia media, sampai menjadi identitas diri ke masyarakat umum [2].

Pada sosial media terutama *Twitter*, banyak penyebaran informasi, hiburan, bahkan opini politik, hal ini tidak luput dari terjadinya pemilu yang merupakan kegiatan politik yang akan dilaksanakan pada tahun 2024, disebut sebagai pesta demokrasi, hal tersebut pastinya ramai diperbincangkan di sosial media. Salah satu cara menyampaikan opini dan lewat media sosial adalah dengan menggunakan tanda pagar tertentu pada *tweet* yang dipost di *Twitter*. *Twitter* sendiri digunakan karena kemudahan dalam membalas *tweet* dengan *like*, komentar bahkan *reply*, sehingga topik yang sedang ramai diperbincangkan dapat dengan mudah tersebar [3].

Oleh karena itu, maka diperlukannya analisa lebih lanjut untuk melihat seberapa besar kesetujuan dan ketidaksetujuan pengguna *Twitter* terhadap pemilihan calon presiden 2024, lalu data akan diperoleh dari *tweet* pengguna *Twitter* yang berkaitan dengan topik ini yaitu pemilihan umum 2024. *Sentiment Analysis* atau dapat disebut *Opinion Mining* merupakan studi yang mempelajari opini, sentimen, evaluasi, tingkah laku dan emosi terhadap suatu entitas [2] [3].

Sentiment Analysis ini digunakan untuk mengetahui *tweet* yang diposting mengandung kata positif, negatif atau netral [4].

Pada penelitian ini, akan menggunakan metode Support Vector Machine (SVM) yaitu salah satu metode klasifikasi pada *machine learning (supervised learning)* yang memprediksi kelas berdasarkan pola dari hasil training. Klasifikasi dilakukan dengan membuat garis pemisah (*hyperlane*) antara kelas positif dan negatif [2]. Sehingga hasil dari penelitian ini adalah sentiment pengguna *Twitter* terhadap pemilihan calon presiden pada tahun 2024.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah disebutkan, maka penulis mendapati rumusan masalah:

1. Bagaimana penerapan *Sentiment Analysis* menggunakan algoritma *Support Vector Machine* (SVM) diimplementasikan pada topik pemilihan calon presiden 2024?
2. Bagaimana akurasi *Sentiment Analysis* untuk tweet pemilihan calon presiden 2024 dengan menggunakan algoritma *10-Fold Cross Validation* dan *Confusion Matrix*?
3. Apakah ada keterkaitan antara hasil analisa dengan survey elektabilitas portal berita?

1.3. Tujuan dan Manfaat

1.3.1. Tujuan Penelitian

Berdasarkan masalah yang dinyatakan diatas, maka tujuan dari penelitian ini adalah:

1. Mengembangkan aplikasi web sederhana sebagai bentuk implementasi *Sentiment Analysis* menggunakan algoritma *Support Vector Machine* (SVM) dengan topik pemilu 2024
2. Menguji akurasi *Sentiment Analysis* dengan algoritma *Support Vector Machine* (SVM) untuk *tweet* pemilu 2024
3. Melakukan perbandingan hasil prediksi dengan website berita terpercaya seperti detik.com

1.3.2. Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Manfaat bagi penulis:
 - a. Hasil penelitian dapat memberikan manfaat bagi penulis tentang penggunaan *Text Mining* dalam memprediksi hasil pemilu 2024 berdasarkan data *Twitter*
 - b. Mengimplementasi ilmu dan teori yang penulis dapatkan selama perkuliahan dalam bidang *Machine Learning*
2. Manfaat bagi akademis:
 - a. Penulis berharap dengan riset yang dilakukan dapat menjadi referensi bacaan dengan topik *Sentiment Analysis* dan *Text Mining*
3. Manfaat bagi masyarakat:
 - a. Sebagai prediksi pemilihan calon presiden 2024 berdasarkan data *Twitter*
 - b. Memberikan referensi dan bacaan bagi masyarakat yang berkaitan dengan tema politik

1.4. Ruang Lingkup

Pada penelitian ini, ruang lingkup yang dibatasi oleh penulis adalah sebagai berikut:

1. Perancangan aplikasi ini dijalankan berbasis website
2. Penerapan *Sentiment Analysis* menggunakan algoritma *Support Vector Machine* (SVM)
3. Tema pemilu yang diambil hanya seputar pemilihan calon presiden, bukan calon lainnya.
4. Pengujian algoritma akan dilakukan dengan metode *10-fold cross validation*
5. Bahasa pemrograman yang digunakan dalam perancangan aplikasi ini adalah bahasa pemrograman python.
6. Sentimen yang diteliti penulis dibagi menjadi 3 yaitu positif, netral, dan negatif
7. Data diambil dari *Twitter* dengan batas 1 Januari 2023 sampai dengan 31 Maret 2023

1.5. Metodologi Penelitian

Metode yang akan dipakai pada penelitian ini adalah algoritma *Support Vector Machine* (SVM) yang akan diterapkan pada data *tweet* dari media sosial *Twitter* dengan tema pemilihan capres 2024, lalu data tersebut akan dilakukan beberapa proses diantaranya:

1. *Data Labeling* melakukan pelabelan setiap data menjadi 3 jenis, positif, netral, dan negatif.
2. *Preprocessing* yaitu pembersihan data dari noise sehingga siap digunakan pada tahap *Transformation*

3. *Transformation* menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) sehingga data diubah dari kalimat atau kata menjadi sebuah nilai
4. *Klasifikasi* dilakukan pemrosesan data menggunakan algoritma *Support Vector Machine* (SVM).
5. *Evaluasi* akan dilakukan untuk mengetahui nilai akurasi, presisi, *recall*, dan *F1-score* menggunakan *10-Fold Cross Validation* dan *Confusion Matrix* dengan membagi dataset kedalam data *training* dan data *testing* secara bertahap. Lalu dilakukan validasi dengan dibandingkan dengan label data sebenarnya.

1.6. Sistematika Penulisan

Sistematika penulisan yang akan digunakan adalah sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini menjelaskan tentang latar belakang masalah yang diangkat, rumusan masalah yang didapatkan, batasan masalah yang dibuat, manfaat dan tujuan penelitian bagi berbagai pihak serta sistematika penulisan yang akan digunakan di penelitian ini.

BAB 2 LANDASAN TEORI

Bab ini menjelaskan tentang teori yang berkaitan tentang *Sentiment Analysis* dan algoritma *Support Vector Machine* yang diambil dari berbagai sumber sebagai landasan dalam pembuatan penelitian.

BAB 3 ANALISIS DAN PERANCANGAN

Pada bagian ini berisi tahapan rancangan aplikasi *Sentiment Analysis* dengan algoritma *Support Vector Machine* berbasis website sederhana.

BAB 4 IMPLEMENTASI DAN EVALUASI

Bab ini berisikan hasil pengembangan yang diimplementasi dalam website dan uji akurasi yang diukur dengan metode *10-Fold Cross Validation*

BAB 5 SIMPULAN DAN SARAN

Bab ini berisikan pemaparan kesimpulan dari hasil uji coba dan pendapat penulis tentang riset yang telah dilakukan.

BAB 2

LANDASAN TEORI

2.1. Landasan Teori

2.1.1. Twitter

Twitter merupakan salah satu sosial media yang digunakan untuk berinteraksi secara daring dengan pengguna lain dan atau bot. Pada twitter sendiri, terdapat istilah tweet yaitu teks status pengguna yang digunakan untuk memberikan informasi [5].

2.1.2. *Text Mining*

Text mining menggunakan proses *preprocessing* dokumen, proses ini dapat menghasilkan sentimen kelas positif, negatif, dan netral [6] *Text mining* menurut Nurhuda dan Sihwi, proses menemukan informasi dari sekumpulan dokumen teks menggunakan metode analisis tertentu [5]. Umumnya, tahapan pada *text mining* diawali dengan tahapan *preprocessing* sebagai berikut:

- A. *Lowercasing* merupakan tahapan merubah semua huruf menjadi huruf kecil [7].
- B. *Punctuation Removal* merupakan tahapan menghapus semua tanda baca karena umumnya tidak mempengaruhi makna sentimen [7]
- C. *Tokenizing* merupakan tahapan pemotongan string input per kata menjadi token [7]. Contohnya “Saya Ingin Makan” akan menjadi “Saya”, “Ingin”, “Makan”.
- D. *Slang Word Conversion* merupakan tahapan mengubah bahasa gaul atau singkatan menjadi bahasa baku [7]

- E. *Stop Word Removal* merupakan tahapan menghapus kata penghubung atau *stopwords* yang tidak membawa makna sentimen apapun [7].
- F. *Stemming* merupakan tahapan menghapus imbuhan dan mengambil kata dasar yang tersisa [7].
- G. *Synonym* merupakan tahapan menggabungkan beberapa kata yang memiliki arti serupa atau sejenis [8].

2.1.3. Sentiment Analysis

Sentiment analysis juga bisa disebut *Opinion Mining* merupakan sebuah studi komputasi yang meneliti opini, pendapat, perilaku, dan emosi seseorang secara tekstual lalu diklasifikasi menjadi kelompok sentiment negatif, netral dan positif terhadap suatu individu, kejadian, atau topik [9]

2.1.4. Pembobotan TF-IDF

Bobot TF-IDF merupakan penilaian bobot setiap kata yang ada pada sebuah dokumen, TF-IDF merupakan metode pembobotan fitur yang kuat yang mengukur frekuensi dan distribusi kata untuk menentukan bobotnya. TF-IDF terdiri dari TF (*Term Frequency*) yaitu frekuensi kemunculan kata dalam dokumen dan IDF (*Inverse Document Frequency*) yaitu kebalikan dari frekuensi dokumen, nilai TF-IDF dapat diperoleh dengan menggunakan persamaan berikut:

1. Nilai *Term Frequency* (TF) diperoleh dengan nilai frekuensi kemunculan fitur t pada dokumen d [10]

$$TF_t = (t, d) \quad (1)$$

2. Nilai *Inverse Document Frequency* (IDF) diperoleh dari logaritma banyaknya dokumen n dibagi dokumen df yang mengandung fitur t [10]

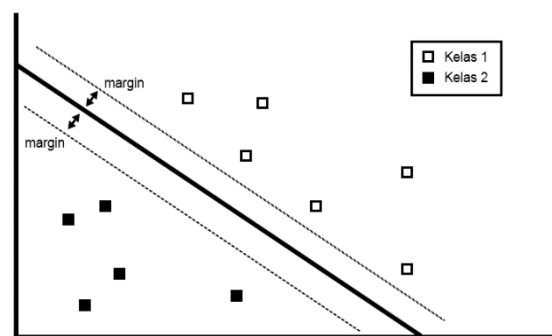
$$IDF_t = \log \frac{n}{df(t)} + 1 \quad (2)$$

3. Nilai Term Frequency Inverse Document Frequency (TF-IDF) didapatkan dengan mengalikan nilai TF dan IDF [10]

$$W_t = TF_t \cdot IDF_t \quad (3)$$

2.1.5. Support Vector Machine (SVM)

Support Vector Machine diciptakan oleh Boser, Guyon, dan Vapnik [8] yang merupakan salah satu metode klasifikasi dengan tujuan menemukan *hyperplane* terbaik untuk memisahkan 2 kelas [11]. Algoritma ini akan menghasilkan *hyperplane* yang memisahkan sebuah plane menjadi 2 bagian pada masing-masing kelas yang dapat dilihat pada gambar 1.1



Gambar 2. 1 Contoh Hyperplane Support Vector Machine

Beberapa kernel yang dapat diterapkan pada SVM:

1. *Linear* Kernel merupakan fungsi kernel yang memisahkan 2 kelas dengan persamaan sebagai berikut:

$$[(w^t \cdot x_i) + b] \geq 1 \text{ for } y_i = 1 \text{ dan } [(w^t \cdot x_i) + b] \leq -1 \text{ for } y_i = -1 \quad (4)$$

Deskripsi :

x_i = dataset *training*

$i = 1, 2, \dots, n$

y_i = label dari x_i [8]

2. *Polynomial* Kernel merupakan fungsi kernel yang tidak terpisah secara *linear*, persamaan sebagai berikut:

$$K(x_i, x_j) = (x_i, x_j + c)^d \quad (5)$$

Deskripsi :

x_i, x_j = *training* data

$c, d > 0$ = konstanta [8]

3. *Radial Basis Function* (RBF) Kernel merupakan fungsi untuk menganalisis data yang tidak terpisah secara linear dan memerlukan 2 jenis parameter (Gamma dan Cost).

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

Deksripsi :

$\|x_i - x_j\|$ = *Euclidean Distance*

2σ = Parameter Independen menentukan tingkat pengurangan

$K(x_i, x_j)$ menuju 0 [8]

2.1.6. Python

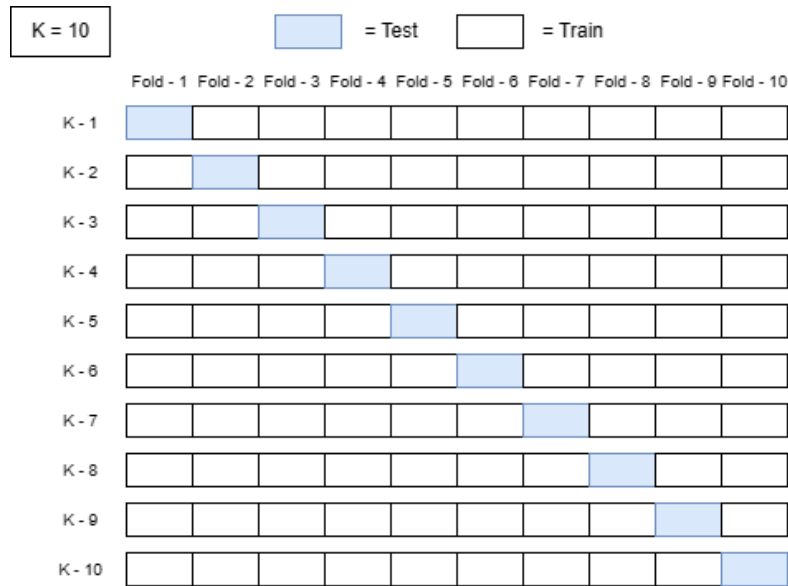
Python merupakan bahasa pemrograman digunakan oleh banyak platform oleh developer seperti website, mobile dan desktop [12]. Namun sejak berkembangnya *Machine Learning* python pun turut digunakan terutama dengan ketersediaan banyaknya *library* seperti scikit-learn, keras, tensorflow, openCV, dan lain-lain.

2.1.7. K-Fold Cross Validation

K-Fold Cross Validation merupakan salah satu bentuk pengujian yang berfungsi untuk menilai kinerja algoritma dengan cara kerja membagi sampel

data secara acak lalu membagi data tersebut sebanyak nilai K dari total *fold*. Selanjutnya, salah satu kelompok K tersebut akan dijadikan sebagai data *testing* dan sisa kelompok akan dijadikan sebagai data *training* [13].

Contoh ilustrasi cara kerja K -Fold Cross Validation ditampilkan pada gambar 1.2



Gambar 2. 2 Ilustrasi Cara Kerja 10-Fold Cross Validation

2.1.8. Confusion Matrix

Confusion Matrix merupakan sebuah metode berupa matriks yang berfungsi untuk menilai kinerja akurasi klasifikasi berdasarkan dataset dan label sebenarnya.

Penerapan confusion matrix terlihat pada tabel berikut:

Tabel
2. 1

	Prediksi Positif	Prediksi Netral	Prediksi Negatif
Label Positif	True Positif	False Netral	False Negatif
Label Netral	False Positif	True Netral	False Negatif
Label Negatif	False Positif	False Netral	True Negatif

Contoh Tabel *Confusion Matrix*

Dengan keterangan True Negatif adalah hasil prediksi dan label sebenarnya negatif. False Negatif adalah hasil prediksi adalah negatif sedangkan label sebenarnya bukan negatif. True Netral adalah hasil prediksi dan label sebenarnya netral. False Netral adalah hasil prediksi netral sedangkan label sebenarnya bukan netral. True Positif adalah hasil prediksi dan label sebenarnya positif. False Positif adalah hasil prediksi positif sedangkan label sebenarnya bukan positif [6].

2.1.9. Pearson Correlation

Pearson Correlation merupakan metode untuk mencari hubungan linear antara 2 variable atau lebih. Hasil dari Pearson Correlation merupakan koefisien korelasi yang berkisar antara angka 0 sampai 1 yang berlaku untuk angka negatif [8]. Dalam menguji validitasnya menggunakan Pearson Product Moment Correlation dengan persamaan sebagai berikut:

$$r_{x,y} = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{\{n \sum x^2 - (\sum x)^2\} \{n \sum y^2 - (\sum y)^2\}}} \quad (7)$$

Dimana:

n = Jumlah Data

$\sum xy$ = Hasil Jumlah antara X dan Y setiap pasang

$\sum x$ = Hasil Jumlah Variable X

$\sum y$ = Hasil Jumlah Variable Y

$\sum x^2$ = Hasil Jumlah Variable X Kuadrat

$\sum y^2$ = Hasil Jumlah Variable Y Kuadrat

2.2. Penelitian Terdahulu

Berikut ini merupakan penelitian yang menjadi referensi penulis dalam melakukan penelitian ini:

Tabel 2. 2 Tabel Penelitian Terdahulu

No.	Penulis	Tahun	Judul	Kesimpulan
1	Dionisia Bhisetya Rarasati, Josef Cristian Adi Putra	2021	Correlation Between Twitter Sentiment Analysis with Three Kernels Using Algorithm Support Vector Machine (SVM) Governor Candidate	Pada algoritma Support Vector Machine dengan studi kasus pemilihan gubernur DKI Jakarta, kernel dengan akurasi terbaik adalah Gaussian RBF 90.58%, diikuti dengan Linear 85.87%, dan Polynomial 78.5

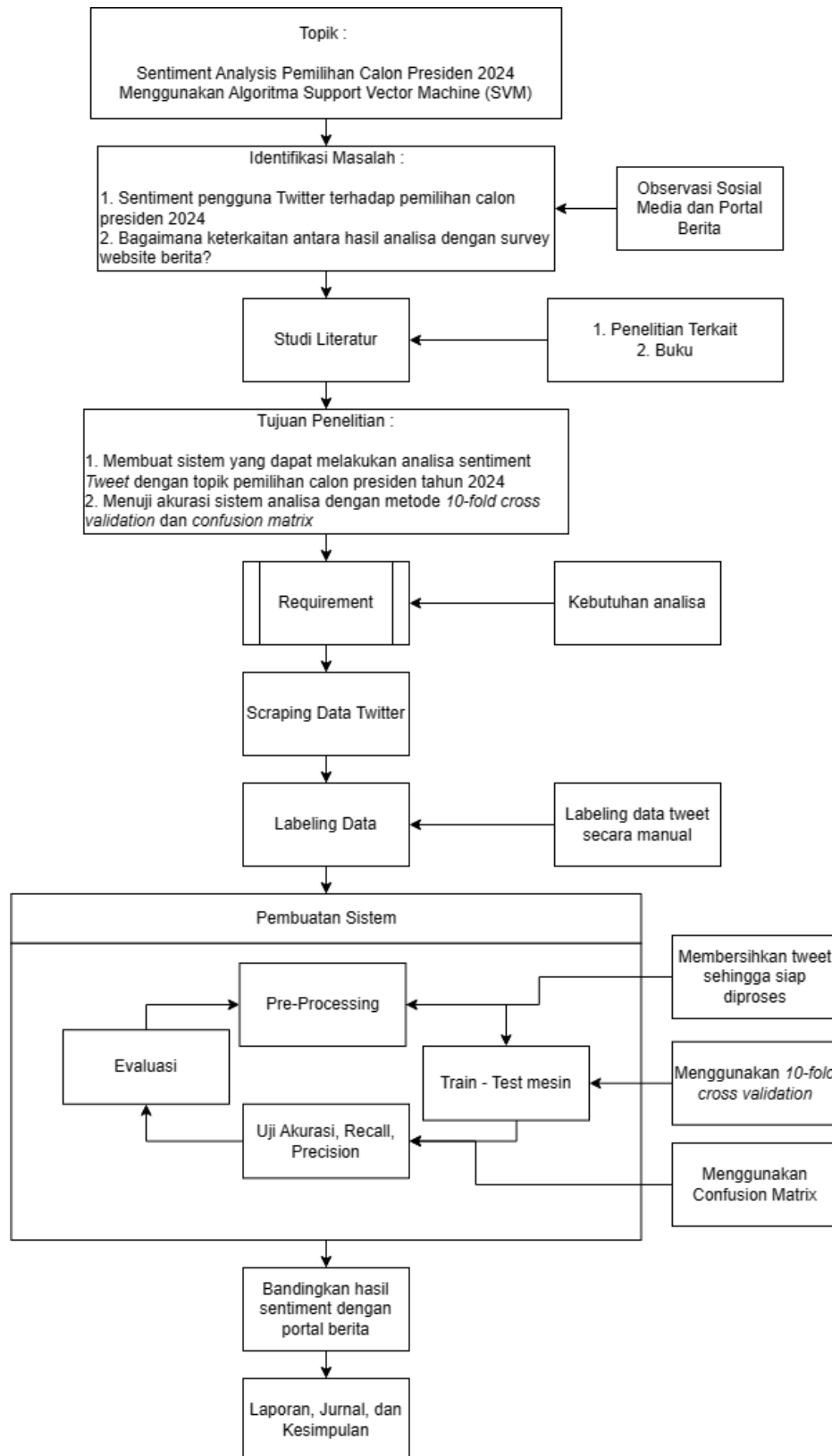
			Electability Level	
2	Hendrik Setiawan, Ema Utami, Sudarmawan	2021	Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunaka n Algoritma Support Vector Machine dan Naive Bayes	Penelitian analisis sentiment pengguna twitter terhadap kuliah online pasca covid-19 memiliki akurasi sebesar 85% dengan algoritma SVM, sedangkan akurasi 81.2% menggunakan <i>Naïve Bayes</i> .
3	Sri Diantika, Windu Gata, Hiya Nalatissifa, Mareanus Lase	2021	Komparasi Algoritma SVM Dan Naive Bayes Untuk Klasifikasi Kestabilan Jaringan Listrik	Akurasi algoritma <i>Support Vector Machine</i> pada data kestabilan jaringan mendapatkan akurasi 98.8% semestara jika menggunakan <i>Naïve Bayes</i> sebesar 97.64%.

4	Elly Indrayuni	2018	KOMPARAS I ALGORITMA A NAIVE BAYES DAN SUPPORT VECTOR MACHINE UNTUK ANALISA SENTIMEN REVIEW FILM	Penelitian analisis sentiment pengguna twitter terhadap review film menggunakan <i>confussion matrix</i> menunjukkan bahwa akurasi algoritma SVM sebesar 90% dan untuk algoritma <i>Naïve Bayes</i> sebesar 84.5%
5	Dedi Darwis, Eka Shintya Pratiwi, A. Ferico Octaviansya h Pasaribu	2020	PENERAPAN N ALGORITMA A SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER	Hasil klasifikasi menggunakan metode SVM dibagi menjadi 3 kelas, yaitu 8% positif, 15% netral, dan 77% negatif. Lalu hasil pengujian nilai akurasi, precession, recall, dan F1-score dapat disimpulkan bahwa sentimen masyarakat terhadap kinerja KPK sangat

			KOMISI PEMBERAN TASAN KORUPSI REPUBLIK INDONESIA	kurang baik dengan presentase negatif sebesar 77% serta pengujian hasil akurasi sebesar 82%, precision 90%, recall 88%, dan f1-score 89%
--	--	--	---	---

Sehingga berdasarkan penelitian terdahulu para penulis menggunakan algoritma SVM dan mendapatkan akurasi terbaik dibandingkan algoritma lain dalam melakukan analisis sentimen dengan data dari twitter. Dan untuk kernel yang dipakai akan menggunakan kernel *Gaussian Radial Basis Function* dikarenakan kernel tersebut memiliki tingkat akurasi terbaik berdasarkan jurna [8]. Dikarenakan pemilihan presiden diikuti oleh masyarakat luas, dan pemilihan presiden akan sangat berkaitan dengan pandangan masyarakat terhadap masing-masing calon, oleh sebab itu maka penulis melakukan penelitian *sentiment analysis* dengan algoritma *Support Vector Machine* dalam melakukan analisis sentimen terhadap pemilihan calon presiden tahun 2024.

2.3. Kerangka Berpikir



Gambar 2. 3 Gambar Kerangka Berpikir

Kerangka alur penelitian yang dilakukan dalam membuat sistem *sentiment analysis* terhadap topik pemilu 2024. Dimulai dari mendapatkan dataset *tweet* dari kata kunci yang relevan seperti “Ganjar Pranowo”, “Prabowo Subianto”, dan “Anies Baswedan”, lalu melakukan *labeling tweet* secara manual oleh penulis, dilanjutkan dengan melakukan *pre-processing* pada dataset seperti melakukan *Cleansing, Tokenizing, Filtering, Stemming, Tagging, dan Analyzing*. Selanjutnya baru setiap kata diberi bobot dengan metode *TF-IDF* dan dilanjutkan dengan membagi data kedalam data *training* dan data *testing* menurut metode *10-Fold Cross Validation* untuk dimasukan ke dalam algoritma *Support Vector Machine*. Lalu dari setiap fold akan dilihat untuk nilai K terbaik dari segi akurasi, *precision*, *recall*, dan *f1-score*. Lalu hasil prediksi setiap capres akan dibandingkan secara manual oleh survey yang diterbitkan oleh portal berita untuk melakukan perbandingan hasil prediksi dengan survey yang disebar.

BAB 3

ANALISIS DAN PERANCANGAN

3.1. Analisis Kebutuhan Fungsional dan Non Fungsional

3.1.1. Analisis Kebutuhan Fungsional

Kebutuhan Fungsional merupakan proses yang ada didalam sistem serta apa saja yang dibutuhkan oleh sistem agar dapat berjalan baik. Berikut kebutuhan fungsional dari sistem yang akan dirancang:

1. Sistem akan terdiri dari 3 segment yaitu Get Data, Train, Test dengan data Training 90 persen dan Testing 10 persen
2. Halaman Get Data berfungsi untuk mengambil data bersumber dari twitter
3. Data input sistem akan berupa file excel / csv
4. Sistem akan dikembangkan dalam bentuk website
5. Output program akan berupa file hasil prediksi *sentiment analysis*

3.1.2. Analisis Kebutuhan Non Fungsional

Kebutuhan Non Fungsional merupakan kebutuhan yang harus dipenuhi dalam menunjang berjalannya sistem. Berikut kebutuhan perangkat lunak dan keras dari sistem yang akan dirancang:

1. Ryzen 5 3400G
2. Rx Vega 11 Graphics
3. 16 GB RAM
4. 256 GB SSD
5. 1 TB HDD
6. Windows 10

3.2. Pemilihan Algoritma Pemecahan

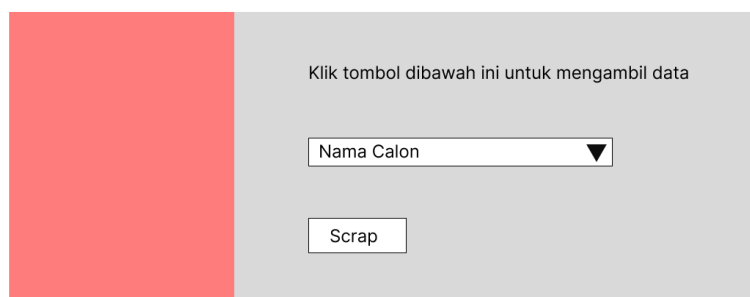
Algoritma yang dipilih adalah Support Vector Machine dikarenakan Support Vector Machine merupakan algoritma text mining yang memiliki akurasi tinggi [14] terlihat pada jurnal ini, algoritma *Support Vector Machine* memiliki akurasi, *precision*, dan *recall* lebih tinggi dibandingkan dengan algoritma *Naïve Bayes*, lalu untuk jurnal [15] diketahui bahwa tingkat akurasi, *precision*, dan *recall* algoritma *Support Vector Machine* memiliki akurasi lebih tinggi daripada algoritma *K-Nearest Neighbor* (K-NN) dan *Naïve Bayes*, sehingga penulis memilih algoritma Support Vector Machine sebagai algoritma untuk memprediksi hasil sentiment pemilihan capres 2024 melalui sosial media twitter. Lalu untuk validasi akan dilakukan dengan *10-Fold Cross Validation*, penulis melakukan 10 iterasi dikarenakan tingkat akurasi pada $k = 10$ memiliki akurasi tertinggi [16] .

3.3. Perancangan User Interface

Berikut Gambar 3.1 sampai 3.3 merupakan rancangan User Interface yang berbentuk wireframe, rancangan ini merupakan gambaran bentuk website akan dibangun

Get Data Train and Test Machine Result

Sentiment Analysis Pemilihan Calon Presiden 2024



Klik tombol dibawah ini untuk mengambil data

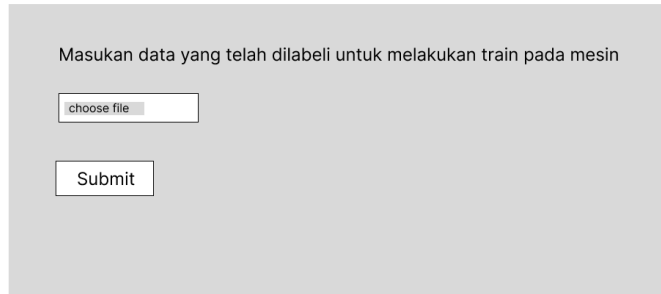
Nama Calon ▼

Scrap

Gambar 3. 1 Rancangan User Interface Halaman Get Data

Get Data **Train and Test Machine** Result

Sentiment Analysis Pemilihan Calon Presiden 2024

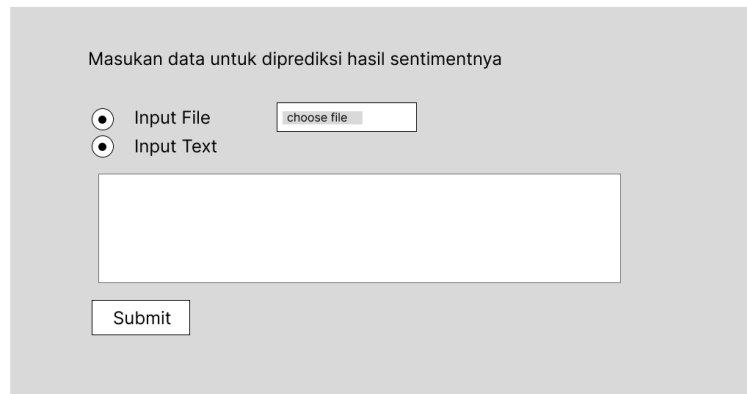


Masukan data yang telah dilabeli untuk melakukan train pada mesin

Gambar 3. 2 Rancangan User Interface Halaman Train

Get Data Train and Test Machine **Predict**

Sentiment Analysis Pemilihan Calon Presiden 2024



Masukan data untuk diprediksi hasil sentimentnya

☒ Input File

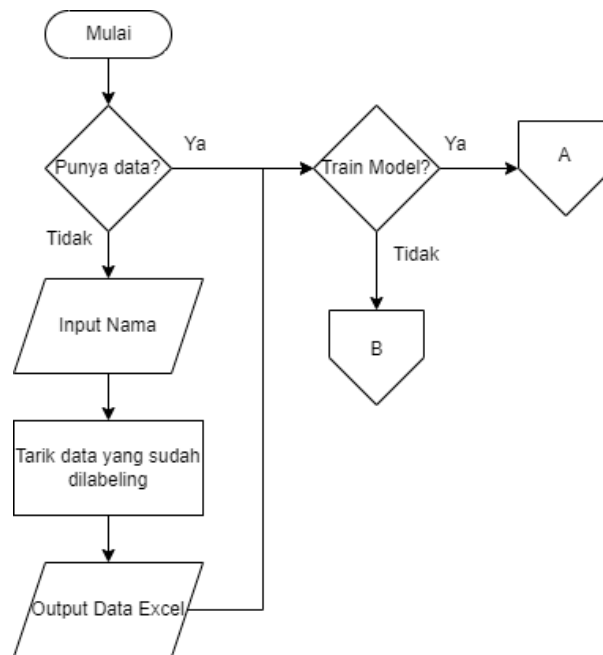
☐ Input Text

Gambar 3. 3 Rancangan User Interface Halaman Predict

3.4. Perancangan Proses

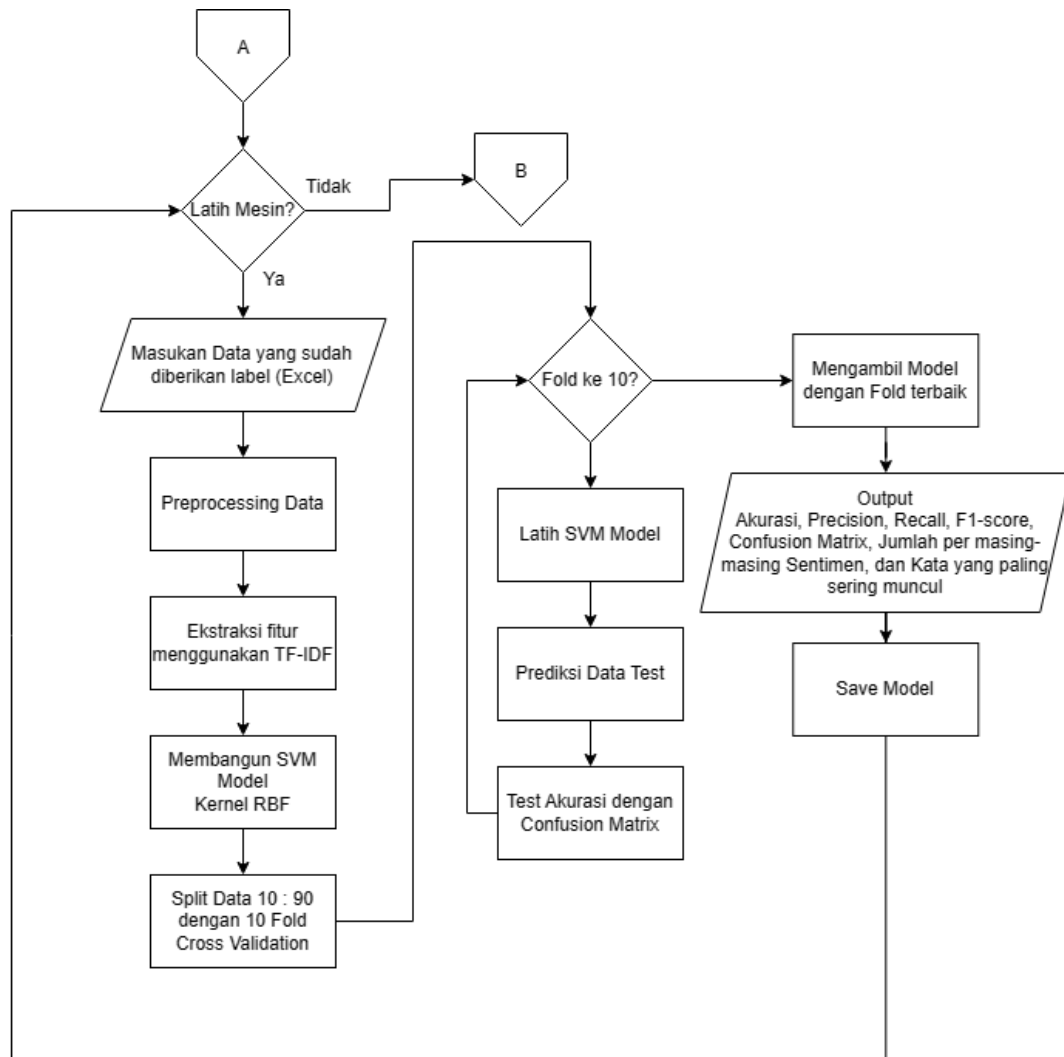
Gambar 3.4 sampai 3.6 memuat gambaran alur proses sistem yang dibagi menjadi 3 bagian inti yaitu “Get Data”, “Train Model” dan “Predict”.

Pada gambar 3.4 dijelaskan pada proses penarikan data yang bersumber dari *Twitter* sehingga user akan diminta memasukan nama calon, yang selanjutnya data untuk calon tersebut yang telah dilabeli akan terdownload untuk user dan dapat dilanjutkan ke proses train model.



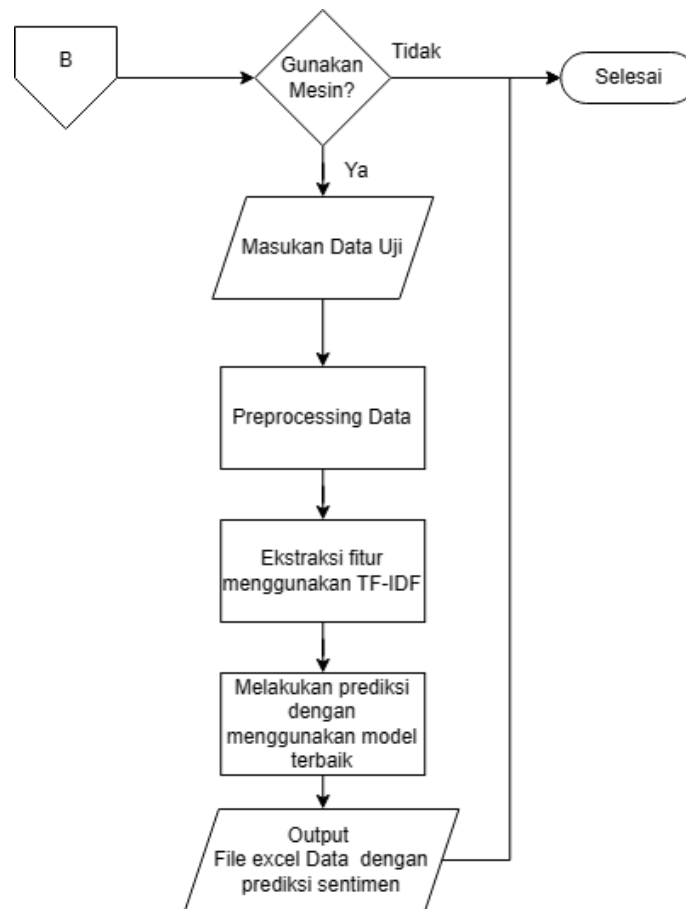
Gambar 3. 4 Flowchart Get Data

Selanjutnya pada gambar 3.5 dijelaskan proses train model SVM yang akan dilakukan penyimpanan model dengan nilai *f1-score* terbesar. Proses dimulai dengan memasukan data yang telah diberikan label dari tahap sebelumnya. Lalu dilakukan pre-processing pada setiap baris data. Lalu fitur data hasil pre-processing akan diekstraksi menggunakan TF-IDF. Hasil fitur ekstraksi lalu akan dimasukan kedalam model SVM dengan kernel Gaussian RBF yang dimasukan dengan metode 10 Fold Cross Validation sehingga data dibagi menjadi 10:90 dengan 90 data training dan 10 data testing, lalu dari train tersebut diambil nilai *f1-score* untuk menentukan fold terbaik dengan confusion matrix. Selanjutnya model terbaik akan disave untuk digunakan pada halaman selanjutnya.



Gambar 3. 5 Flowchart Train

Selanjutnya pada gambar 3.6 ditunjukan cara mesin melakukan prediksi dengan model yang sudah di save saat melakukan train mesin. Awalnya, data yang akan diprediksi dimasukan dalam bentuk teks atau file, lalu data tersebut dilakukan pre-processing, dan fiturnya diekstraksi menggunakan TF-IDF. Hasil fitur ekstraksi tersbut lalu diprediksi menggunakan model yang telah disave pada tahap sebelumnya. Output dari predict ini adalah hasil teks pre-procesing dan sentiment untuk inputan teks, sedangkan file excel berisikan *tweet* dan label hasil prediksi mesin.



Gambar 3. 6 Flowchart Predict

3.4.1. Scrap Data

Tahapan awal dari penelitian ini adalah melakukan *scraping data* bersumber dari *Twitter* dengan kata kunci “(Nama Calon) Presiden” untuk mempersempit konteks dari *Tweet* pengguna *Twitter*. Proses pengumpulan data dilakukan dengan bantuan library *snsrape* oleh python. Data dikumpulkan mulai dari tanggal 1 Januari 2023 sampai 31 Maret 2023.

Proses *scraping* ini menghasilkan total 18.267 data dengan 7.810 data yang duplikat dan tidak relevan dengan tema “Pemilihan Calon Presiden 2024” sehingga menghasilkan total 10.457 data bersih dari 3 calon presiden.

3.4.2. Labeling

Pada tahap ini, dilakukan proses pemberian label sentiment secara manual oleh penulis terhadap data bersih *tweet*. Pemberian label berdasarkan asumsi penulis terhadap *tweet* sehingga proses ini dapat berbeda pada setiap pengguna.

Pada penelitian ini, penulis menggunakan 3 jenis sentimen yaitu positif dilambangkan dengan angka 1, netral dilambangkan dengan angka 0, dan negatif yang dilambangkan dengan angka -1.

3.4.3. Pre-processing

Tahapan *pre-processing* berfungsi untuk membersihkan dan mentransformasi data sehingga dapat diproses oleh sistem dengan mudah dan akurat. Beberapa tahapan *pre-processing* pada penelitian ini diantaranya:

3.4.3.1. Lowercasing

Lowercasing merupakan tahapan untuk merubah semua huruf menjadi huruf kecil sehingga kata yang diproses menjadi seragam. Tabel 3.1 menampilkan contoh proses *lowercasing* dari beberapa *tweet* Anies Baswedan:

Tabel 3. 1 Tabel Hasil *Lowercasing*

No	Kalimat Awal	Hasil Proses
1	PKS Mendukung Anies Baswedan Menjadi Calon Presiden 2024	pks mendukung anies baswedan menjadi calon presiden 2024
2	Anies Baswedan Presiden Indonesia	anies baswedan presiden indonesia
3	Walau sekarang didukung Demokrat dan PKS, ada empat alasan Anies Baswedan gak bakal bisa menang jadi presiden !!	walau sekarang didukung demokrat dan pks, ada empat alasan anies baswedan gak bakal bisa menang jadi presiden !!
4	Anies Baswedan lh yg pantas jadi presiden	anies baswedan lh yg pantas jadi presiden
5	Pak Anies Baswedan calon presiden 2024	pak anies baswedan calon presiden 2024

3.4.3.2. Remove Special Character and Number

Tahapan ini berfungsi menghasilkan kalimat tanpa tanda baca, *emoticon*, dan angka. Tabel 3.2 menampilkan contoh proses *Remove Special Character and Number*:

Tabel 3. 2 Tabel Hasil *Remove Special Character and Number*

No	Kalimat Awal	Hasil Proses
1	pks mendukung anies baswedan menjadi calon presiden 2024	pks mendukung anies baswedan menjadi calon presiden
2	anies baswedan presiden indonesia	anies baswedan presiden indonesia
3	walaupun sekarang didukung demokrat dan pks, ada empat alasan anies baswedan gak bakal bisa menang jadi presiden !!	walaupun sekarang didukung demokrat dan pks ada empat alasan anies baswedan gak bakal bisa menang jadi presiden
4	anies baswedan lh yg pantas jadi presiden	anies baswedan lh yg pantas jadi presiden
5	pak anies baswedan calon presiden 2024	pak anies baswedan calon presiden

3.4.3.3. Stemming

Tahap ini, semua kata berimbuhan akan diproses menjadi kata dasar, seperti ditampilkan pada tabel 3.3 berikut:

Tabel 3. 3 Tabel Hasil *Stemming*

No	Kalimat Awal	Hasil Proses
1	pks mendukung anies baswedan menjadi calon presiden	pks dukung anies baswedan jadi calon presiden
2	anies baswedan presiden indonesia	anies baswedan presiden indonesia
3	walaupun sekarang didukung demokrat dan pks, ada empat alasan anies baswedan gak bakal bisa menang jadi presiden	walaupun sekarang dukung demokrat dan pks ada empat alasan anies baswedan gak bakal bisa menang jadi presiden
4	anies baswedan lh yg pantas jadi presiden	anies baswedan lh yg pantas jadi presiden
5	pak anies baswedan calon presiden	pak anies baswedan calon presiden

3.4.3.4. Tokenize

Tahap ini berfungsi untuk memecah kalimat menjadi token per kata. Proses ini dilakukan dengan tujuan agar proses selanjutnya dapat diterapkan dalam level kata. Proses ditunjukkan pada tabel 3.4 berikut:

Tabel 3. 4 Tabel Hasil *Tokenize*

No	Kalimat Awal	Hasil Proses
1	pks dukung anies baswedan jadi calon presiden	[[pks] [dukung] [anies] [baswedan] [jadi] [calon] [presiden]]
2	anies baswedan presiden indonesia	[[anies] [baswedan] [presiden] [Indonesia]]
3	walaupun sekarang dukung demokrat dan pks, ada empat alasan anies baswedan gak bakal bisa menang jadi presiden	[[walaupun] [sekarang] [dukung] [demokrat] [dan] [pks] [ada] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [bisa] [menang] [jadi] [presiden]]
4	anies baswedan lh yg pantas jadi presiden	[[anies] [baswedan] [lh] [yg] [pantas] [jadi] [presiden]]
5	pak anies baswedan calon presiden	[[pak] [anies] [baswedan] [calon] [presiden]]

3.4.3.5. *Remove Stopwords*

Stopwords merupakan untuk menghapus kata yang tidak memiliki dampak pada sentimen, namun sering muncul pada kalimat. Tabel 3.5 menunjukkan proses *Remove Stopwords* sebagai berikut:

Tabel 3. 5 Tabel Hasil *Remove Stopwords*

No	Kalimat Awal	Hasil Proses
1	[[pks] [dukung] [anies] [baswedan] [jadi] [calon] [presiden]]	[[pks] [dukung] [anies] [baswedan] [calon] [presiden]]
2	[[anies] [baswedan] [presiden] [Indonesia]]	[[anies] [baswedan] [presiden] [Indonesia]]
3	[[walaupun] [sekarang] [dukung] [demokrat] [dan] [pks] [ada] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [bisa] [menang] [jadi] [presiden]]	[[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [menang] [presiden]]
4	[[anies] [baswedan] [lh] [yg] [pantas] [jadi] [presiden]]	[[anies] [baswedan] [lh] [pantas] [presiden]]
5	[[pak] [anies] [baswedan] [calon] [presiden]]	[[pak] [anies] [baswedan] [calon] [presiden]]

3.4.3.6. Slangwords Filtering

Tahapan ini melakukan penghilangan kata-kata gaul menjadi kata baku.

Kamus gaul diambil dari repository github yang dipost oleh [louisowen6](#) yang ditunjukkan pada Tabel 3.6 sebagai berikut:

Tabel 3. 6 Tabel Hasil *Slangwords Filtering*

No	Kalimat Awal	Hasil Proses
1	[[pks] [dukung] [anies] [baswedan] [calon] [presiden]]	[[pks] [dukung] [anies] [baswedan] [calon] [presiden]]
2	[[anies] [baswedan] [presiden] [Indonesia]]	[[anies] [baswedan] [presiden] [Indonesia]]
3	[[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [gak] [bakal] [menang] [presiden]]	[[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [tidak] [akan] [menang] [presiden]]
4	[[anies] [baswedan] [pantas] [presiden]]	[[anies] [baswedan] [pantas] [presiden]]
5	[[pak] [anies] [baswedan] [calon] [presiden]]	[[pak] [anies] [baswedan] [calon] [presiden]]

3.4.3.7. Remove Synonim

Pada tahapan ini, penulis melakukan penghapusan sinonim dalam kalimat yang sama jika ada. Contoh penerapan proses ini, yaitu:

Tabel 3. 7 Tabel Hasil *Remove Synonim*

No	Kalimat Awal	Hasil Proses
1	[[pks] [dukung] [anies] [baswedan] [calon] [presiden]]	[[pks] [dukung] [anies] [baswedan] [calon] [presiden]]
2	[[anies] [baswedan] [presiden] [Indonesia]]	[[anies] [baswedan] [presiden] [Indonesia]]
3	[[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [tidak] [akan] [menang] [presiden]]	[[sekarang] [dukung] [demokrat] [pks] [empat] [alasan] [anies] [baswedan] [tidak] [akan] [menang] [presiden]]
4	[[alhamdulillah] [mudah] [mudah] [anies] [baswedan] [takdir] [allah] [swt] [presiden] [aamiin] [yra]]	[[alhamdulillah] [mudah] [anies] [baswedan] [takdir] [allah] [swt] [presiden] [aamiin] [yra]]
5	[[pak] [anies] [baswedan] [calon] [presiden]]	[[pak] [anies] [baswedan] [calon] [presiden]]

3.4.4. Support Vector Machine

3.4.4.1. Feature Extraction

Pada tahapan *Feature Extraction*, penulis menggunakan metode TF-IDF (Term Frequency – Inverse Document Frequency) yang melakukan pembobotan data hasil preprocessing berdasarkan banyaknya kemunculan suatu kata pada dokumen. Hasil TF-IDF sendiri berfungsi untuk mengubah data menjadi bentuk numerikal agar fitur-fitur yang telah diproses dapat dimuat kedalam persamaan. Sehingga metode ini dipilih untuk mengekstrak data tersebut. Berikut merupakan gambaran bagaimana ekstraksi fitur menggunakan TF-IDF :

Pertama-tama, memuat corpus hasil *pre-processing* sebagai berikut, tabel sebagai contoh memuat 2 dokumen.

Tabel 3. 8 Tabel Contoh Corpus

Dokumen	Corpus
1	mari dukung anies baswedan presiden republik indonesia warga negara indonesia kakek pahlawan jasa genius kenal publik prestasi hitung hidup negara satu republik indonesia
2	anies baswedan bangga rakyat indonesia gubernur daerah khusus ibukota jakarta cocok presiden republik

Digambarkan pada tabel 3.8 berisi *corpus* untuk mendapatkan nilai TF-IDF dari setiap *term* pada suatu dokumen. Pertama-tama dicari nilai *tf* dari setiap *term* pada suatu dokumen yang digunakan. Pencarian *tf* dilakukan dengan mencari jumlah kemunculan *term* pada suatu dokumen, lalu membaginya dengan jumlah total *term* pada dokumen tersebut. Sedangkan nilai *idf* didapat dengan membagi jumlah dokumen dengan jumlah dokumen yang memuat suatu *term* sehingga nilai *idf* dapat dicari dengan persamaan berikut ini: [10]

$$IDF_t = \log \frac{n}{df(t)} + 1 \quad (7)$$

Dimana:

IDF_t : Nilai *idf* dari suatu *term*

n : Jumlah dokumen

$df(t)$: Jumlah dokumen yang memuat suatu *term*

Selanjutnya setelah nilai *tf* dan *idf* didapatkan, maka nilai TF-IDF dapat dihitung dengan menggunakan persamaan: [10]

$$W_t = TF_t \cdot IDF_t \quad (8)$$

Dimana:

W_t : Bobot suatu *term* pada dokumen

TF_t : Nilai *tf* suatu *term* pada suatu dokumen

IDF_t : Nilai *idf* dari suatu *term*

Berikut ditampilkan hasil perhitungan TF-IDF berdasarkan 2 dokumen diatas.

Tabel 3. 9 Tabel Perhitungan TF-IDF

Term	<i>tf</i>		<i>df</i>	<i>idf</i>	tf-idf	
	Dokumen 1	Dokumen 2			Dokumen 1	Dokumen 2
mari	0,04	0	1	1,30	0,05	0
dukung	0,04	0	1	1,30	0,05	0
anies	0,04	0,08	2	1,00	0,05	0,10
baswedan	0,04	0,08	2	1,00	0,05	0,10
presiden	0,04	0,08	2	1,00	0,05	0,10
republik	0,09	0,08	2	1,00	0,12	0,10
indonesia	0,13	0,08	2	1,00	0,17	0,10
warga	0,04	0	1	1,30	0,05	0
negara	0,09	0	1	1,30	0,12	0
kakek	0,04	0	1	1,30	0,05	0
pahlawan	0,04	0	1	1,30	0,05	0
jasa	0,04	0	1	1,30	0,05	0
genius	0,04	0	1	1,30	0,05	0
kenal	0,04	0	1	1,30	0,05	0
publik	0,04	0	1	1,30	0,05	0
prestasi	0,04	0	1	1,30	0,05	0
hitung	0,04	0	1	1,30	0,05	0

hidup	0,04	0	1	1,30	0,05	0
satu	0,04	0	1	1,30	0,05	0
bangga	0	0,08	1	1,30	0	0,10
rakyat	0	0,08	1	1,30	0	0,10
gubernur	0	0,08	1	1,30	0	0,10
daerah	0	0,08	1	1,30	0	0,10
khusus	0	0,08	1	1,30	0	0,10
ibukota	0	0,08	1	1,30	0	0,10
jakarta	0	0,08	1	1,30	0	0,10
cocok	0	0,08	1	1,30	0	0,10

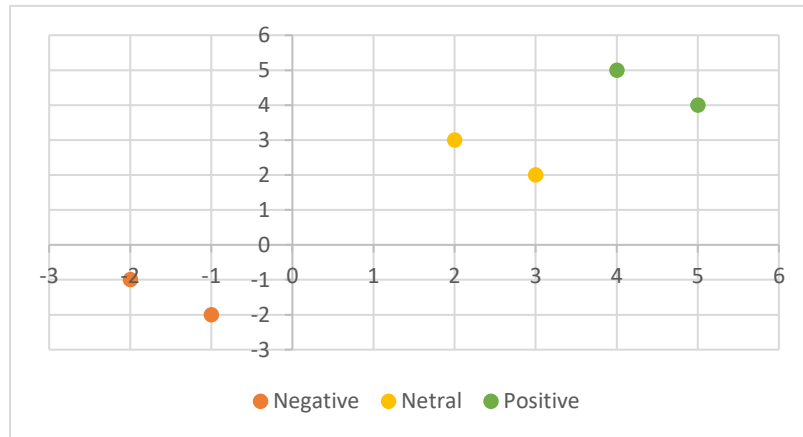
3.4.4.2.SVM

Untuk menggunakan algoritma Support Vector Machine, disini perlu dicari \vec{w} terbaik untuk memisahkan tiap kelas. Normalnya dengan data yang tidak terlalu tersebar, hyperplane berupa linear atau polynomial. Namun jika masuk pada kasus penulis, persebaran data yang sangat luas dan dimensi fitur yang tanpa batas, maka diperlukannya penggunaan *kernel trick*, implementasi yang penulis lakukan untuk mengatasi masalah ini adalah menggunakan Gaussian RBF (Radial Basis Function) yang mampu mengatasi masalah diatas.

Jika diberikan dataset dengan 3 label seperti pada tabel 3.10 juga berikut visualisasi dataset yang ditunjukkan pada gambar 3.7.

Tabel 3. 10 Contoh Data 2 Dimensi

No	X	Y	Label
1	-1	-2	Positif
2	-2	-1	Positif
3	4	5	Netral
4	5	4	Netral
5	7	8	Negatif
6	8	7	Negatif



Gambar 3. 7 Visualisasi Data 2 Dimensi

Dari hasil observasi dataset diatas, dalam penerapan kernel Gaussian RBF (Radial Basis Function) pertama-tama kita perlu menemukan jarak setiap titik dengan titik yang lain. Hal ini dapat dilakukan dengan persamaan *Euclidean Distance*.

$$d_{p,q} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (9)$$

Dimana:

$d_{p,q}$: Jarak *Euclidean Distance*

n : Jumlah titik

p, q : Titik vector data

Hasil perhitungan *Euclidean Distance* ditampilkan pada tabel 3.11 sebagai berikut:

Tabel 3. 11 Tabel Jarak *Euclidean Distance* Per Titik

	1	2	3	4	5	6
1	0,00	1,41	8,60	8,49	12,81	12,73
2	1,41	0,00	8,49	8,60	12,73	12,81

3	8,60	8,49	0,00	1,41	4,24	4,47
4	8,49	8,60	1,41	0,00	4,47	4,24
5	12,81	12,73	4,24	4,47	0,00	1,41
6	12,73	12,81	4,47	4,24	1,41	0,00

Selanjutnya, dilakukan perhitungan per data kedalam fungsi kernel Gaussian RBF dengan persamaan sebagai berikut:

$$K_{(x,x')} = e^{-\gamma \|x-x'\|^2} \quad (10)$$

Dimana:

$K_{(x,x')}$: Nilai Kernel Gaussian RBF

γ : Nilai Parameter (Gamma)

x, x' : Titik vector data

Kita asumsikan data baru yang akan diklasifikasi memiliki vektor (2, 2) sehingga hasil perhitungan setiap data dengan data baru pada tabel 3.12 sebagai berikut:

Tabel 3. 12 Tabel Hasil Nilai Kernel

Data	Kernel Value
1	1,38879E-11
2	1,38879E-11
3	2,26033E-06
4	2,26033E-06
5	3,22134E-27
6	3,22134E-27

Lalu semua data dijumlahkan dengan masing-masing labelnya, pada kasus ini perhitungan akan ditampilkan sebagai berikut:

$$\text{Label Positif} = \text{Data 1} + \text{Data 2} = 2,77759\text{E-11}$$

$$\text{Label Netral} = \text{Data 3} + \text{Data 4} = 4,52066\text{E-06}$$

$$\text{Label Negatif} = \text{Data 5} + \text{Data 6} = 6,44268\text{E-27}$$

Dari ketiga label diatas, angka terbesar ada pada Label Netral, sehingga data baru dengan vektor (2,2) masuk kedalam Label Netral.

3.4.4.3. 10-Fold Cross Validation

Setelah mesin SVM dilatih, untuk menemukan tingkat untuk menemukan tingkat untuk menemukan tingkat untuk menemukan nilai f1-score paling baik dari model SVM, penulis menggunakan metode 10-Fold Cross Validation. Metode ini membagi sebuah dataset menjadi 10 bagian dengan perbandingan data 10% banding 90%. Seperti yang dijelaskan pada subbab 2.7, penggunaan metode ini mengacak posisi data training dan testing berdasarkan pembagian tersebut.

Selanjutnya, akan dilihat dari 10 pembagian data, mana pembagian yang paling baik f1-scorenya untuk diambil model dari *fold* tersebut. Berikut ditampilkan contoh dataset jika menerapkan metode 10-Fold Cross Validation

Tabel 3. 13 Contoh Dataset untuk 10-Fold Cross Validation

No	Fitur	Label	No	Fitur	Label
1	Baik	1	11	Biasa	0
2	Biasa	0	12	Buruk	-1
3	Buruk	-1	13	Sangat Buruk	-1
4	Sangat Baik	1	14	Biasa	0
5	Sangat Buruk	-1	15	Sangat Baik	1
6	Biasa	0	16	Baik	1
7	Buruk	-1	17	Biasa	0
8	Sangat Baik	1	18	Buruk	-1
9	Sangat Baik	1	19	Sangat Buruk	-1
10	Baik	1	20	Biasa	0

Misalkan, pada tabel ... memuat dataset utuh, lalu pada tabel ... memuat dataset yang telah dipecah sebagai data training dan data testing. Dengan data yang dihuruf tebal sebagai data testing, dan data sisanya sebagai data training.

Tabel 3. 14 Tabel Hasil Pembagian Data Training dan Data Testing

Fold 1			
No	Fitur	No	Fitur
1	Baik	11	Biasa
2	Biasa	12	Buruk
3	Buruk	13	Sangat Buruk
4	Sangat Baik	14	Biasa
5	Sangat Buruk	15	Sangat Baik
6	Biasa	16	Baik
7	Buruk	17	Biasa
8	Sangat Baik	18	Buruk
9	Sangat Baik	19	Sangat Buruk
10	Baik	20	Biasa

Data tersebut selanjutnya dimasukkan kedalam model sebagai data training dan di uji validasinya pada data dengan nomor 2 dan 15 sebagai data testing.

3.4.4.4. Confusion Matrix

Confusion matrix merupakan tabel dimana untuk mengevaluasi model klasifikasi, nilai f1-score sebagai nilai yang digunakan untuk menentukan model terbaik [17]. Sehingga penulis menerapkan evaluasi pada penelitian dengan metode confusion matrix. Ilustrasi penulis dalam menentukan f1-score adalah sebagai berikut:

Tabel 3. 15 Tabel Confusion Matrix

	Prediksi Positif	Prediksi Netral	Prediksi Negatif
Label Positif	50	4	0
Label Netral	20	30	2
Label Negatif	13	10	8

Selanjutnya mencari nilai precision, recall, dan f1-score didapat dengan menggunakan persamaan:

$$Precision_{(A)} = \frac{TP_A}{TP_A + FP_A} \quad (11)$$

$$Recall_{(A)} = \frac{TP_A}{TP_A + FN_A} \quad (12)$$

$$f1_{(A)} = \frac{2 * Precision_{(A)} * Recall_{(A)}}{Precision_{(A)} + Recall_{(A)}} \quad (13)$$

Dimana:

TP = True Positive

FP = False Positive

FN = False Negative

F1-score dicari pada setiap labels sehingga hasil f1-score pada setiap label akan ditampilkan sebagai berikut:

Tabel 3. 16 Hasil Perhitungan f1-score

Positif	Precision	0,926
	Recall	0,602
	F1-score	1,806
Netral	Precision	0,577
	Recall	0,682
	F1-score	2,046
Negatif	Precision	0,258
	Recall	0,8
	F1-score	2,4

Selanjutnya, untuk menemukan nilai F1-score pada model dengan label lebih dari 2, penulis menggunakan macro f1, dimana nilai yang diambil adalah

hasil rata-rata nilai f1-score semua kelas, penulis menggunakan nilai macro dikarenakan nilai akan lebih merata untuk dataset yang tidak seimbang [18]. Sehingga nilai f1-score pada confusion matrix diatas adalah 2,084. Lalu dapat dilanjutkan dengan membandingkan nilai f1 pada model dengan nilai f1 pada model yang lain.

3.5. Perancangan Sarana Pendukung

Sarana pendukung pada penelitian ini berupa kamus data yang berisikan 3 kamus dibawah ini. Kamus data ini berfungsi sebagai pendukung model dalam melakukan *pre-processing* sehingga data yang telah terfilter dapat diproses dengan lebih baik. Kamus ini bersumber dari Kamus Besar Bahasa Indonesia dan Tesaurus Tematis Bahasa Indonesia.

3.5.1. Kamus Kata Dasar

Kamus kata dasar merupakan sekumpulan kata yang bersumber dari Kamus Besar Bahasa Indonesia yang merupakan bentuk awal dari suatu kata, kamus diambil dari website [KBBI](#) dengan mengimplementasikan kamus berasal dari:

Tabel 3. 17 Tabel Sumber Kamus Kata Dasar

Library	Sumber
Sastrawi	sastrawi/kata-dasar.txt at master · sastrawi/sastrawi
Root Words	NLP bahasa resources/combined root words.txt at master · louisowen6/NLP bahasa resources

3.5.2. Kamus Sinonim Kata

Kamus sinonim merupakan kumpulan kata dengan kata lain dengan arti yang sama, kamus sinonim bersumber dari Tesaurus Tematis Bahasa Indonesia. Kamus diambil dari website [TTBI](#) dengan bentuk pengimplementasian berasal dari tabel 3.18

Tabel 3. 18 Tabel Sumber Sinonim

Library	Sumber
Tesaurus	tesaurus/dict.json at master · victoriasovereigne/tesaurus

3.5.3. Kamus Kata Tidak Baku

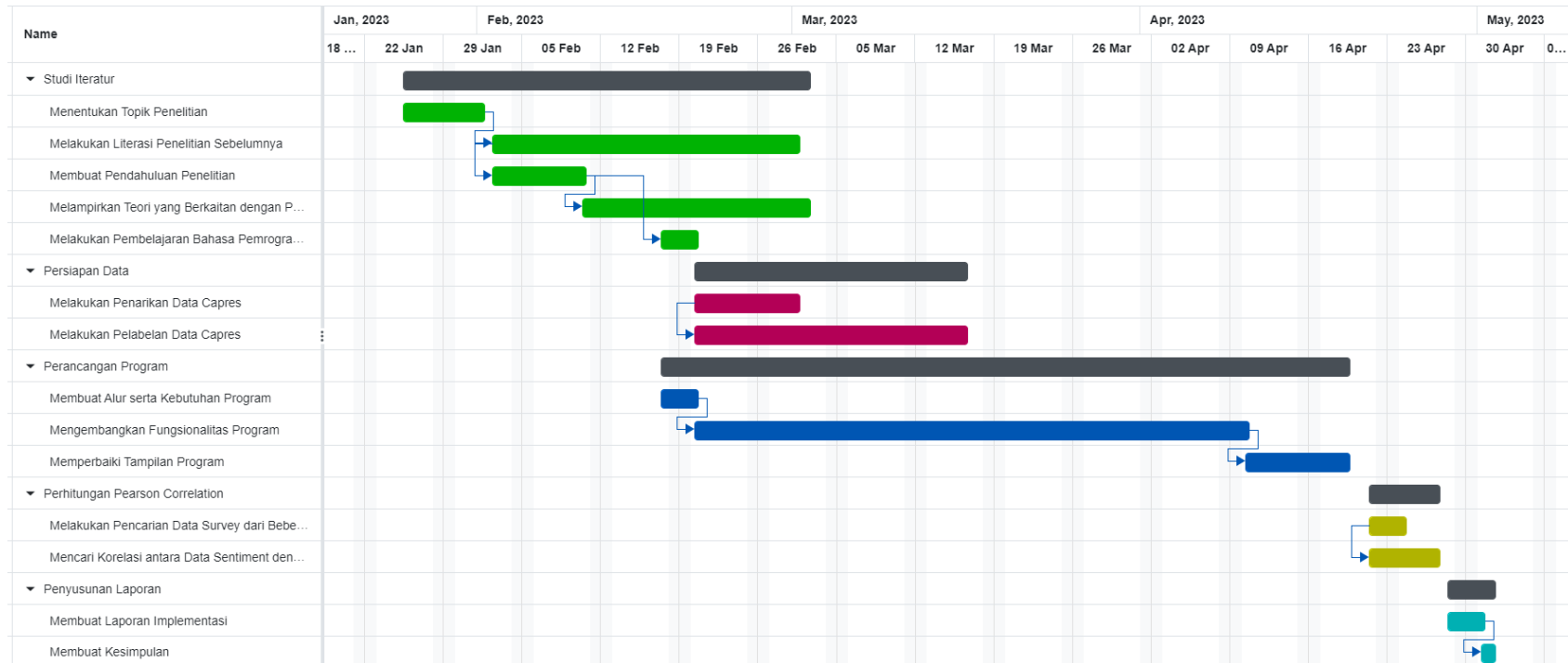
Untuk kamus data kata tidak baku merupakan sekumpulan kata yang sering digunakan namun tidak ada pada Kamus Besar Bahasa Indonesia. Kamus diambil dari repository github [louisowen6](#) sebagai berikut:

Tabel 3. 19 Tabel Sumber Kata Tidak Baku

Library	Sumber
Slang Words	NLP_bahasa_resources/combined_slang_words.txt at master · louisowen6/NLP_bahasa_resources

3.6. Jadwal Pengerjaan

Adapun jadwal pengerjaan pada penelitian ini dijabarkan dalam bentuk Gantt Chart sebagai berikut.



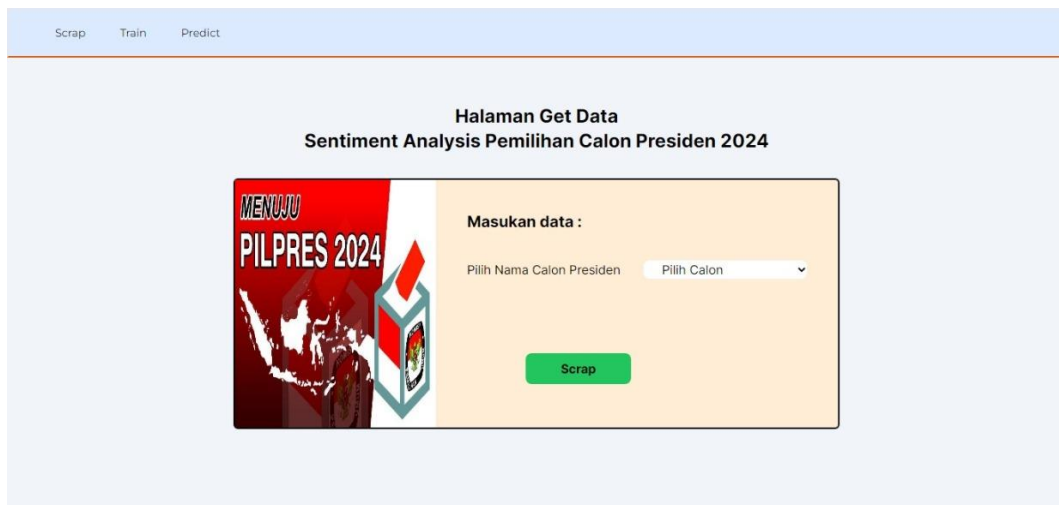
Gambar 3. 8 Gantt Chart Jadwal Pengerjaan

BAB 4

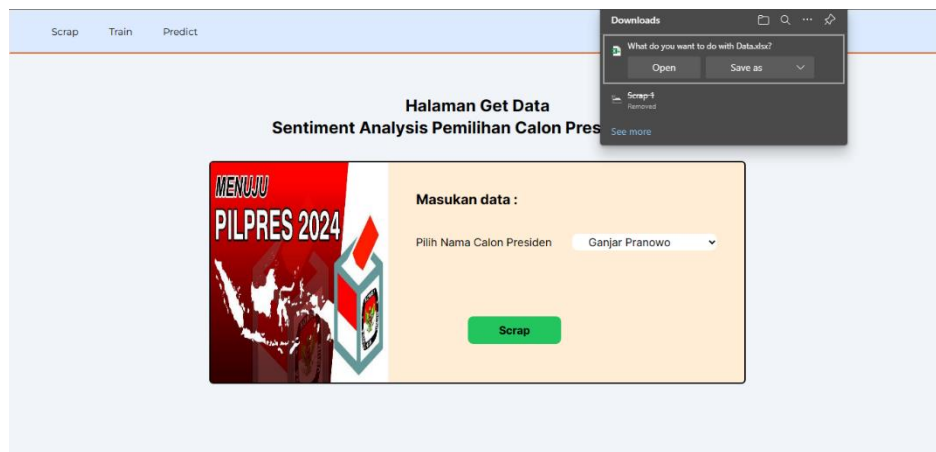
IMPLEMENTASI DAN EVALUASI

4.1. Implementasi User Interface

Gambar 4.1 sampai 4.7 menunjukkan tampilan User Interface dari website yang telah dirancang. Gambar 4.1 merupakan halaman scrap. Halaman ini berfungsi untuk melakukan scraping tweet dari twitter dengan meminta inputan nama, tanggal mulai dan tanggal akhir tarik. Proses scraping ini dilakukan dengan menggunakan library python “snsrcape”. Proses ini meminta inputan berupa query untuk dilakukan pencarian pada media sosial twitter, selanjutnya data akan dikembalikan dalam bentuk object berupa tanggal, dan tweet pengguna.



Gambar 4. 1 Gambar Tampilan Get Data

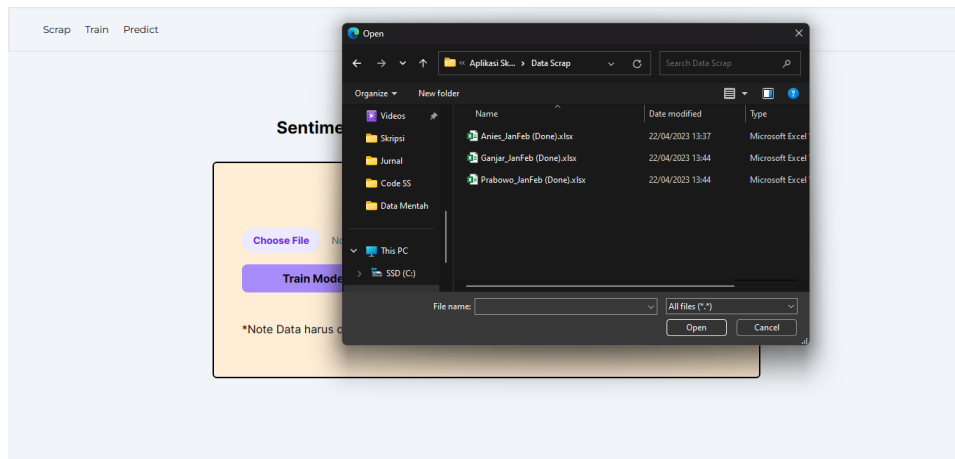


Gambar 4. 2 Gambar Tampilan Get Data Download

Selanjutnya pada halaman train, halaman ini berfungsi untuk melatih model SVM dengan inputan berupa data excel dari data scrap yang telah ditarik sebelumnya. Data scrap tentunya harus sudah diisi label sebenarnya pada kolom "Sentiment". Hasil dari halaman ini berupa tampilan grafik tentang fold terbaik beserta skornya, banyaknya prediksi dari data test tiap label, tabel confusion matrix dari fold terbaik, grafik hyperplane SVM yang sudah dilatih, wordcloud untuk kata terbanyak dari setiap label (Positif, Netral, Negatif).

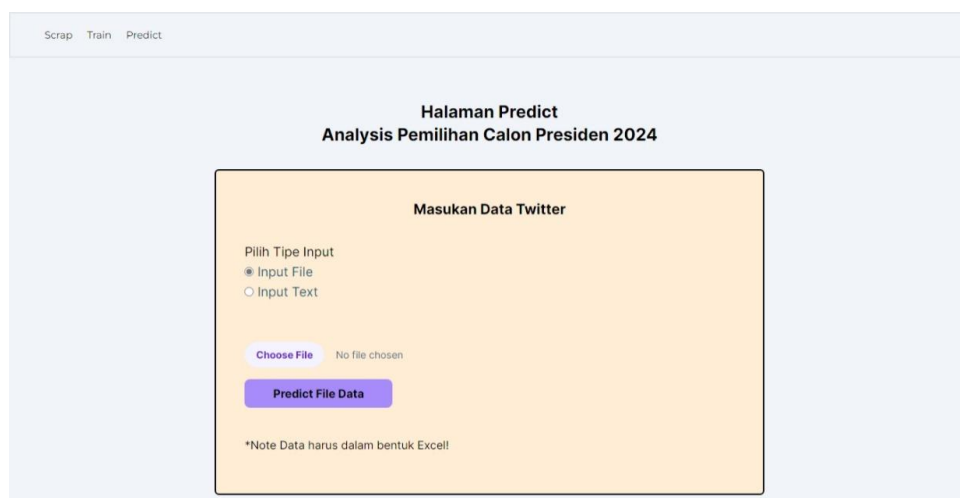


Gambar 4. 3 Gambar Tampilan Train



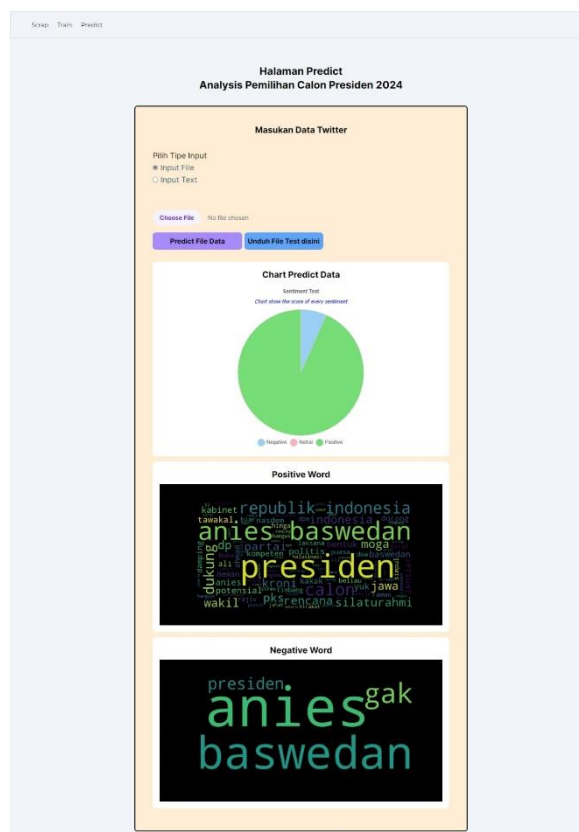
Gambar 4. 4 Gambar Tampilan Train Upload File

Halaman predict berfungsi untuk menggunakan model yang telah dilatih, pada halaman ini inputan dapat berupa 2 tipe, yang pertama dengan menggunakan ekstensi file excel dengan data tweet pada kolom bernama “Tweet”, dan tipe kedua menggunakan text area. Pada cara kedua, user hanya perlu meng *copy paste* tweet atau langsung mengetiknya pada kolom yang disediakan. Output dari halaman ini adalah berupa grafik jumlah prediksi tiap label, wordcloud setiap label, dan file hasil prediksi model pada tipe input menggunakan file excel, dari hasil preprocessing beserta prediksi untuk input text.



Gambar 4. 5 Gambar Tampilan Predict

Gambar 4. 6 Gambar Tampilan Predict Input Text

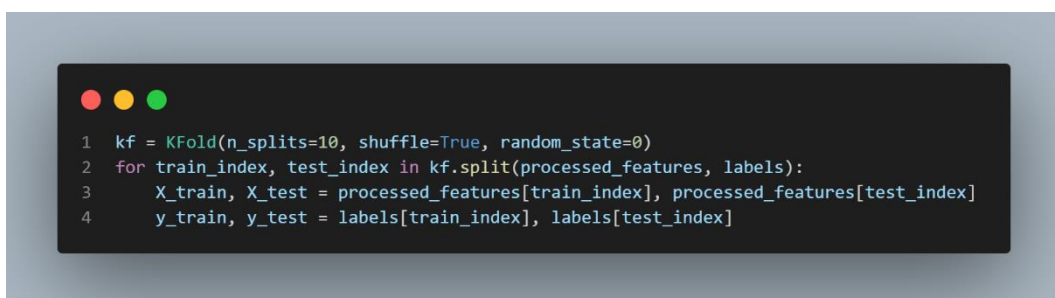


Gambar 4. 7 Gambar Tampilan Predict Input File

4.2. Implementasi Metode dan Algoritma

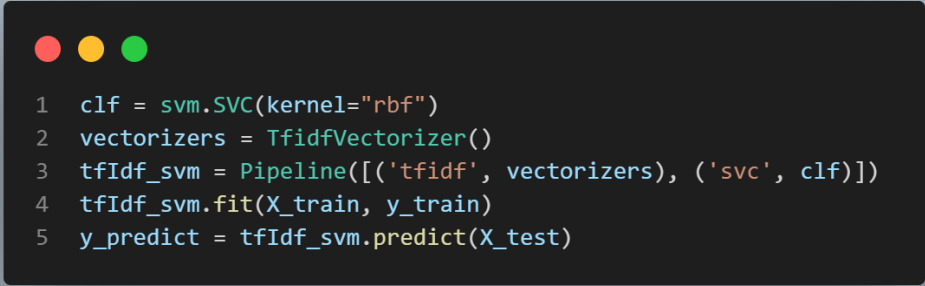
Implementasi algoritma SVM (Support Vector Machine) dilakukan persis seperti yang dilampirkan pada bab 3 subbab perancangan proses, pembangunan sistem SVM diawali dengan tahap pembersihan data, lalu transformasi data kedalam bentuk numerik, lalu dilakukan pelatihan model dengan pembagian data split 10:90 dengan metode 10-Fold Cross Validation, dan diuji model terbaik dengan mengambil nilai f1-score dengan metode Confusion Matrix.

Tahap pertama dalam melakukan train model setelah melakukan pembersihan data adalah melakukan data split yang ditunjukkan pada gambar ... Dimana data displit secara acak secara 10 fold melalui fungsi KFold dengan n_splits sebagai jumlah fold dan shuffle sebagai parameter acak.



Gambar 4. 8 Gambar Split Data dengan K-Fold Cross Validation

Lalu kita lakukan transformasi data menjadi bentuk numerik dengan menggunakan metode TF-IDF dan train kedalam model SVM dengan kernel RBF. Pada objek TfidfVectorizer. Lalu penggunaan Pipeline ditujukan agar model yang disimpan nantinya akan memuat train dari vectorizers dan juga train dari model SVM saat akan disimpan dan dipanggil pada halaman lain. Lalu dilakukan train pada Pipeline dengan parameter X_train dan y_train. Train disini dilakukan pada 2 method yaitu tf-idf dan SVM. Selanjutnya dilakukan predict pada model dan label prediksi disimpan pada variabel y_predict.



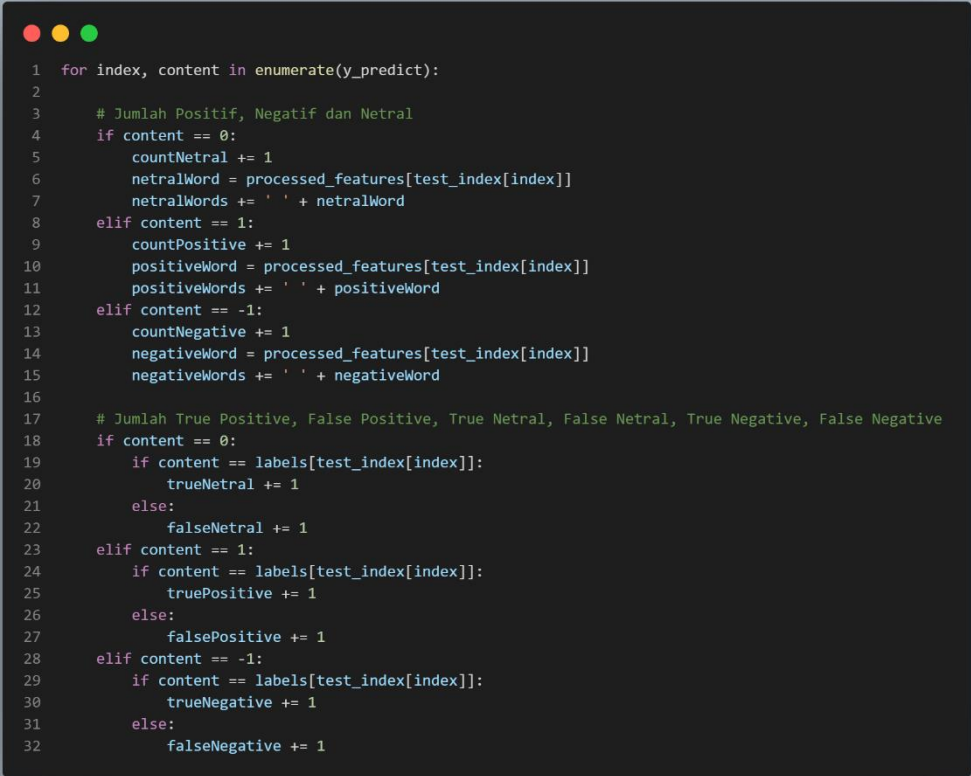
```

1  clf = svm.SVC(kernel="rbf")
2  vectorizers = TfidfVectorizer()
3  tfIdf_svm = Pipeline(['tfidf', vectorizers), ('svc', clf)])
4  tfIdf_svm.fit(X_train, y_train)
5  y_predict = tfIdf_svm.predict(X_test)

```

Gambar 4. 9 Gambar Pembobotan dan Train Mesin

Setelah itu dilakukan perhitungan jumlah masing-masing label beserta term apa yang dimuat pada label tersebut. Pengimplementasian ada pada gambar 4.10. Hal ini dilakukan untuk ditampilkan pada grafik dan wordcloud nantinya



```

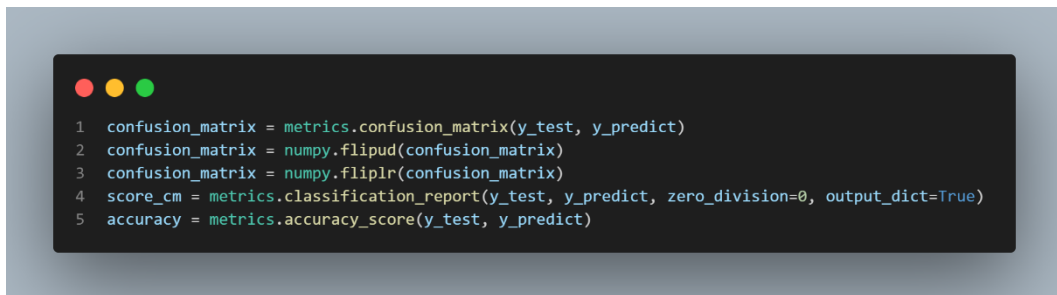
1  for index, content in enumerate(y_predict):
2
3      # Jumlah Positif, Negatif dan Netral
4      if content == 0:
5          countNetral += 1
6          netralWord = processed_features[test_index[index]]
7          netralWords += ' ' + netralWord
8      elif content == 1:
9          countPositive += 1
10         positiveWord = processed_features[test_index[index]]
11         positiveWords += ' ' + positiveWord
12     elif content == -1:
13         countNegative += 1
14         negativeWord = processed_features[test_index[index]]
15         negativeWords += ' ' + negativeWord
16
17     # Jumlah True Positive, False Positive, True Netral, False Netral, True Negative, False Negative
18     if content == 0:
19         if content == labels[test_index[index]]:
20             trueNetral += 1
21         else:
22             falseNetral += 1
23     elif content == 1:
24         if content == labels[test_index[index]]:
25             truePositive += 1
26         else:
27             falsePositive += 1
28     elif content == -1:
29         if content == labels[test_index[index]]:
30             trueNegative += 1
31         else:
32             falseNegative += 1

```

Gambar 4. 10 Gambar Tampilan Persiapan Data

Selanjutnya, setelah semua proses train SVM dijalankan, maka fungsi ini akan menghasilkan confusion matrix berdasarkan `y_predict` dan `y_test` (label

sebenarnya), dan score pada model. Score ini dihitung berdasarkan jumlah label yang diprediksi dengan label test.



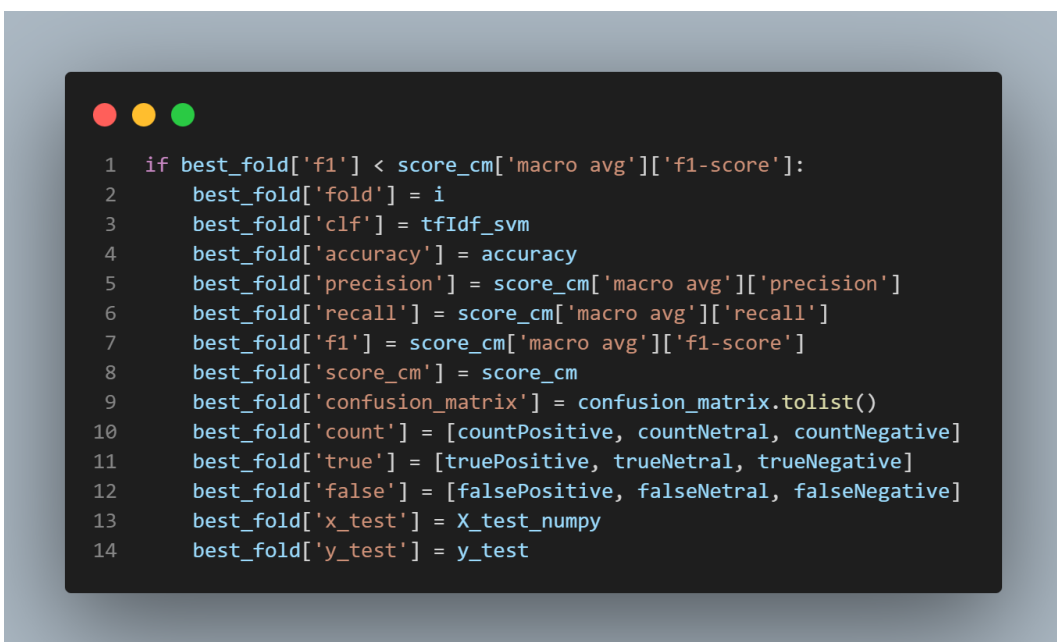
```

1 confusion_matrix = metrics.confusion_matrix(y_test, y_predict)
2 confusion_matrix = numpy.flipud(confusion_matrix)
3 confusion_matrix = numpy.fliplr(confusion_matrix)
4 score_cm = metrics.classification_report(y_test, y_predict, zero_division=0, output_dict=True)
5 accuracy = metrics.accuracy_score(y_test, y_predict)

```

Gambar 4. 11 Gambar Implementasi Confusion Matrix

Lalu data score akan dibandingkan fold terbaiknya berdasarkan nilai f1 dan disave dalam 1 variable bernama best_fold yang akan memuat semua data yang ada dalam sebuah fold tersebut, data tersebut antara lain f1-score, accuracy, model, precision, recall, confusion matrix, jumlah prediksi per label, jumlah prediksi salah dan benar, x_test dan y_test. Fold terbaik ditentukan berdasarkan f1-score saat data False Negative (FN) dan False Positive (FP) sangat tidak mendekati [19].



```

1 if best_fold['f1'] < score_cm['macro avg']['f1-score']:
2     best_fold['fold'] = i
3     best_fold['clf'] = tfIdf_svm
4     best_fold['accuracy'] = accuracy
5     best_fold['precision'] = score_cm['macro avg']['precision']
6     best_fold['recall'] = score_cm['macro avg']['recall']
7     best_fold['f1'] = score_cm['macro avg']['f1-score']
8     best_fold['score_cm'] = score_cm
9     best_fold['confusion_matrix'] = confusion_matrix.tolist()
10    best_fold['count'] = [countPositive, countNetral, countNegative]
11    best_fold['true'] = [truePositive, trueNetral, trueNegative]
12    best_fold['false'] = [falsePositive, falseNetral, falseNegative]
13    best_fold['x_test'] = X_test_numpy
14    best_fold['y_test'] = y_test

```

Gambar 4. 12 Gambar Pencarian Fold Terbaik berdasarkan f1-score

Data `y_test` dan `y_predict` diolah menjadi confusion matrix pada gambar 4.13 dengan fungsi `ConfusionMatrixDisplay`. Juga hyperplane SVM dibentuk menggunakan `DecisionBoundaryDisplay` dengan data sumbu X yaitu fitur pada data test dan sumbu Y merupakan label prediksi pada data test.



```

1  matplotlib.use('agg')
2  result = numpy.column_stack((best_fold['x_test'].data, best_fold['y_test']))
3  model = clf.fit(result[:, :2], best_fold['y_test'])
4  display = DecisionBoundaryDisplay.from_estimator(model, result[:, :2], response_method="predict", alpha=0.5)
5  display.plot(plot_method="contourf", xlabel="Test Features", ylabel="Predicted Labels")
6  display.ax_.scatter(result[best_fold['y_test'] == -1, 0], result[best_fold['y_test'] == -1, 1], edgecolors="black", marker='x')
7  display.ax_.scatter(result[best_fold['y_test'] == 0, 0], result[best_fold['y_test'] == 0, 1], edgecolors="black", marker='o')
8  display.ax_.scatter(result[best_fold['y_test'] == 1, 0], result[best_fold['y_test'] == 1, 1], edgecolors="black", marker='+')
9  plt.savefig('TrainData/TrainChart.png')
10
11 cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix=confusion_matrix, display_labels=["Positive", "Netral", "Negative"])
12 cm_display.plot()
13 plt.savefig('TrainData/TrainPlot.png')
14 matplotlib.pyplot.close()

```

Gambar 4. 13 Gambar Plot Confusion Matrix dan Hyperplane SVM

Langkah selanjutnya adalah menyimpan data pada gambar 4.10 menjadi Wordcloud yang diimplementasikan melalui fungsi `generateWordCloud` pada gambar 4.14 Wordcloud dibatasi dengan maksimal 100 kata dengan tinggi 400px dan lebar 800px. Lalu semua akan di plot menggunakan interpolasi nearest agar huruf pada setiap kata tidak ada yang diperhalus dengan blur.



```

1  def generateWordCloud(text, name, mode):
2      list_stopwords = stopwords.words('indonesian')
3      new_stopwords = open('../Function/lib/NLP_bahasa_resources/combined_stop_words.txt').read().split("\n")
4      list_stopwords.extend(new_stopwords)
5
6      wordcloud = WordCloud(max_words=100, height=400, width=800, background_color="black").generate(text)
7      matplotlib.use('agg')
8      plt.figure(figsize=(20,10), facecolor='k')
9      plt.imshow(wordcloud, interpolation="nearest")
10     plt.axis("off")
11     plt.savefig(f'{mode}Data/{name + mode}WordCloud.png')
12     matplotlib.pyplot.close()

```

Gambar 4. 14 Gambar Implementasi WordCloud

Data terakhir yang harus disiapkan pada step train ini adalah tabel DF (Data Frequency) yang diimplementasikan pada gambar 4.15 lalu `tfidf_svm.named_steps['tfidf'].get_feature_names_out()` mengambil element pada Pipeline yang sudah di train untuk dapat divisualisasikan setiap termnya dengan menggunakan fungsi `get_features_names_out`.




```

1 df_value = []
2 words_list = tfidf_svm.named_steps['tfidf'].get_feature_names_out()
3 for index, word in enumerate(words_list):
4     df = numpy.sum(all_tfidf[:, index] > 0)
5     df_value.append({'word': word, 'df': df})

```

Gambar 4. 15 Gambar Implementasi Tabel DF

Lalu model yang sudah ditrain dan ditentukan yang terbaik oleh f1-score akan disimpan menggunakan library `joblib`.



```

1 if not os.path.exists('Model'):
2     os.mkdir('Model')
3
4 joblib.dump(best_fold['clf'], 'Model/svm.pkl')
5

```

Gambar 4. 16 Gambar Save Model

4.3. Pengujian Sistem

Subbab ini menjelaskan bagaimana hasil proses training dengan mencari model dengan fold terbaik, dilanjutkan dengan analisa keterkaitan hasil prediksi model dengan menggunakan Pearson Moment Product Correlation.

4.3.1. Hasil Train Model

Data train diambil dari dataset training sebanyak 15 data calon Anies Baswedan dari dataset sebanyak 283 data test. Data hasil training ditampilkan pada tabel 4.1.

Tabel 4. 1 Tabel Sample Data Test dari Dataset Anies Baswedan

No	Tweet	Predict	Actual
1	Kita Butuh Pemimpin Yang Benar-benar Cakap, Paham Masalah, Dan Berpihak Pada Rakyat. 2024 Anies Baswedan Presiden RI ID	1	1
2	ANIES BASWEDAN PRESIDEN RI 2024	1	1
3	Survey menyatakan warga Jakarta puas kinerjanya Anies Baswedan. Atas dasar itulah Rakyat menilai Anies pantas jadi Presiden PD 2024.	1	1
4	Dukung Anies Baswedan untuk Presiden RI ke VIII di pemilu 2024	1	1
5	RELAWAN DPD Kabupaten Bangkalan mengajak seluruh komponen bangsa untuk mengukung Bapak Anies Baswedan sebagai calon presiden tahun 2024,” Dr H Engkun Iskandar membacakan pernyataan deklarasi.	1	1
6	Semoga Selalu sehat dan panjang umur, supaya 2024	1	1

	nanti bisa menghadiri acara pelantikan Anies Baswedan sebagai presiden RI. Aamiin		
7	Anies Baswedan Presiden!!!! MERDEKA!! 🇮🇩 🇮🇩	1	1
8	Mengerikan kalau Anies Baswedan jadi Presiden. Maka HTI dan FPI akan bangkit lagi dan meminta rehabilitasi. Buntutnya, mereka eksis kembali gelar demo ala 212 berjilid-jilid.	-1	-1
9	Nauzubillah gobloknya ... Pilpres 2024, InsyaAllah Anies Baswedan Presiden RI ID	1	1
10	Dr H Engkan Iskandar Pimpin Deklarasi Relawan Dukung Anies Baswedan Presiden	1	1
11	Inilah mantan Goodbener DKI pak Anies Baswedan, Calon Presiden RI yg sangat toleran thd umat beragama non muslim, dan yg terpenting tdk hobi nonton bokep sprti capres idolanya si densi alias 🤪	1	1
12	Tetap akan kalah sama Anies Baswedan insyaallah jadi presiden RI	1	1
13	In sha Allah Pak Anies Baswedan, akan jadi Presiden, semangat Pak Anies... Harkat martabat mu tak akan jatuh, hanya karna di caci di maki, di hina oleh siapa pun 😊	1	1
14	Ketika masyarakat yakin bahwa Anies adalah satu2 calon presiden yang bisa mensejahterakan rakyat Indonesia maka keyakinan itu	1	1

	dibuktikan bersama. Pilih Anies Baswedan!		
15	Anies Baswedan presiden RI 2024	1	1

Hasil Train tersebut diambil berdasarkan model terbaik dengan kernel RBF pada fold ke 8 dengan hasil sebagai berikut:

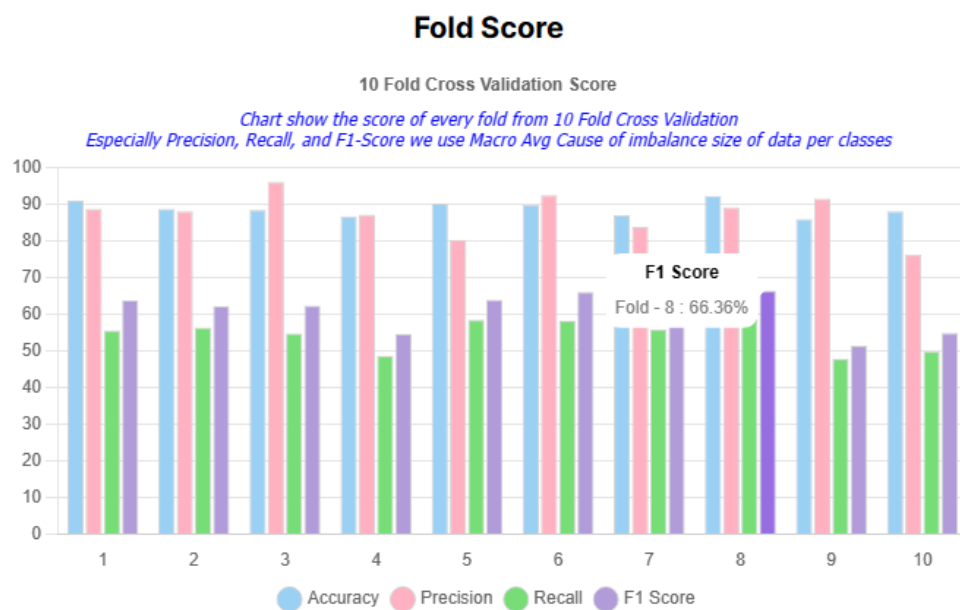
```

Urutan ke 1 Fold ke 8 f1-score 0.6635658914728683
Urutan ke 2 Fold ke 6 f1-score 0.6598707228844215
Urutan ke 3 Fold ke 5 f1-score 0.6397727984256899
Urutan ke 4 Fold ke 1 f1-score 0.6371151371151371
Urutan ke 5 Fold ke 3 f1-score 0.6230948803883046
Urutan ke 6 Fold ke 2 f1-score 0.6212196540963664
Urutan ke 7 Fold ke 7 f1-score 0.6184236598579228
Urutan ke 8 Fold ke 10 f1-score 0.5486580447781949
Urutan ke 9 Fold ke 4 f1-score 0.5458523672312413
Urutan ke 10 Fold ke 9 f1-score 0.5139831658973663

```

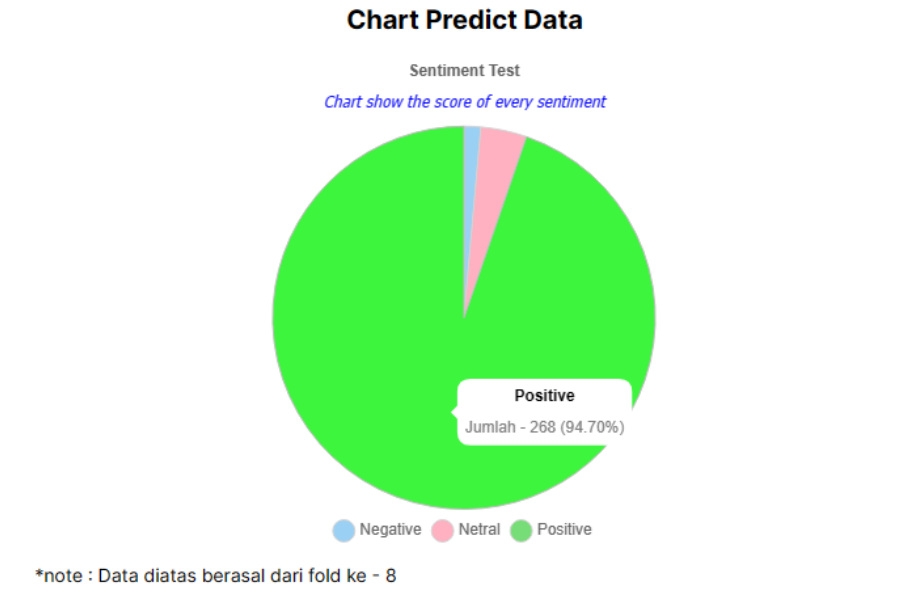
Gambar 4. 17 Gambar f1-score Anies Baswedan

Score tiap fold juga ditampilkan dalam bentuk grafik seperti pada gambar 4.18 dengan f1-score tertinggi pada angka 66,36%



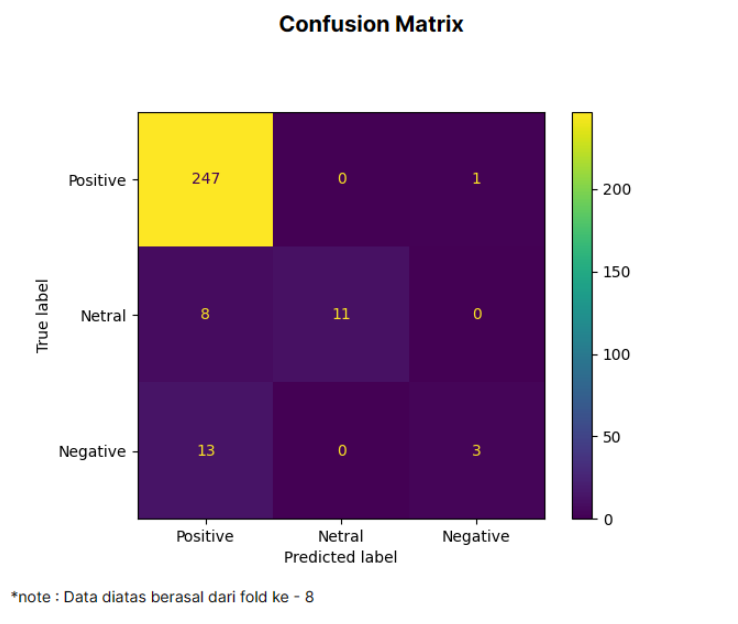
Gambar 4. 18 Grafik Score setiap Fold Anies Baswedan

Lalu dari hasil train tersebut, didapat bahwa prediksi sebesar 268 label positif, 11 label netral, dan 4 label negatif.



Gambar 4. 19 Grafik Pie Jumlah Tiap Sentiment Anies Baswedan

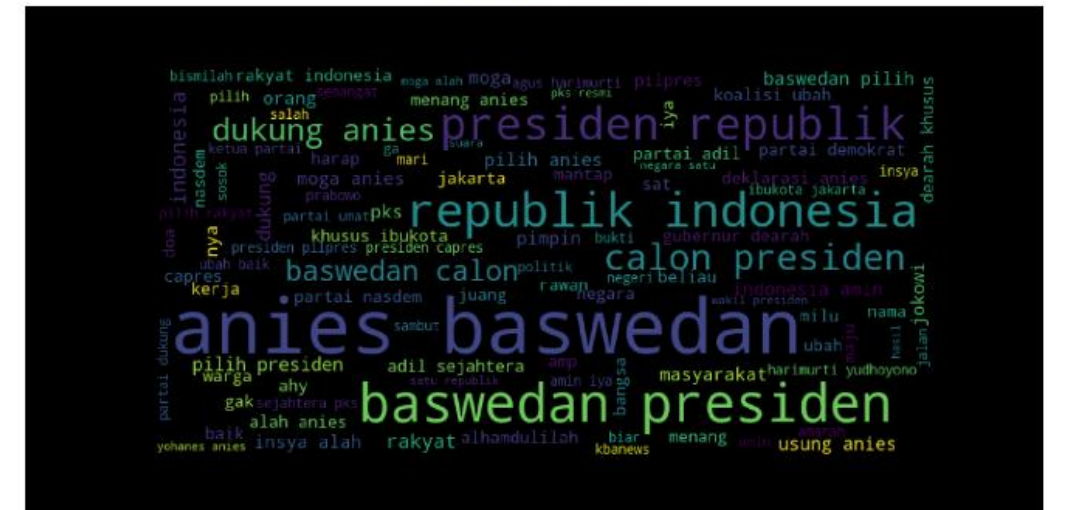
Lalu confusion matrix dari proses train ini menunjukkan tabel dengan nilai True Positive 247, True Netral 11, True Negative 3, False Positive 21, False Netral 0, dan False Negative 1.



Gambar 4. 20 Confusion Matrix Anies Baswedan

Juga didapat wordcloud untuk kata positif, netral, dan negatif sebagai berikut:

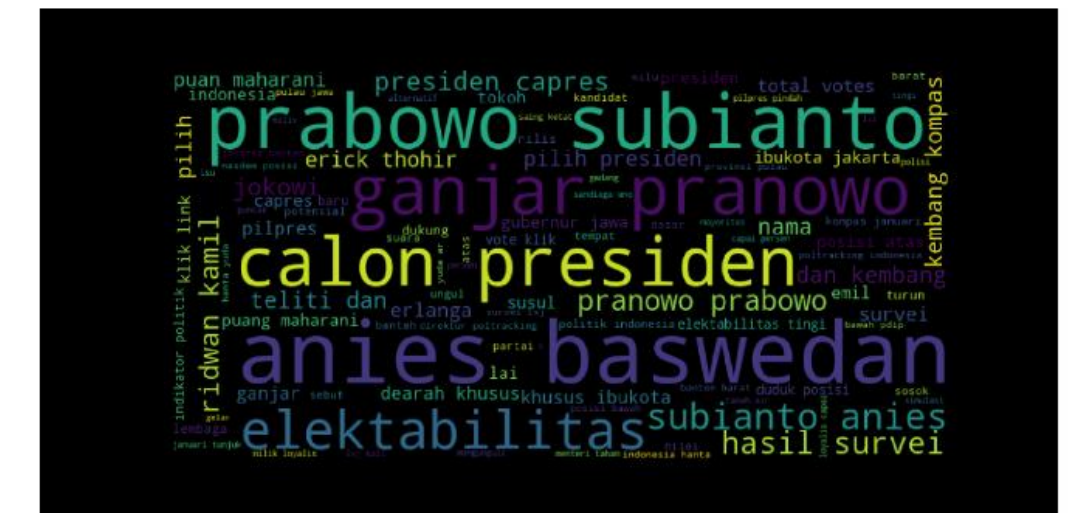
Positive Word



*note : Data diatas berasal dari fold ke - 8

Gambar 4. 21 Gambar WordCloud Positif Anies Baswedan

Netral Word



*note : Data diatas berasal dari fold ke - 8

Gambar 4. 22 Gambar WordCloud Netral Anies Baswedan

	Ini lebih masuk akal dan berjenjang untuk kemajuan dan kaderisasi Pemimpin RI berikutnya.		
3	sukses selalu untuk Pak Ganjar Pranowo semoga bisa menjadi Presiden RI Ke-8 amiin	1	1
4	Emang pantas jadi presiden. Luar biasa nih pak Ganjar Pranowo	1	1
5	Gubernur Ganjar Pranowo melantik Hevearita G Rahayu sebagai Wali Kota Semarang, sisa masa jabatan 2021-2026, di Grhadhika Bhakti Praja. Pelantikan dihadiri Presiden RI kelima, Megawati Soekarnoputri. Greek Cencelo Indonesia Terus Maju	1	1
6	Gubernur Jawa Tengah Ganjar Pranowo melantik Hevearita G Rahayu sebagai Wali Kota Semarang, sisa masa jabatan 2021-2026, di Grhadhika Bhakti Praja. Pelantikan dihadiri Presiden RI kelima, Megawati Soekarnoputri. Greek Cencelo Indonesia Terus Maju	1	1
7	Komunitas Sopir Truk DKI Jakarta Dukung Ganjar Pranowo Menjadi Presiden RI 2024 & 2029	1	1
8	Gelombang Besar Dukungan Ganjar Pranowo di Kota Pekalongan, Jawa Tengah, Sinyal Kuat Presiden 2024.	1	1
9	Satu lagi, komunitas yang dukung Ganjar Pranowo Komunitas Sopir Truk DKI Jakarta Perkuat konsolidasi dalam menjaring dukungan bagi	1	1

	Ganjar Pranowo sebagai Presiden RI 2024 – 2029		
10	Ketua Umum Kadin, ditengarai tengah berambisi menjadi menteri jika nanti Ganjar Pranowo terpilih menggantikan Joko Widodo sebagai presiden	1	1
11	Pak Ganjar Pranowo sosok pemimpin hebat yg pantas untuk menjadi Presiden RI	1	1
12	Alhamdulillah semoga pak Ganjar Pranowo menjadi presiden Indonesia	1	1
13	Beredar Kabar PDIP Resmi Usung Ganjar Pranowo dan Khofifah di Pilpres, Ini Faktanya: Beredar kabar yang menyebutkan PDIP resmi mengusung Ganjar Pranowo sebagai calon presiden atau Capres dan Khofifah Indar...	1	1
14	selalu mendukung Pak Ganjar Pranowo untuk menjadi Presiden RI	1	1
15	semoga pak Ganjar Pranowo menjadi presiden Indonesia	1	1

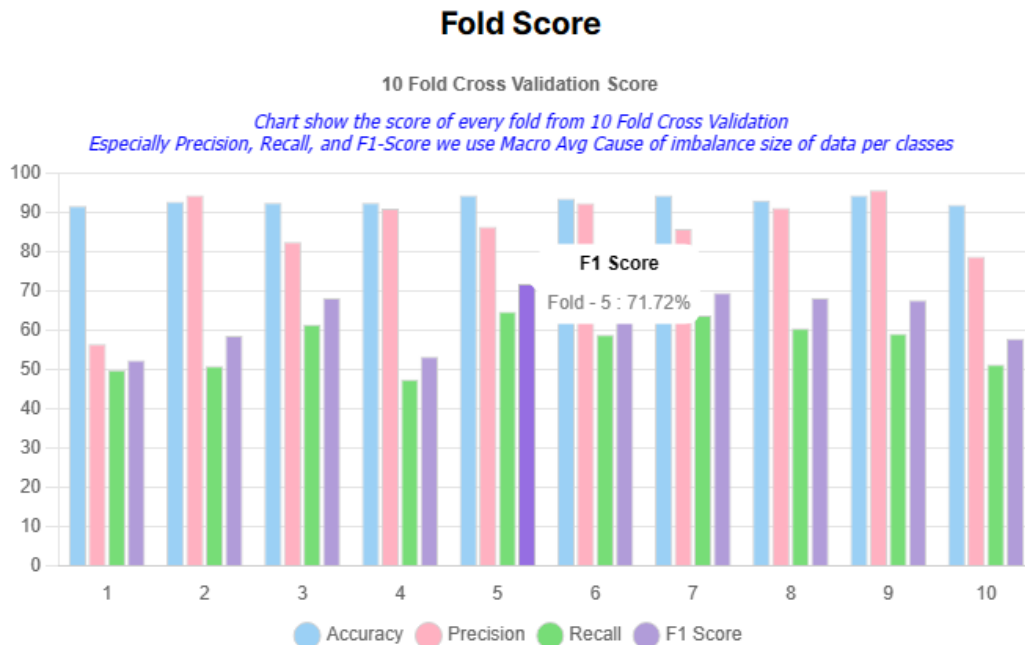
Hasil Train tersebut diambil berdasarkan model terbaik dengan kernel

RBF pada fold ke 5 dengan hasil sebagai berikut:

```
Urutan ke 1 Fold ke 5 f1-score 0.7171783787406588
Urutan ke 2 Fold ke 7 f1-score 0.6951154052603328
Urutan ke 3 Fold ke 6 f1-score 0.6824218865035192
Urutan ke 4 Fold ke 8 f1-score 0.6817319877422179
Urutan ke 5 Fold ke 3 f1-score 0.6815987933634992
Urutan ke 6 Fold ke 9 f1-score 0.6767451102356762
Urutan ke 7 Fold ke 2 f1-score 0.5856682084314672
Urutan ke 8 Fold ke 10 f1-score 0.5785851336871745
Urutan ke 9 Fold ke 4 f1-score 0.5327829391229968
```

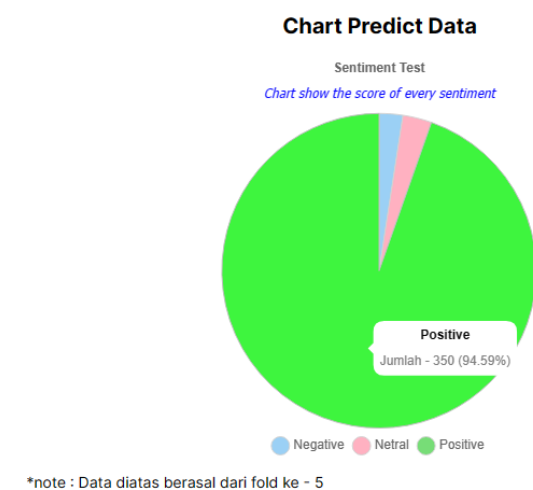
Gambar 4. 24 Gambar f1-score Ganjar Pranowo

Score tiap fold juga ditampilkan dalam bentuk grafik seperti pada gambar 4.25 dengan f1-score tertinggi pada angka 71,72%



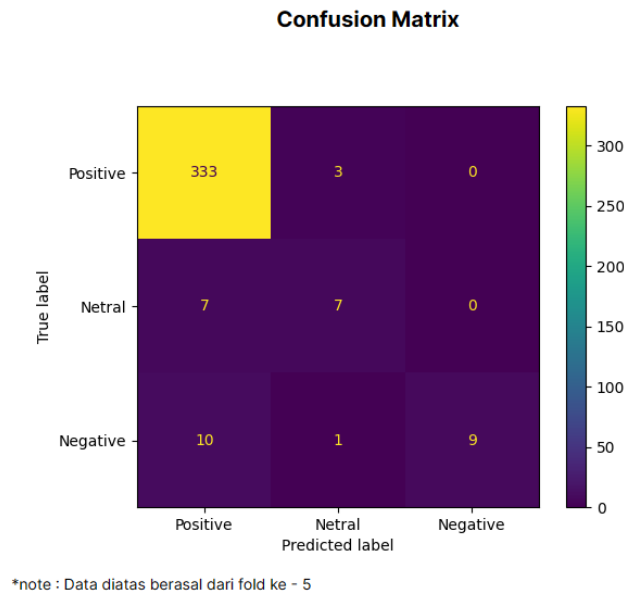
Gambar 4. 25 Grafik Score setiap Fold Ganjar Pranowo

Lalu dari hasil train tersebut, didapat bahwa prediksi sebesar 350 label positif, 11 label netral, dan 9 label negatif.



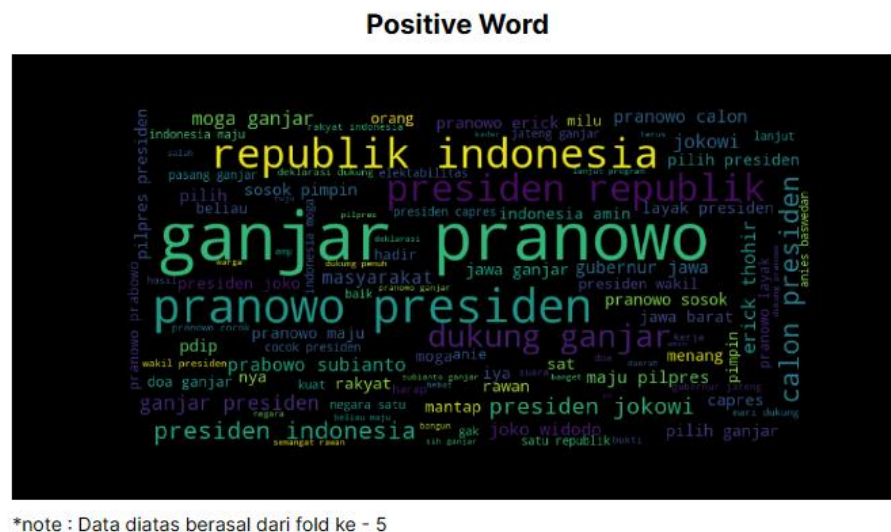
Gambar 4. 26 Grafik Pie Jumlah Tiap Sentiment Ganjar Pranowo

Lalu confusion matrix dari proses train ini menunjukkan tabel dengan nilai True Positive 333, True Netral 7, True Negative 9, False Positive 17, False Netral 4, dan False Negative 0.



Gambar 4. 27 Confusion Matrix Ganjar Pranowo

Juga didapat wordcloud untuk kata positif, netral, dan negatif sebagai berikut:



Gambar 4. 28 Gambar WordCloud Positif Ganjar Pranowo

Tabel 4. 3 Tabel Sample Data Test dari Dataset Prabowo Subianto

No	Tweet	Predict	Actual
1	Prabowo Subianto Ngarep Omongan Gus Dur jadi Kenyataan, jadi Presiden di Usia Tua - RMOL	-1	-1
2	Prabowo Subianto: Prabowo Subianto Ngarep Omongan Gus Dur jadi Kenyataan, jadi Presiden di Usia Tua - RMOL	-1	-1
3	Tetap Prabowo Subianto presiden Republik Indonesia 2024	1	1
4	Sandiaga Uno mengungkapkan ada perjanjian atau kesepakatan antara Ketua Umum Partai Gerindra Prabowo Subianto dengan Anies Baswedan terkait pemilihan presiden (pilpres). capres2024 pemilu2024 dekade08 mendingprabowo	0	0
5	emak-emak berharap bapak prabowo menjadi presiden,” tutupnya Prabowo Subianto	1	1
6	Ketua PD Tidar Sumut ini menyampaikan bahwa hingga saat ini di setiap kunjungannya ke masyarakat dalam menjalankan tugas kedewanan emak-emak menyampaikan harapan dan aspirasi dukungan ke Menhan Prabowo Subianto agar bisa melanjutkan kepemimpinan Presiden Joko Widodo.	1	1

7	Emak-Emak Masih Idolakan Prabowo Subianto Jadi Presiden	1	1
8	emak-emak menyampaikan harapan dan aspirasi dukungan ke Menhan Prabowo Subianto agar bisa melanjutkan kepemimpinan Presiden Joko Widodo (Jokowi). Prabowo Subianto	1	1
9	Presiden Jokowi Restu Prabowo Subianto Maju Pilpres 2024, Netizen: Prabowo-Ganjar Sangat Cocok - Metro Sulteng - Metro Sulteng	1	1
10	emak-emak menyampaikan harapan dan aspirasi dukungan ke Menhan Prabowo Subianto agar bisa melanjutkan kepemimpinan Presiden Joko Widodo (Jokowi). Prabowo Subianto	1	1
11	"Emak-emak di Kota Medan tetap setia dan semakin sayang kepada pak prabowo, emak-emak berharap Bapak Prabowo menjadi presiden," tutupnya. Prabowo Subianto	1	1
12	Sementara warganet lainnya menyampaikan kecupan itu bukti emak-emak di Medan mendukung Prabowo untuk maju sebagai Presiden 2024. " Prabowo Subianto	1	1
13	Emak-emak di Kota Medan Dukung Prabowo Subianto Presiden 2024 - Tribun Timur	1	1
14	Emak-Emak Medan Doakan Prabowo Subianto saat Hadiri Acara Zikir: Sehat Terus Pak, Presiden-Presiden! - Sumut	1	1
15	Berita Prabowo : Partai Gerindra Buka Pintu Lebar	1	1

	Buat Putra Bungsu Presiden Jokowi, Kaesang Pangarep: Prabowo Subianto Percaya, Generasi Muda Membawa Perubahan dan Kemakmuran -		
--	---	--	--

Hasil Train tersebut diambil berdasarkan model terbaik dengan kernel RBF pada fold ke 4 dengan hasil sebagai berikut:

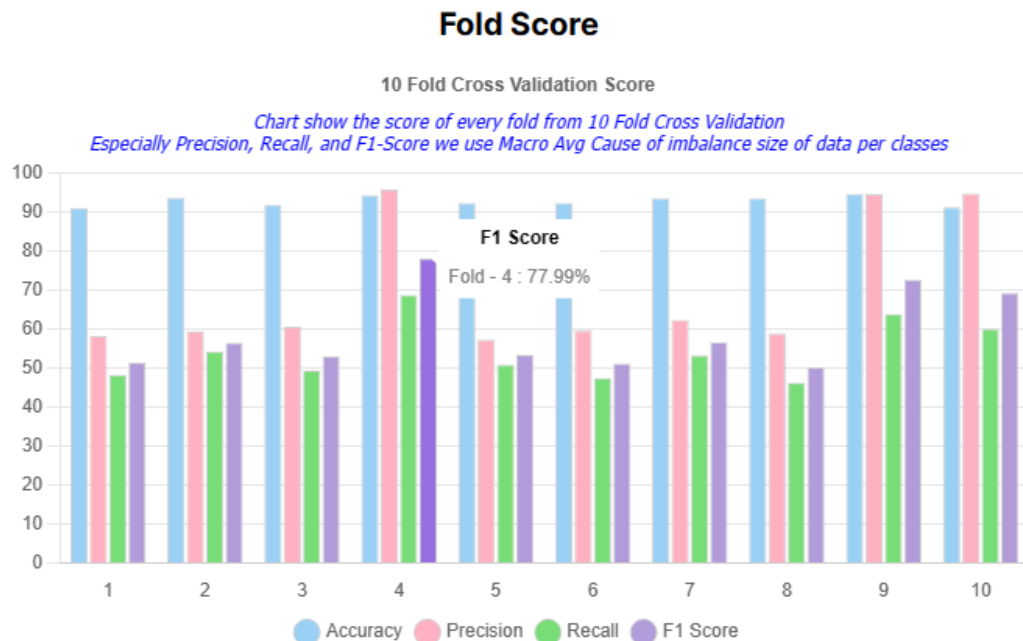
```

Urutan ke 1 Fold ke 4 f1-score 0.7798509958134371
Urutan ke 2 Fold ke 9 f1-score 0.7259683578832515
Urutan ke 3 Fold ke 10 f1-score 0.6927331248016682
Urutan ke 4 Fold ke 7 f1-score 0.5659992550391715
Urutan ke 5 Fold ke 2 f1-score 0.564375764819716
Urutan ke 6 Fold ke 5 f1-score 0.533059533059533
Urutan ke 7 Fold ke 3 f1-score 0.5293011889656469
Urutan ke 8 Fold ke 1 f1-score 0.5140291806958474
Urutan ke 9 Fold ke 6 f1-score 0.5110882193035168
Urutan ke 10 Fold ke 8 f1-score 0.5012624915537537

```

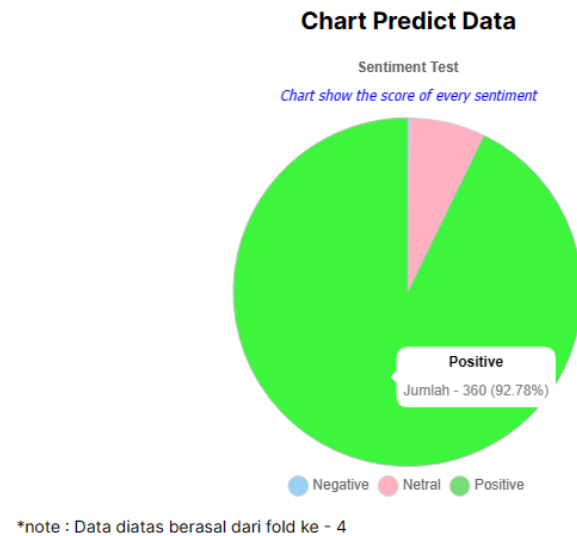
Gambar 4. 31 Gambar f1-score Prabowo Subianto

Score tiap fold juga ditampilkan dalam bentuk grafik seperti pada gambar 4.32 dengan f1-score tertinggi pada angka 71,72%



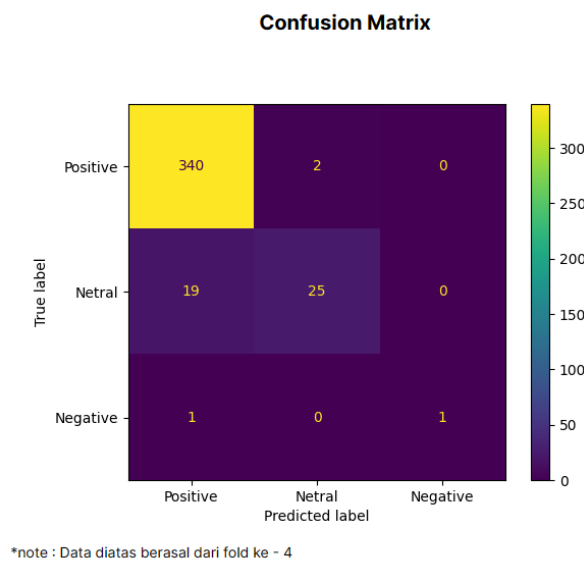
Gambar 4. 32 Grafik Score setiap Fold Prabowo Subianto

Lalu dari hasil train tersebut, didapat bahwa prediksi sebesar 360 label positif, 27 label netral, dan 1 label negatif.



Gambar 4. 33 Grafik Pie Jumlah Tiap Sentiment Prabowo Subianto

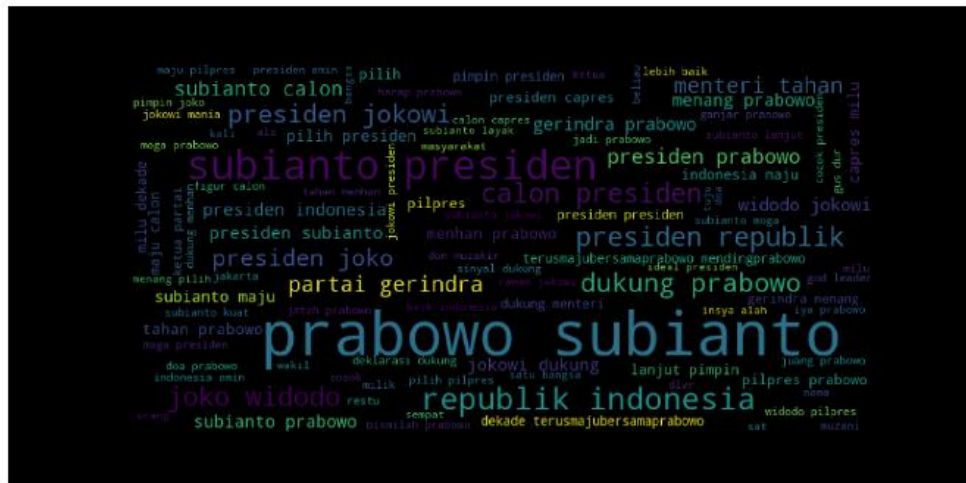
Lalu confusion matrix dari proses train ini menunjukkan tabel dengan nilai True Positive 340, True Netral 25, True Negative 1, False Positive 20, False Netral 2, dan False Negative 0.



Gambar 4. 34 Confusion Matrix Prabowo Subianto

Juga didapat wordcloud untuk kata positif, netral, dan negatif sebagai berikut:

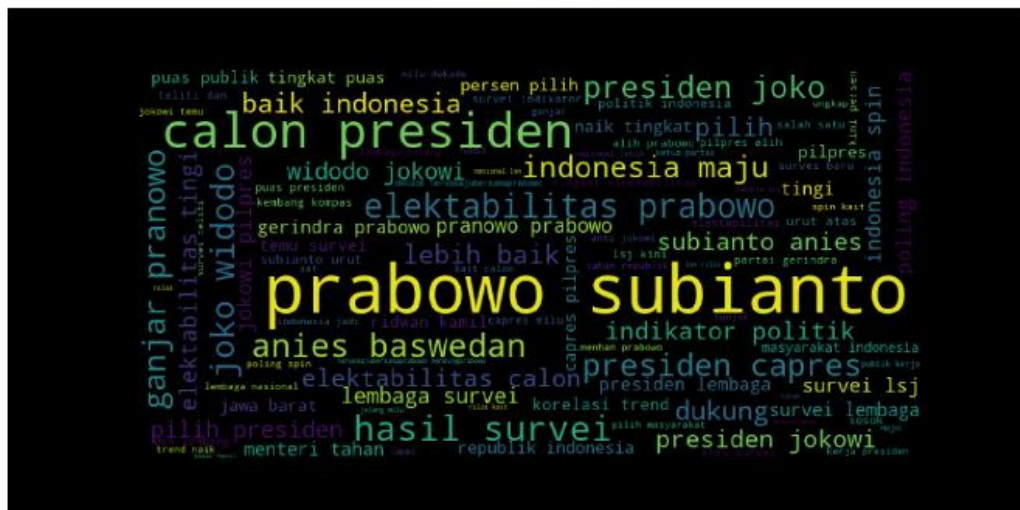
Positive Word



*note : Data diatas berasal dari fold ke - 4

Gambar 4. 35 Gambar WordCloud Positif Prabowo Subianto

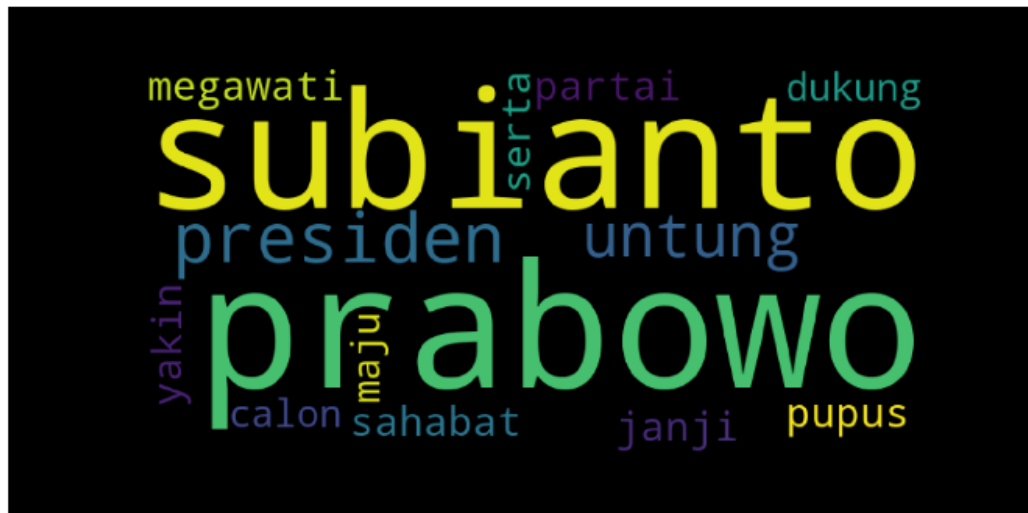
Netral Word



*note : Data diatas berasal dari fold ke - 4

Gambar 4. 36 Gambar WordCloud Netral Prabowo Subianto

Negative Word

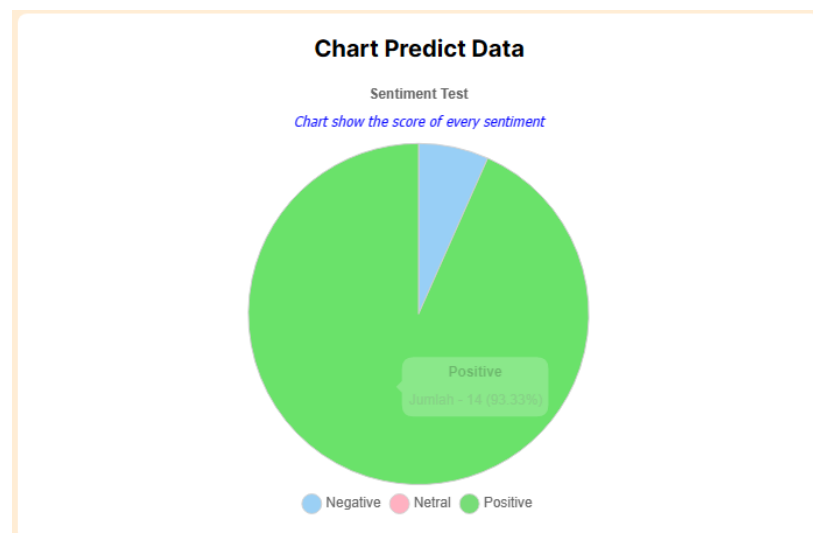


*note : Data diatas berasal dari fold ke - 4

Gambar 4. 37 Gambar WordCloud Negatif Prabowo Subianto

4.3.2. Hasil Test Model

Data Test diambil dari twitter berupa 15 baris data acak dengan keyword “Anies Baswedan Presiden” lalu dimasukan kedalam model yang sudah di train



Gambar 4. 38 Grafik Pie Sentiment Test Twitter Anies Baswedan

dan hasil prediksi mesin ditunjukkan pada gambar 4.38 dan detailnya pada tabel

4.4

Tabel 4. 4 Tabel Datatest Twitter Anies Baswedan

No	Sentiment	Tweet
1	1	Kita Butuh Pemimpin Yang Benar-benar Cakap, Paham Masalah, Dan Berpihak Pada Rakyat. 2024 Anies Baswedan Presiden RI ID
2	1	ANIES BASWEDAN PRESIDEN RI 2024
3	1	Survey menyatakan warga Jakarta puas kinerjanya Anies Baswedan. Atas dasar itulah Rakyat menilai Anies pantas jadi Presiden PD 2924.
4	1	Dukung Anies Baswedan untuk Presiden RI ke VIII di pemilu 2024
5	1	RELAWAN DPD Kabupaten Bangkalan mengajak seluruh komponen bangsa untuk mengusung Bapak Anies Baswedan sebagai calon presiden tahun 2024,” Dr H Engkun Iskandar membacakan pernyataan deklarasi.
6	-1	Semoga Selalu sehat dan panjang umur, supaya 2024 nanti bisa menghadiri acara pelantikan Anies Baswedan sebagai presiden RI. Aamiin
7	1	Anies Baswedan Presiden!!!! MERDEKA!! 🇮🇩 🇮🇩
8	1	Mengerikan kalau Anies Baswedan jadi Presiden. Maka HTI dan FPI akan bangkit lagi dan meminta rehabilitasi. Buntutnya, mereka eksis kembali gelar demo ala 212 berjilid-jilid.
9	1	Nauzubillah gobloknya ... Pilpres 2024, InsyaAllah Anies Baswedan Presiden RI ID
10	1	Dr H Engkan Iskandar Pimpin Deklarasi Relawan Dukung Anies Baswedan Presiden
11	1	Inilah mantan Goodbener DKI pak Anies Baswedan, Calon Presiden RI yg sangat toleran thd umat beragama non muslim, dan yg terpenting tdk hobi nonton bokep sprti capres idolanya si densi alias 🤪
12	1	Tetap akan kalah sama Anies Baswedan insyaallah jadi presiden RI
13	1	In sha Allah Pak Anies Baswedan, akan jadi Presiden, semangat Pak Anies... Harkat martabat mu tak akan jatuh, hanya karna di caci di maki, di hina oleh siapa pun 😊
14	1	Ketika masyarakat yakin bahwa Anies adalah satu2 calon presiden yang bisa mensejahterakan rakyat Indonesia maka keyakinan itu dibuktikan bersama. Pilih Anies Baswedan!
15	1	Anies Baswedan presiden RI 2024

Adapun wordcloud untuk kata positif dan negatif dari hasil prediksi diatas adalah sebagai berikut:

Positive Word



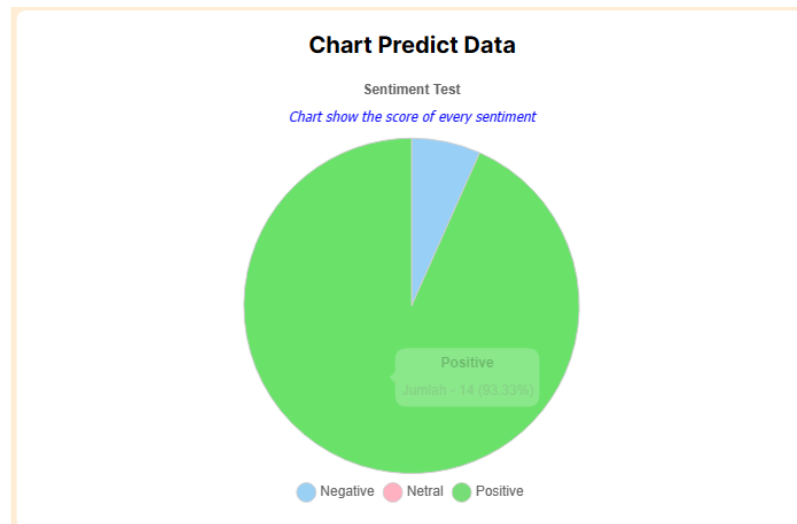
Gambar 4. 40 Gambar WordCloud Positif Twitter Anies Baswedan

Negative Word




Gambar 4. 39 Gambar WordCloud Negatif Twitter Anies Baswedan

Selanjutnya Data Test diambil dari twitter berupa 15 baris data acak dengan keyword “Ganjar Pranowo Presiden” lalu dimasukan kedalam model yang sudah di train dan hasil prediksi mesin ditunjukkan pada gambar 4.41 dan detailnya pada tabel 4.5



Gambar 4. 41 Grafik Pie Sentiment Test Twitter Ganjar Pranowo

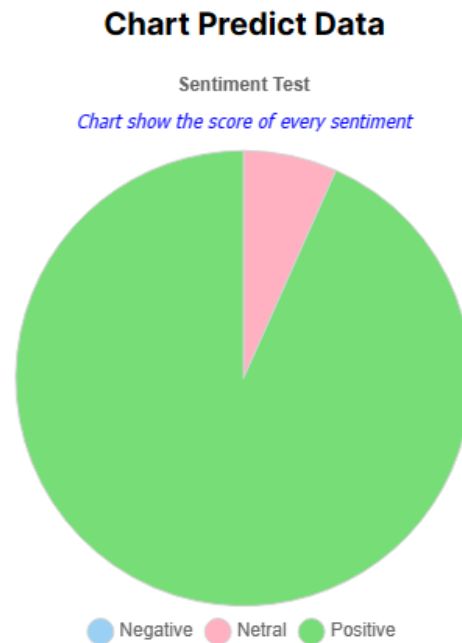
Tabel 4. 5 Tabel Datatest Twitter Ganjar Pranowo

No	Sentiment	Tweet
1	-1	Karena "Dosa" di Wadas, Netizen Tolak Calon Presiden dari PDIP, Ganjar Pranowo
2	1	Mari gunakan photo kalian di Twibbon dalam rangka mendukung Ganjar Pranowo sebagai Presiden Indonesia 2024-2029, Untuk membuat Twibbon mendukung Ganjar Pranowo sebagai Presiden Indonesia 2024-2029, ada beberapa langkah yang harus diikuti. Pembuatan Twibbon terbilang cepat dan...
3	1	Alasan PPP Dukung Ganjar Pranowo sebagai Calon Presiden 2024
4	1	Ketua Umum PDIP Megawati Soekarnoputri resmi mengumumkan Ganjar Pranowo sebagai calon presiden dari PDIP pada Pemilu 2024 mendatang.
5	1	Ganjar Pranowo the next presiden Indonesia maju ID  ID
6	1	Tidak ada yang salah Ganjar Pranowo suka sekali nonton Bokep, masalahnya jadi tak elok soal etika dan martabat karena sekarang Ganjar Pranowo seorang Capres. Jabatan

		Presiden itu harus bersih dan menjaga norma ² Budi Pekerti. #BongkarTPPUKorupsi349T #BongkarTPPUKorupsi349T
7	1	Mari gunakan photo kalian di Twibbon dalam rangka mendukung Ganjar Pranowo sebagai Presiden Indonesia 2024-2029,
8	1	Aktivis politik yang juga influencer serta konten kreator diberbagai platform sosial media, Tengku Zanzabella mendukung ganjar pranowo sebagai presiden di tahun 2024. Tengku Zanzabella menyebut Ganjar Pranowo adalah sosok tokoh yang sangat merakyat dan nasionalis
9	1	Ayo PAN dan GOLKAR mari merapat dan bergabung bersama ² mendukung Ganjar Pranowo sbg calon Presiden.
10	1	Muslim Arbi: Oligarki dan Taipan akan Menjadikan Ganjar Pranowo sebagai Presiden Indonesia 2024!
11	1	GANJAR PRANOWO, calon Presiden 2024 pilihan PDI Perjuangan 🇮🇩
12	1	SAH..... PPP, HANURA RESMI DUKUNG GANJAR PRANOWO PRESIDEN RI 2024.
13	1	Betul!! Harapan bangsa Indonesia adalah Ganjar Pranowo menjadi Presiden RI 2024. Sosok yang ramah dan dekat dgn masyarakatnya. #GanjarMenangTotal
14	1	Bismillah... Ganjar Pranowo Calon Presiden RI 2024
15	1	Aktivis politik yang juga influencer serta konten kreator diberbagai platform sosial media, Tengku Zanzabella mendukung ganjar pranowo sebagai presiden di tahun 2024. *Semua Akan Ganjar Pada Waktunya 🇮🇩 Tetap solid dan Jangan Mau Diadu Domba 🇮🇩

Adapun wordcloud untuk kata positif dan negatif dari hasil prediksi diatas adalah sebagai berikut:

Selanjutnya Data Test diambil dari twitter berupa 15 baris data acak dengan keyword “Prabowo Subianto Presiden” lalu dimasukan kedalam model yang sudah di train dan hasil prediksi mesin ditunjukkan pada gambar 4.44 dan detailnya pada tabel 4.6

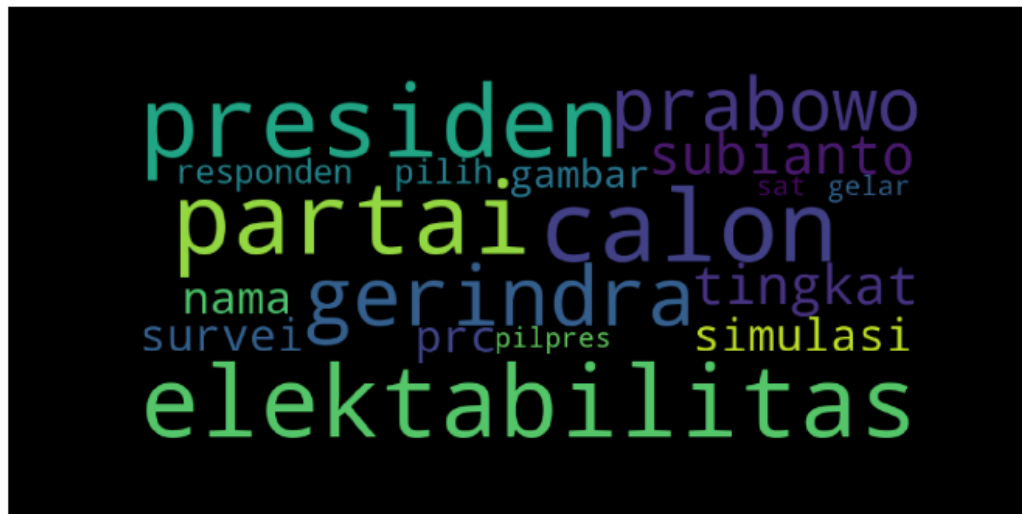


Gambar 4. 44 Grafik Pie Sentiment Test Twitter Prabowo Subianto

Tabel 4. 6 Tabel Datatest Twitter Prabowo Subianto

No	Sentiment	Tweet
1	1	Teriakan 'Presiden' Sambut Prabowo saat Tiba di Bandara Sumbar #PrabowoDicintaiMilenial #IkutPakde Prabowo Subianto
2	1	"Prabowo adalah capres Gerindra yang kita tetapkan melalui rapat pimpinan nasional. Menyebutkan bahwa calon presiden hanya tunggal, satu nama, namanya Prabowo Subianto. Calon presiden, bukan calon wakil presiden," kata Sekjen Gerindra, Ahmad Muzani.
3	1	Presiden Joko Widodo disebut-sebut ingin menduetkan Ganjar Pranowo dan Prabowo Subianto sebagai pasangan bakal calon presiden dan wakil presiden di Pemilu 2024.#newsupdate #news #update #text
4	1	Pak Prabowo Subianto tetap Calon Presiden. Inshaallah beliau Presiden RI 2024 Amiinn
5	1	Sementara dari simulasi 3 nama calon Presiden paling populer, jika pemilihan Presiden diadakan sekarang, yang akan dipilih sebagai Presiden oleh responden lagi-lagi adalah Prabowo Subianto (32.7%).

Netral Word



Gambar 4. 46 Gambar WordCloud Netral Twitter Prabowo Subianto

4.3.3. Pearson Moment Product Correlation

Pearson Moment Product Correlation atau disebut Correlation digunakan untuk mencari koefisien hubungan linear, pada penelitian ini koefisien digunakan untuk mencari keterkaitan antara data prediksi sentiment pada twitter dengan survey elektabilitas beberapa portal berita yang ditampilkan pada tabel 4.7.

Tabel 4. 7 Tabel Data Survei Elektabilitas

No	Bulan	Nama Survey	Anies Baswedan	Ganjar Pranowo	Prabowo Subianto	Lain-lain
1	Januari	LSJ (Lembaga Survei Jakarta)	15,6%	40,5%	40,6%	3,3%
2	Januari	LSI (Lembaga Survei Indonesia)	16,8%	27,2%	16%	40%
3	Januari	Lembaga survei Algoritma Research	18,7%	25,1%	16,6%	39,6%

		& Consulting				
4	Januari	Populi	10,8%	19,8%	17,1%	52,3%
5	Januari	Litbang	13,1%	25,3%	18,1%	43,5%
6	Februari	Populi	10,8%	19,8%	17,1%	52,3%
7	Februari	Litbang	13,1%	25,3%	18,1%	43,5%
8	Februari	SPIN	20%	20,6%	33%	26,4%
9	Februari	Polstat	19,4%	20,6%	33%	27%
10	Februari	Litbang	13,1%	25,3%	18,1%	43,5%
11	Februari	Indo Barometer	25,3%	30,3%	28,4%	16%
12	Maret	IPO (Indonesia Political Opinion)	31,6%	24,9%	21,1%	22,4%
13	Maret	Indikator Politik Indonesia	26,8%	36,8%	27%	9,4%
14	Maret	PolMark Research Center (PRC)	13,9%	22,8%	17,4%	45,9%
15	Maret	Indikator Politik Indonesia	21,7%	30,8%	21,7%	25,8%

Dikarenakan data survei tidak seimbang setiap bulannya yang dalam kenyataannya total lembaga yang melakukan survei berbeda tiap bulannya, maka penulis melakukan rata-rata perbulan sebelum melakukan tahapan berikutnya. Sehingga, setelah memiliki data sentiment analysis dan rata-rata data elektabilitas calon presiden per bulannya, maka dapat dihitung Korelasi Pearson Product Moment berdasarkan rumus pada bab sebelumnya.

Data yang diambil adalah data positif dan negatif, data netral diabaikan karena tidak mempengaruhi survei elektabilitas atau tidak memilih calon manapun.

Diketahui data sentiment positif, negatif, dan survei elektabilitas terhadap data sentiment calon Anies Baswedan ditampilkan pada tabel 4.8.

Tabel 4. 8 Tabel Data Sentiment dan Elektabilitas Anies Baswedan

	Positif	Negatif	Elektabilitas
Januari	88,11%	11,89%	15,00%
Februari	90,69%	9,31%	16,95%
Maret	91,92%	8,08%	23,50%

Dari data diatas, diketahui data elektabilitas tertinggi pada bulan maret dengan data survei sebesar 23,5% dan data survei terendah pada bulan januari sebesar 15%. Juga untuk data positif tertinggi pada bulan maret sebesar 91,92% dan data positif terendah pada bulan januari sebesar 88,11%. Serta data negatif tertinggi pada bulan januari sebesar 11,89% dan data negatif terendah pada bulan maret sebesar 8,08%.

Selanjutnya dibawah ini adalah rangkuman hasil Korelasi Pearson untuk calon presiden Anies Baswedan.

Tabel 4. 9 Tabel Korelasi Anies Baswedan

Korelasi Pearson	
Korelasi Sentimen Positif	0,876
Korelasi Sentimen Negatif	-0,876

Pada tabel 4.9 terlihat hasil Korelasi Pearson dengan calon presiden Anies Baswedan didapatkan nilai Korelasi Sentimen Positif sebesar 0,876 dan sebaliknya nilai Korelasi Sentimen Negatif sebesar -0,876.

Sementara itu, diketahui data sentiment positif, negatif, dan survei elektabilitas terhadap data sentiment calon Ganjar Pranowo ditampilkan pada tabel 4.10.

Tabel 4. 10 Tabel Data Sentiment dan Elektabilitas Ganjar Pranowo

	Positif	Negatif	Elektabilitas
Januari	96,22%	3,78%	27,58%
Februari	97,30%	2,70%	23,65%
Maret	94,32%	5,68%	28,83%

Dari data diatas, diketahui data elektabilitas tertinggi pada bulan maret dengan data survei sebesar 28,83% dan data survei terendah pada bulan februari sebesar 23,65%. Juga untuk data positif tertinggi pada bulan februari sebesar 97,3% dan data positif terendah pada bulan maret sebesar 94,32%. Serta data negatif tertinggi pada bulan maret sebesar 5,68% dan data negatif terendah pada bulan februari sebesar 2,7%.

Selanjutnya dibawah ini adalah rangkuman hasil Korelasi Pearson untuk calon presiden Ganjar Pranowo.

Tabel 4. 11 Tabel Korelasi Ganjar Pranowo

Korelasi Pearson	
Korelasi Sentimen Positif	0,894
Korelasi Sentimen Negatif	-0,894

Pada tabel 4.11 terlihat hasil Korelasi Pearson dengan calon presiden Ganjar Pranowo didapatkan nilai Korelasi Sentimen Positif sebesar 0,894 dan sebaliknya nilai Korelasi Sentimen Negatif sebesar -0,894.

Sedangkan, diketahui data sentiment positif, negatif, dan survei elektabilitas terhadap data sentiment calon Prabowo Subianto ditampilkan pada tabel 4.12.

Tabel 4. 12 Tabel Data Sentiment dan Elektabilitas Prabowo Subianto

	Positif	Negatif	Elektabilitas
Januari	99,05%	0,95%	21,68%
Februari	99,23%	0,77%	24,62%
Maret	99,79%	0,21%	23,12%

Dari data diatas, diketahui data elektabilitas tertinggi pada bulan februari dengan data survei sebesar 24,62% dan data survei terendah pada bulan januari sebesar 21,68%. Juga untuk data positif tertinggi pada bulan maret sebesar 99,79% dan data positif terendah pada bulan januari sebesar 99,05%. Serta data negatif tertinggi pada bulan januari sebesar 0,95% dan data negatif terendah pada bulan maret sebesar 0,21%.

Selanjutnya dibawah ini adalah rangkuman hasil Korelasi Pearson untuk calon presiden Prabowo Subianto.

Tabel 4. 13 Tabel Korelasi Prabowo Subianto

Korelasi Pearson	
Korelasi Sentimen Positif	0,970
Korelasi Sentimen Negatif	-0,970

Pada tabel 4.13 terlihat hasil Korelasi Pearson dengan calon presiden Prabowo Subianto didapatkan nilai Korelasi Sentimen Positif sebesar 0,97 dan sebaliknya nilai Korelasi Sentimen Negatif sebesar -0,97.

BAB 5

SIMPULAN DAN SARAN

5.1.Simpulan

Berdasarkan butir-butir yang telah ditetapkan pada rumusan masalah dan tujuan penelitian, maka kesimpulan dari penelitian ini adalah sebagai berikut:

1. Implementasi Mesin Sentiment Analysis dengan algoritma Support Vector Machine menggunakan kernel Gaussian RBF ditemukan fold terbaik untuk masing-masing calon presiden yaitu fold ke-8 dengan f1-score 0,66 untuk calon Anies Baswedan dengan total 2.554 data training dan 283 data testing, fold ke-5 dengan f1-score 0,72 untuk calon Ganjar Pranowo dengan total 3.330 data training dan 370 data testing, dan fold ke-4 dengan f1-score 0,78 untuk calon Prabowo Subianto dengan total 3487 data training dan 387 data testing.
2. Didapatkan akurasi untuk Mesin Sentiment Analysis dengan algoritma Support Vector Machine menggunakan kernel Gaussian RBF dengan rata-rata akurasi 88,75% untuk calon presiden Anies Baswedan, akurasi 93,05% untuk calon presiden Ganjar Pranowo, akurasi 92,79% untuk calon presiden Prabowo Subianto.
3. Untuk calon presiden Anies Baswedan, korelasi antara survei elektabilitas dengan hasil analisis sentiment positif adalah 0,876 dan sentiment negatif adalah -0,876. Sehingga kesimpulan yang diperoleh korelasi memiliki hubungan kuat yang searah. Untuk calon presiden Ganjar Pranowo, korelasi antara survei elektabilitas dengan hasil analisis sentiment positif adalah 0,894 dan sentiment negatif adalah -0,894. Sehingga kesimpulan yang

diperoleh korelasi memiliki hubungan kuat yang searah. Untuk calon presiden Prabowo Subianto, korelasi antara survei elektabilitas dengan hasil analisis sentiment positif adalah 0,97 dan sentiment negatif adalah -0,97. Sehingga kesimpulan yang diperoleh korelasi memiliki hubungan kuat yang searah. Sehingga dapat disimpulkan dari 3 calon presiden bahwa semakin tinggi sentimen positif, maka semakin tinggi elektabilitas.

5.2.Saran

Berdasarkan penelitian yang sudah dilakukan, maka saran yang dapat penulis berikan untuk pengembangan penelitian selanjutnya adalah:

1. Berdasarkan hasil proses scraping, ada kendala untuk melakukan penarikan data pada twitter sehingga disarankan untuk mencari alternatif lain saat mencari sumber data.
2. Untuk topik sentiment analysis, disarankan untuk mendalami topik parameter tuning disertai dengan penggunaan kernel lain seperti linear, polynomial, dan sigmoid dengan tema calon presiden pada penelitian selanjutnya.

DAFTAR PUSTAKA

- [1] Bps, “Jumlah Pemilih 2004-2014.”
- [2] B. W. Sari and F. F. Haranto, “IMPLEMENTASI SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP PELAYANAN TELKOM DAN BIZNET,” *Jurnal Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 171–176, Sep. 2019, doi: 10.33480/pilar.v15i2.699.
- [3] Fatihah Rahmadayana and Yuliant Sibaroni, “Sentiment Analysis of Work from Home Activity using SVM with Randomized Search Optimization,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 5, pp. 936–942, Oct. 2021, doi: 10.29207/resti.v5i5.3457.
- [4] S. Fendyputra Pratama, R. Andrean, and A. Nugroho, “Analisis Sentimen Twitter Debat Calon Presiden Indonesia Menggunakan Metode Fined-Grained Sentiment Analysis,” *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 4, no. 2, pp. 2541–3619, 2019, doi: 10.31328/jo.
- [5] D. W. Seno and A. Wibowo, “Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Dengan Metode Lexicon Based Dan Support Vector Machine,” *Jurnal Ilmiah FIFO*, vol. 11, no. 2, p. 144, Nov. 2019, doi: 10.22441/fifo.2019.v11i2.004.
- [6] D. Darwis, E. Shintya Pratiwi, A. Ferico, and O. Pasaribu, “PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA,” 2020.
- [7] A. S. Arief, “SENTIMENTANALYSIS REVIEW APLIKASI MENGGUNAKAN ALGORITMA SVM PADA APLIKASI MYPERTAMINA,” 2023.
- [8] D. B. Rarasati and J. C. A. Putra, “Correlation Between Twitter Sentiment Analysis with Three Kernels Using Algorithm Support Vector Machine (SVM) Governor Candidate Electability Level,” *COIESE*, pp. 249–256, 2021.
- [9] E. Indrayuni, “KOMPARASI ALGORITMA NAIVE BAYES DAN SUPPORT VECTOR MACHINE UNTUK ANALISA SENTIMEN REVIEW FILM,” *Jurnal PILAR Nusa Mandiri*, vol. 14, no. 2, p. 175, 2018, [Online]. Available: <http://www.bsi.ac.id>
- [10] N. Arifin, U. Enri, and N. Sulistiyowati, “PENERAPAN ALGORITMA SUPPORT VECTOR MACHINE (SVM) DENGAN TF-IDF N-GRAM

- UNTUK TEXT CLASSIFICATION,” *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 6, no. 2, pp. 129–136, 2021.
- [11] F. Rahutomo, P. Y. Saputra, and M. A. Fidyawan, “IMPLEMENTASI TWITTER SENTIMENT ANALYSIS UNTUK REVIEW FILM MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE,” *Jurnal Informatika Polinema*, vol. 4, no. 2, pp. 93–100, 2018.
 - [12] M. Romzi and B. Kurniawan, “PEMBELAJARAN PEMROGRAMAN PYTHON DENGAN PENDEKATAN LOGIKA ALGORITMA,” *Jurnal Teknik Informatika Mahakarya*, no. 2, pp. 37–44, 2020.
 - [13] A. Hutapea and M. Tanzil Furqon, “Penerapan Algoritme Modified K-Nearest Neighbour Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia,” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, pp. 3957–3961, 2018, [Online]. Available: <http://j-ptiik.ub.ac.id>
 - [14] S. Diantika, W. Gata, H. Nalatissifa, and M. Lase, “Komparasi Algoritma SVM Dan Naive Bayes Untuk Klasifikasi Kestabilan Jaringan Listrik,” *JURNAL ILMIAH ELEKTRONIKA DAN KOMPUTER*, vol. 14, no. 1, pp. 10–15, 2021.
 - [15] R. Risnantoyo, A. Nugroho, and K. Mandara, “Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm,” *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, vol. 4, no. 1, pp. 86–96, Jul. 2020, doi: 10.31289/jite.v4i1.3798.
 - [16] H. Florenci Tapikap, B. S. Djahi, and T. Widiastuti, “KLASIFIKASI SPAM E-MAIL MENGGUNAKAN METODE TRANSFORMED COMPLEMENT NAÏVE BAYES (TCNB),” *J-ICON*, vol. 7, no. 1, pp. 21–26, 2019.
 - [17] F. Istighfarizkya, N. A. S. ER, I. M. Widiarthaa, L. G. Astutia, I. G. N. A. C. Putra, and I. K. G. Suhartana, “Klasifikasi Jurnal menggunakan Metode KNN dengan Mengimplementasikan Perbandingan Seleksi Fitur,” *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 11, pp. 167–176, 2022, [Online]. Available: <https://scholar.google.com>
 - [18] O. H. Anidjar, A. Barak, B. Ben-Moshe, E. Hagai, and S. Tuvyahu, “A Stethoscope for Drones: Transformers Based Methods for UAVs Acoustic Anomaly Detection,” *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3262702.
 - [19] F. Satria, Zamhariri, and M. A. Syaripudin, “Prediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma C4.5 Pada Fakultas Dakwah Dan Ilmu Komunikasi UIN Raden Intan Lampung,” *Jurnal Ilmiah MATRIK*, vol. 22, pp. 28–35, 2020.

RIYAWAT HIDUP

Nama	:	Michael Alfonso
N I M	:	32190039
Tempat/tgl lahir	:	Tangerang / 01-06-2001
Jenis Kelamin	:	Laki-laki
Alamat	:	Perumahan Simprug diPoris Blok F1/10a, Kelurahan Poris Gaga Baru, Kecamatan Bataceper, Kota Tangerang, Banten, 15122
No. Telp	:	085863599547



Riwayat Pendidikan

Tahun 2007 s/d 2014	SD, SD Mutiara Bangsa 1, Tangerang
Tahun 2014 s/d 2017	SMP, SMP Mutiara Bangsa 1, Tangerang
Tahun 2017 s/d 2019	SMK, SMK Mutiara Bangsa 1, Tangerang
Tahun 2019 s/d 2023	S1, Universitas Bunda Mulia, Tangerang

Pengalaman Kerja

Tahun 2022 s/d 2022	Service Development Specialist, PT Sumber Alfaria Trijaya TBK
Tahun 2022 s/d Sekarang	LA Support, Universitas Bunda Mulia

LAMPIRAN