BAB II
TINJAUAN PUSTAKA

A. Penelitian Terdahulu

Berikut ini merupakan penelitian dengan topik analisis sentimen dengan menggunakan metode *Support Vector Machine* (SVM) seperti pada Tabel 2.1

Tabel 2.1 Penelitian terdahulu

No	Peneliti	Hasil Penelitian	Metode
1.	(Tineges dkk., 2020)	"Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine" penelitian itu menghasilkan sebuah metode SVM memiliki tingkat keakurasian sebesar 87% dengan ketepatan antara hasil prediksi sebenarnya sebesar 86%, kemudian tinggkat keberhasilan system sebesar 95% kemudian tingkat kesalahan data (error rate) 13% dengan nilai rata-rata *precission* dan *recall* sebesar 90%. Dengan hasil klasifikasi bernilai positif sebesar 18,4% dan negative 81,6%.	*Support Vector Machine (SVM)*
2.	(Fitriana dkk., 2021)	"Analisis Sentimen Opini masyarakat Terhadap Vaksin Covid-19 pada Sosial Media Twitter menggunakan Support Vector Machine dan Naïve Bayes" pada penelitian in metode SVM memiliki tingkat keakurasian yang lebih baik daripada metode NBC tetapi dari segi efisiensi eaktu dalam melakukan proses kalsifikasi metode NBC leboh unggul dari metode SVM. Hasil akurasi dari metode SVM mencapai 90,47% sedangkan metode NBC 88,64%.	*Support Vector Machine (SVM)*

Tabel 2.1 Penelitian terdahulu lanjutan

No.	Peneliti	Hasil Penelitian	Metode
3.	(Lukmana dkk., 2019)	"Analisis Sentiemn Pada Calon Presiden 2019 Dengan Support Vector Machine di Twitter" pada penelitian ini klasifikasi menggunakan algoritma SVM dengan *kernel* dapat melakukan klasifikasi dengan tingkat keakurasian 86.82% untuk tweet dengan kata kunci "jokowi" dan 86.27% untuk tweet prabowo. Pada penelitian ini menggunakan data tweet sebanyak 20.000 dengan masing-masing calon sebanyak 10.000 tweet.	*Support Vector Machine (SVM)*
4.	(Baita dkk., 2021)	"Analisis Sentimen Mengenai Vaksin Sinovac Menggunakan Algoritma Support Vector Machine (SVM) dan K-Nearest Neighbor (KNN)" pada penelitian ini algoritma SVM memiliki rata-rata hasil akurasi sebesar 0.7 sedangkan KNN 0.56. hasil dari penelitian ini menunjukan bahwa algortima SVM memiliki performa yang lebih baik ketika dipadukan dengan fungsi linear. Namun nilai akurasi tergolong rendah dari kedua metode hal tersebut dikarenakan pelabelan otomatis dengan text blob.	*SVM dan KNN*
5.	(Fikri dkk., 2020)	"Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter" hasil dari penelitian ini menunjukan bahwa algoritma NBC dan SVM memiliki nilai akurasi yang cukup baik antara SVM dan NBC, dimana algoritma NBC mendapatkan hasil akurasi 73,65% dan algoritma SVM 70,20% dengan menggunakan data sebanyak 2030 record.	*SVM* dan NBC

Tabel 2.1 Penelitian terdahulu lanjutan

No.	Peneliti	Hasil Penelitian	Metode
6.	(Setiawan & Utami, 2021)	"Analisi Sentimen Mengenai Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naïve Bayes" pada penelitian ini algoritma NBC lebih unggul dari segi waktu pemrosesan dari algoritma SVM, algoritma NBC mendapatkan hasil terbaik pada iterasi ke 1 sedangkan algoritma SVM mendapatkan hasil terbaik pada iterasi ke 423, tetapi algortima SVM mendapat nilai akurasi yang cukup bagus dengan memperoleh nilai akurasi sebesar 85% dan algortima NBC sebesar 81,20%. Dengan hasil akurasi 85% dengan waktu 31,60 detik dengan tingkat akurasi *recall* sebesar 84% dan *precission* 83,60% maka algoritma SVM dapat dikatakan lebih unggul dari algoritma NBC dalam melakukan klasifikasi.	SVM dan *Naive Bayes Classifier* (*NBC*)
7.	(Brian, Pandemic & Laurensz, Sediyono, 2021)	"Analisis Sentimen Masyarakat Terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19" hasil dari penelitian ini menunjukan bahwa analisis sentimen dengan menggunakan algoritma SVM memiliki nilai akurasi lebih baik dari algoritma NBC, nilai akurasi yang diperoleh berdasarkan data yang sama yaitu "Vaksin Sinovac" dan "Vaksin Merah Putih" metode SVM lebih unggul dalam hal tinggat akurasi dibandingkan dengan metode SVM.	*Support Vector Machine (SVM)* dan NBC

Tabel 2.1 Penelitian terdahulu lanjutan

No.	Peneliti	Hasil Penelitian	Metode
8.	(Sujadi, 2022).	"Analisis Sentimen Pengguna Media Sosial Twitter Terhadap Wabah Covid-19 Dengan Menggunakan Metode Naïve Bayes Classifier dan Support Vector Machine" juga menghasilkan nilai akurasi yang baik dimana algoritma SVM memeproleh nilai akurasi sebesar 81.6%. Sedangkan NBC dengan akurasi 78.3% dengan nilai *Precision* 77.9%, *Recall* 78.2% dan nilai rata-rata *f1-score* 78%. Pada penelitian ini juga menggunakan metode pembobotan per kata menggunakan metode TF-IDF.	SVM dan *Naive Bayes Classifier* (*NBC*)
9.	(Kevin dkk., 2020)	"Analisis Sentimen Terhadap Transportasi Online Menggunakan Support Vector Machine Berbasis *Particle Swarm Optimization*" dalam penelitian ini menghasilkan kesimpulan bahwa algoritma SVM dapat melakukan klasifikasi dengan tingkat keakursian yang lebih baik darpida metode SVM-PSO. Namun dari hasil pengujian menggunakan 10 *k-fold cross validation* mendapatkan nilai akurasi dan AUC metode SVM sebesar 95,46% AUC0,979 sednagkan pada metode SVM-PSO mendapatkan 96,04% dan AUC 0,993.	*SVM dan SVM-PSO*

Tabel 2.1 Penelitian terdahulu lanjutan

No.	Peneliti	Hasil Penelitian	Metode
10.	(Arsi & Waluyo, 2021)	"Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine" pada penelitian ini dengan menggunakan data tweet sebanyak 1.236 dengan 404 tweet positif dan 832 tweet negative menghasilkan analisis sentimen dengan tingkat keakurasian 96,68%. Selain itu tingkat *precission* dan *recall* cukup tinggi yaitu 95.82% dan 94.04% dengan nilai AUC 0,979. Dengan hasil klasifikasi bahwa masyarakat cenderung memiliki sentimen negative terkait pemindahan inu kota Indonesia.	*Support Vector Machine (SVM)*
11.	(Aldisa & Maulana, 2022)	"Analisis Sentimen Opini Masyarakat Terhadap Vaksinasi Booster COVID-19 Dengan Perbandingan Metode Naïve Bayes, Becission Tree dan SVM" pada penelitian ini algortima NBC menghasilkan nilai yang dapat dikatakan baik yaitu 83.82% dengan menggunakan dataset positif sebanyak 2248 dan untuk data set negative sebanyak 751.	*SVM*, NBC, Decission Tree
12.	(Siswanto dkk., 2022)	"Analisis Sentimen Publik Mengenai Perekonomian Indonesia Pada Masa Pandemi Covid-19 di Twitter Menggunakan Metode Klasifikasi K-NN dan SVM" dari penelitian iini algoritma SVM lebih unggul dalam akurasi untukmelakukan analisis sentimen. Nilai yang diperoleh dari algoritma SVM sebesar 78% dan untuk K-NN sebesar 76%, dengan hasil tersebut algoritma SVM dapat dikatakan lebih baik dari algoritma K-NN walaupun nilai	*Support Vector Machine (SVM)* dan *KNN*

		dari hasil akurasi K-NN tidak terpau jauh dari metode SVM.	

Tabel 2.1 Penelitian terdahulu lanjutan

No.	Peneliti	Hasil Penelitian	Metode
13.	(Raihan dkk., 2022)	Penelitian dengan judul "Analisis Sentimen Terhadap Bakal Calon Presiden 2024 Dengan Algirotma Naïve Bayes" menghasilkan sebuah kesimpulan bahwa setiap data set yang ada warganet cenderung memiliki sentimen positif terhadap 3 tokoh yaitu Ganjar Pranwo, Ridwan Kamil, Anies dan Prabowo. Dengan menggunakan algoritma NBC akurasi yang didapat pada penelitian ini mencapai nilai tertinggi pada skor 73,68% yaitu pada dataset Ganjar lebih tinggi dari dataset Prabowo yang memiliki nilai akurasi 60%, kemudian Anies 71,43% dan Ridwan Kamil 62,5%.	*Naïve Bayes Classifier (NBC*
14.	(Soer, 2022)	" Analisis Sentimen Terhadap Pemerintahan Ridwan Kamil Sebagai Gubernur Jawa Barat Menggunakan Algoritma Naïve Bayes " pada peneilitian menghasilkan kesimpulan bahwa nilai akurasi yang didapat dengan menggunakana algoritma NBC dengan penggunaan teknik cross validation menghasilkan nilai akurasi sebesar 84,38% dengan tingkat respon positif masyarakat terhadap ridwan kamil sebesar 49%.	*Naïve Bayes Classifier (NBC)*

Berdasarkan Tabel 2.1 diketahui bahwa berbagai macam topik penelitian menggunakan metode SVM menghasilkan tingkat akurasi yang berbeda-beda. Hasil akurasi terbaik diperoleh dengan tingkat akurasi sebesar 96.68% dengan menggunakan algoritma *Support Vector Machine (*SVM*)*.

B. NLP

Natural Language Processing (NLP) adalah disiplin ilmu komputer yang memiliki tujuan untuk memahami konsep bahasa dari manusia. Tidak seperti manusia yang mahir dalam memahami tata bahasa serta sebuah hubungan yang tersirat, sementara komputer memiliki kendala dalam pengolahan *query* bahasa alami. Dalam kasusnya analisis sentimen merupakan versi mini dari NLP. NLP telah banyak digunakan dari tahun ketahun dalam penyelesaian kasus analisis sentimen. Dengan memanfaatkan library yang ada pada NLP sehingga memudahkan dalam pemrosesan bahasa (Liu, 2015).

C. Analisis Sentimen

Analaisis Sentimen merupakan sebuah proses untuk menentukan sentimen dari seseorang yang nantinya direpresentasikan kedalam bentuk teks dan dikategorikan kedalam sentimen positif, negative atau neutral. Analisis sentimen juga biasa disebut dengan istilah penambangan opini. Tujuan dari analisis sentiment adalah untuk mendefinisikan alat otomatis yang mampu mengekstrak sebuah informasi dari sebuah teks kedalam bahasa alami seperti pendapat dan sentimen(Pozzi dkk., 2016).

Fungsi dari analisis ini adalah untuk mendapatkan informasi dari sebuah data. Proses ini dilakukan dengan menggunakan *Mechine Learning* karena dapat melakukan prediksi sentimen (positif, negative, netral) dengan memanfaatkan data testing dan data training (Kevin dkk., 2020).

D. Twitter

Twitter merupakan sebuah *platform* sosial media *microblogging* yang memungkinkan para penggunanya mengirimkan sebuah pesan dan postingan twitter secara realtime. Pesan dalam twitter umunya disebut dengan *tweet*, *tweet* merupakan pesan singkat yang memiliki batasan panjang karakter

mencapai 140 karakter saja pada tiap *tweet*-nya (Nurrun Muchammad Shiddieqy dkk., 2016).

Twitter merupakan *platform* sosial media yang banyak digunakan oleh masyarakat di berbagai dunia untuk mengutarakan berbagai pendapat mengenai suatu hal, sehingga opini pun sering kali muncul terhadap sebuah *tweet* tersebut. Opini itulah yang dapat dimanfaatkan untuk melakukan sebuah penelitian dengan menggunakan sebuah algoritma untuk melakukan klasifikasi. Oleh karena itu, penulis memilih twitter sebagai tempat untuk mencari sumber data.

E. Scrapping

Scrapping data atau web *scrapping* merupakan sebuah teknik yang digunakan untuk melakukan sebuah ekstraksi data berupa informasi dari sebuah website kemudian data yang berhasil diambi akan disimpan kedalam sebuah format tertentu (Rakhmawati dkk., 2020). Kemudian data yang berhasil dikumpulkan tadi selanjutnya akan diproses dan dilakukan analisa oleh penulis. Pada penelitian ini penulis akan melakukan analisis sentimen terhadap opini masyarakat berdasarkan data tweet yang berhasil dikumpulan tadi kedalam tiga kelas sentimen yaitu positif, neagtif dan netral.

F. Pengumpulan Data

Pengumpulan data dilakukan dengan memanfaatkan library Snscrape. Data yang diambil berupa data *tweet* masyarakat yang di ambil secara acak dangan melakukan pencarian berdasarkan kata kunci #GanjarPranowo yang memiliki keterkaitan dengan Gubernur Jawa Tengah Ganjar Pranowo.

Pengambilan data dilakukan dari tanggal 17/05/2019 – 1/9/2022. Dari data yang berhasil dikumpulkan, data tersebut nantinya akan dibuat sebuah *dataset* dengan mengklasifikanya menjadi dua kelas sentimen dengan nilai positif dan negative. Hasil dari klasifikasi *dataset* tersebut selanjutnya akan digunakan untuk membuat pelabelan data dimana hasil dari pelabelan data

tersebut nantinya akan dibagi menjadi tiga kelas sentimen yang memiliki nilai positif, negative, dan netral.

G. Prepocessing

Prepocesing merupakan tahap pengolahan data berdasarkan data yang berhasil di ambil dari twitter. Data yang berhasil diambil dapat dikatakan sebagai data mentah karena data yang didapat belum sepenuhnya dapat langsung digunakan. Pada tahapan ini kumpulan data yang berhasil diambil atau data yang masih mentah nantinya akan diproses dengan beberapa tahap *text procesing* sehingga data tersebut menjadi sebuah bentuk data yang lebih mudah untuk dipahami. Tahapan text processing meliputi *cleansing, case folding, tokenizing, stopword removal, stemming, remove duplicate.*

a) Cleansing

Tahap ini merupakan tahapan dimana data text yang berhasil kita ambil pada twitter selanjutnya akan diproses dengan melakukan cleansing data. Dimana data yang akan digunakan untuk dilakukan analisis sentimen dibersikan terlebih dahulu. Pembersihan data meliputi penghapusan link, menghapus tagar(#), menghapus mentions, menghapus retweet, dan menghapus berbagai symbol yang terdapat pada text tersebut. Contoh symbol tersebut anatara lain %,!,*,&,[ ],?,_,\, dll.

b) Case Folding

Tahap ini merupakan tahap untuk mnegubah huruf besar ke huruf kecil semua (*lowercase*).

c) Tokenizing

Tokenizing merupakan tahap pemotongan atau pemisahan kata dari sebuah kalimat.

d) Stopword Removal

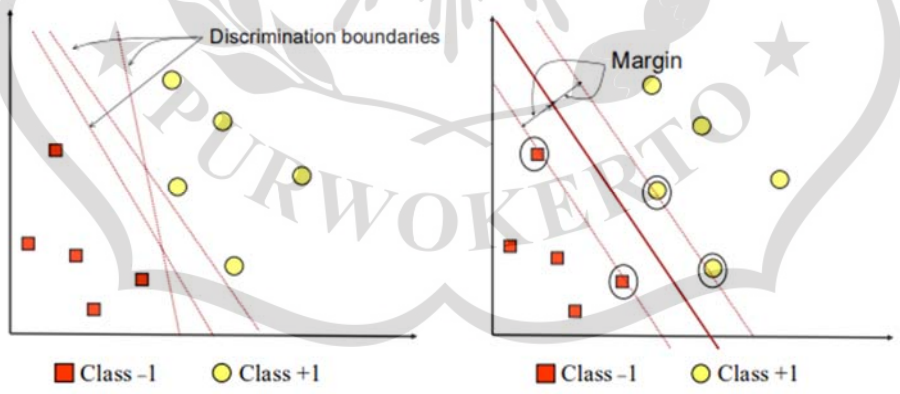
Stopword removal adalah tahap penghapusan atau penghilangan kata yang dianggap tidak penting atau kata yang tidak memiliki makna yang berarti menggunakan library NLTK.

e) Stemming

Stemming merupakan metode yang digunakan untuk merubah kata yang memiliki imbuhan menjadi kata dasar. Pada proses ini menggunakan library sastrawi.

H. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah metode *supervised learning* yang membutuhkan tahap *sequintal training* terlebih dahulu untuk melakukan implementasinya dan harus melalui tahap pengujian sebelum melakukan implementasi. Metode *Support Vector Machine* (SVM) juga memiliki kelebihan dalam melakukan identifikasi hyperplane agar dapat memaksimalkan margin dari kelas yang berbeda. *Support Vector Machine* (SVM) juga memiliki kekurangan yaitu fitur yang sama dapat mempengaruhi hasil akurasi dari metode SVM (Brian, Pandemic & Laurensz, Sediyono, 2021). Konsep dari *Support Vector Machine* (SVM) adalah hyperplane yang berfungsi untuk memisah kelas data positif dan negative. Tahap awal SVM adalah dengan mengkonversi data text kedalam bentuk vector untuk kemudian dilakukan pembobotan (Kevin dkk., 2020).



Gambar 2.1 Hyperplane SVM (Pisner & Schnyer, 2019)

SVM akan melakukan prediksi suatu kelas dari data yang ada yaitu dengan melakukan pelabelan daerah kelas yang sama berdasarkan tempat dari data tersebut. Prinsip dari metode SVM adalah dengan membangun

hyperplane dengan sebuah margin yang sama dan tidak saling mendekati daerah salah satu kelas dengan kelas lain (Fikri dkk., 2020).

Berikut ini merupakan persamaan metode SVM menurut (Widayani, 2021) seperti pada rumus persamaan (2.1) dan (2.2).

$$[(w^T.x_i + b)] \geq 1 \; untuk \; y_i = +1 \quad (2.1)$$

$$[(w^T.x_i + b)] \leq 1 \; untuk \; y_i = -1 \quad (2.2)$$

w adalah normal bidang sedangkan *b* adalah posisi dari bidang relative terhadap titik koordinat.

I. Confusion *Matrix*

Confusion matrix merupakan table yang merepresentasikan sebuah informasi berupa prediksi dari sebuah perbandingan yang dilakukan oleh system dari hasil klasifikasi. Dalam table confusion matrix menyajikan jumlah dari data uji yang memiliki nilai klasifikasi dengan nilai true dan jumlah data uji yang memiliki nilai klasifikasi bernilai false. Dalam confusion matrix juga memiliki istilah seperti True Positif (TP), True Negative (TN), Fasle Positif (FP) dan False Negatif (FN) isitilah tersebut biasanya digunakan untuk menghitung nilai *akurasi*, *recall*, *precision* dan *f1-score* (Fikri dkk., 2020).

Tabel 2.2 Confussion matrix

	Nilai Prediksi		
Nilai Actual	TPos	FPosNeg	FPosNeg
	FNegPos	TNeg	FNegNet
	FNetPos	FNetNeg	TN

Berikut merupakan perhitungan untuk mengetahui nilai *akurasi*, *precission*, *recall*, dan *f1-score*.

$$\text{Akurasi} = \frac{TPos + TNeg + TNet}{Total\ Data}\ x\ 100\% \tag{2.3}$$

Akurasi adalah nilai yang ditampilkan dar hasil pemodelan yang telah kita buat. Untuk rumus persamaan pencarian nilai akurasi dapat dilihat pada persamaan (2.3).

$$\text{Precission} = \frac{TP}{TP + FP}\ x\ 100\% \tag{2.4}$$

Precission adalah tingkat ketepatan informasi yang diminta dengan jawaban dari system. Untuk prsamaannya dapat dilihat pada persamaan (2.5).

$$\text{Recall} = \frac{TP}{TP + FN}\ x\ 100\% \tag{2.5}$$

Recall merupakan tingkat keberhasilan system dalam menemukan sebuah informasi baru lagi. Untuk perhitungan nilai recall dapat dilihat pada persamaan (2.5).

$$F1 -\ score =\ 2\ x\ \frac{Precission\ x\ recall}{Precission\ +recall} \tag{2.6}$$

F1-score merupakan perhitungan yang mana penggabungan data recall dan precission seperti pada persamaan (2.6).

J. Visualisasi Data

Dalam tahap ini merupakan tahap dimana hasil dari proses analisis divisualisasikan kedalam bentuk *wordcloud*. *Wordcloud* merupakan visualisasi data yang menampilkan output dari analisis sentimen yang digambarkan dengan karakterisstik yang terdapat pada text, data yang ditampilkan merupakan kumpulan kata yang berbeda ukuran besar hurufnya. Semakin besar kata pada *wordcloud* berarti tingkat kemunculan kata tersebut semakin besar tingkat kemunculannya (Lukmana dkk., 2019). Data yang ditampilkan merupakan data hasil akhir dari analisis sentimen yang sudah dilakukan sebelumnya.