PERBANDINGAN METODE NAÏVE BAYES DAN SVM UNTUK SENTIMEN ANALISIS MASYARAKAT TERHADAP SERANGAN RANSOMWARE PADA DATA KIP-K

Nabil Safiq Ramadan[1], Dedi Darwis[2]
Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
Jl. ZA. Pagar Alam No.9 -11, Labuhan Ratu, Kec. Kedaton, Bandar Lampung
e-mail: *[1]nabil_safiq_ramadan@teknokrat.ac.id, [2]darwisdedi@teknokrat.ac.id

Abstract

This research examines ransomware attacks on KIP-K data by analyzing the opinions of Social Media X users, using the naïve bayes classifier (NBC) and support vector machine (SVM) methods. The rapid development of technology not only brings great benefits but also increases the risk of digital attacks by certain parties. One example is a ransomware attack that caused a KIP-K data leak. In this study, sentiment analysis was applied to identify public opinions or responses obtained from Social Media X, with the help of python programming and google colab. Of the total 2,648 raw data collected, pre-processing was carried out resulting in 1,738 cleaned data. The study compared two methods, namely Naïve Bayes and Support Vector Machine, to determine what method is more effective in analyzing public sentiment related to ransomware attacks on KIP-K data. The focus of this study is to understand the percentage of Social Media X users' comments and responses related to the KIP-K ransomware taken from media sosial X. The stages of sentiment analysis in this study include crawling, labeling, preprocessing, method classification, and visualization. Before the classification process was carried out, the data was divided into two parts, namely 30% for test data and 70% for training data. Data labeling resulted in 1,313 negative data, 957 positive data and 377 neutral data. The classification results show that the NBC method has an accuracy of 70%, while the SVsM achieves an accuracy of 88%. Based on these results, SVM is proven to be superior in data analysis compared to NBC, especially for big data.

Keyword: *Kip-K, Naive Bayes, Ransomware, SVM, Sentiment Analysis*

PENDAHULUAN

Munculnya media sosial telah mengubah pola perilaku individu dalam hal budaya, etika, dan konvensi yang berlaku. Indonesia, yang dicirikan oleh jumlah penduduk yang besar dan beragamnya suku, ras, dan agama, memiliki potensi yang signifikan untuk transformasi sosial di berbagai lapisan masyarakat. Mayoritas masyarakat Indonesia memanfaatkan media sosial untuk memperoleh dan menyebarkan informasi kepada masyarakat (Fatmawati, 2021). Ransomware adalah jenis perangkat lunak berbahaya yang mengunci atau mengenkripsi data korban dan kemudian meminta bayaran sebagai imbalan untuk memungkinkan mereka mengaksesnya kembali. Dikutip dari Kompas.com "akibat peretasan tersebut, data di 282 layanan kementerian/lembaga negara hilang dan sulit dikembalikan, termasuk bidang pendidikan, yang bertanggung jawab atas penyimpanan data penting seperti Kartu Indonesia Pintar-Kuliah" (Anjelina, 2024). Penelitian ini membandingkan dua metode, yaitu "Naïve Bayes dan Support Vector Machine", guna menentukan metode apa yang lebih efektif dalam menganalisis sentimen masyarakat terkait serangan ransomware pada data kip-k. Melalui perbandingan ini, diharapkan dapat diketahui mana dari kedua metode tersebut yang mampu mengelola dan mengklasifikasikan sentimen publik dengan lebih akurat dan efisien.

Berdasarkan "Undang-Undang No. 12 Tahun 2012 mengenai pendidikan tinggi, pemerintah indonesia bertanggung jawab untuk memperluas akses dan kesempatan belajar di perguruan tinggi, serta mempersiapkan generasi indonesia yang cerdas dan kompetitif". Pemerintah akan terus berupaya menjamin agar anak muda Indonesia dari keluarga prasejahtera, khususnya yang memiliki kemampuan luar biasa, dapat menempuh pendidikan tinggi melalui Program Indonesia Pintar (PIP) (Winston Talakua et al., 2023). PIP merupakan bantuan keuangan pendidikan yang diberikan kepada anak usia 6 sampai 21 tahun dari keluarga prasejahtera. Program ini merupakan pengembangan dari Bantuan Mahasiswa Tidak Mampu

(BSM) dan dilaksanakan sesuai dengan Kebijakan Presiden Nomor 7 Tahun 2014 untuk mewujudkan dan membina keluarga produktif (Agusman, 2019). Pendidikan tinggi PIP bagi mahasiswa difasilitasi melalui Kartu Indonesia Pintar Kuliah (KIP Kuliah).

Berdasarkan petunjuk teknis pengelolaan KIP Kuliah, perguruan tinggi diharuskan membentuk tim verifikator untuk memeriksa kelayakan calon mahasiswa penerima (Winston Talakua et al., 2023). Bidikmisi yang sudah diterima di perguruan tinggi. Proses verifikasi meliputi pemeriksaan terhadap kondisi ekonomi, potensi akademik, daerah asal, evaluasi dokumen pendukung, serta pertimbangan khusus lainnya. Dalam beberapa tahun terakhir, terlihat bahwa pemerintah indonesia terus berupaya meningkatkan jumlah penerima KIP Kuliah (Amilia et al., 2020). Pada tahun 2018, kuota yang disediakan mencapai 90.000. Sementara itu, perguruan tinggi yang mengelola KIP Kuliah masih menggunakan metode manual untuk mengevaluasi berkas, mengolah data, dan menentukan penerima beasiswa. Cara ini bisa memakan waktu lebih lama seiring dengan bertambahnya jumlah pendaftar, menjadi kurang efektif, dan dapat menghasilkan keputusan yang kurang objektif. Oleh karena itu, diperlukan sistem yang mampu membuat keputusan yang tepat mengenai penerima KIP Kuliah berdasarkan data dari para pendaftar (Nopriandi et al., 2023).

Salah satu aspek penting pada pemodelan *machine learning* adalah data yang digunakan untuk melatih model, terutama dalam pemodelan berbasis teks seperti analisis sentimen. Dalam analisis sentimen, proses pelatihan sering kali lebih menantang dibandingkan dengan area *machine learning* lainnya (Ferdiana et al., 2019). Data yang dihasilkan tidak selalu siap untuk diproses (Syah & Witanti, 2022). Hal ini disebabkan oleh sifat data yang subjektif, seperti opini, yang tidak memiliki nilai konkret. Selain itu, data tersebut berasal dari manusia, dan setiap individu memiliki cara yang unik untuk mengekspresikan pendapat mereka (Amrullah et al., 2020).

Metode penelitian ini menggunakan perbandingan dua metode yakni "*support vector machine* (SVM) dan *naïve bayes classifier* (NBC)" (Rahat et al., 2020). "SVM adalah metode *machine learning* yang dapat digunakan untuk melakukan prediksi." SVM juga cocok dipakai pada peramalan deret waktu karena memiliki fungsi kernel yang mampu menangani masalah *non-linear* (Lumbanraja et al., 2019). Sedangkan metode NBC membuat asumsi yang sangat kuat mengenai independensi dari setiap kelas kejadian yang diberi label. NBC ini digunakan untuk mengklasifikasikan sentimen dari data yang sudah dikumpulkan (Cindo et al., 2019). *Naive bayes* adalah algoritma *machine learning* yang didasarkan pada *teorema bayes* dan mengasumsikan setiap fitur dataset berkontribusi secara independen terhadap hasil atau label. Meskipun asumsi ini tidak selalu tepat, NBC tetap populer dan efektif untuk klasifikasi, terutama dalam analisis teks seperti analisis sentimen, karena kesederhanaan dan efisiensinya (Apriani & Gustian, 2019).

Terdapat penelitian sebelumnya yang membahas mengenai komparasi antara SVM dan NBC dengan "klasifikasi berbasis *particle swarm optimization* pada analisis sentimen ekspedisi barang" (Sharazita Dyah Anggita & Ikmah, 2020). Hasil dari penelitian yang di lakukan adalah untuk nilai akurasi pada metode NBC meningkat sebesar 15,11% menjadi 80,81% setelah dioptimalkan dengan PSO. Sedangkan nilai akurasi SVM sebesar 80.3% mengalami peningkatan 1,74% dibanding SVM klasik yang tidak dioptimalkan memakai PSO. Kesimpulan dari penelitian yang di lakukan menggunakan metode NBC lebih baik dibandingkan dengan metode SVM.

Penelitian dilakukan untuk mencari hasil perbandingan metode yang digunakan. Data penelitian didapat dari hasil *crawling* pada media sosial X dengan memanfaatkan kode token API, *tools google collaboratory* dan bahasa pemrograman *pyhton*. Hasil penelitian diperoleh metode *naïve bayes* lebih baik dibanding *support vector machine* yang mana hasil akurasi *naïve bayes* 88.24%, sedangkan *support vector machine* 78.77% dengan jumlah data yang sama (Zamachsari et al., 2020).

Penelitian oleh Dias Saputri et al. (2020) menggunakan pengumpulan data opini pengguna melalui *google play store* dengan jumlah 2.000 data yang terbagi menjadi sentimen negatif dan positif. Data *e-wallet* yang dipakai untuk penelitian yaitu DANA dan OVO. Hasil data tersebut untuk metode *naïve bayes* lebih unggul dengan nilai 94.90% sedangkan *support vector machine* hanya mampu 90.00%. Akan tetapi, jika pada kurva ROC, AUC untuk algoritma SVM lebih tinggi sebesar 0.986%, sedangkan *naïve bayes* 0.778%. Namun kedua metode tersebut menunjukan hasil yang baik untuk *e-wallet* OVO dibandingkan DANA.

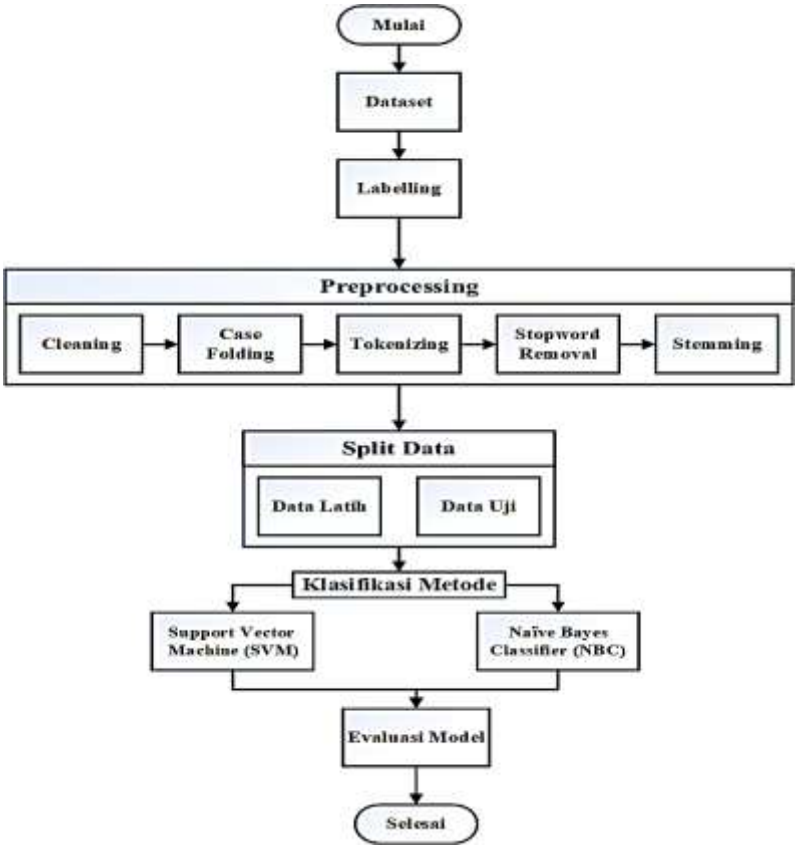
Penelitian bertujuan menentukan metode menganalisis data terkait permasalahan pada serangan *ransomware* pada KIP-K di tiktok, dengan data yang diperoleh melalui *scraping* di

media sosial X. Tujuan utama penelitian ini adalah mengetahui menganalisis persentase komentar dan respons pengguna media sosial X mengenai permasalahan *ransomware* KIP-K, yang dapat menjadi masukan untuk meningkatkan kualitas perusahaan. Penelitian ini diharapkan dapat membantu pengelola data KIP-K meningkatkan keamanan sistem, sehingga dapat mengurangi kekhawatiran masyarakat terhadap serangan *ransomware* dan meningkatkan kepercayaan publik terhadap keamanan data.

METODE PENELITIAN

Rancangan Penelitian

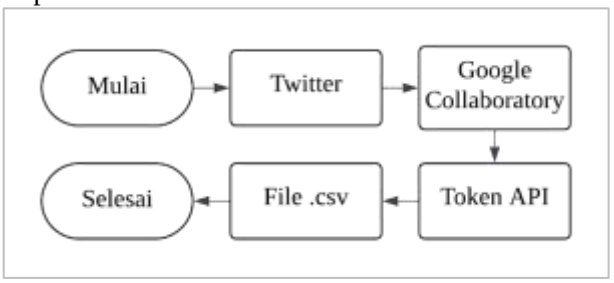
Rancangan penelitian bertujuan membandingkan kinerja antara metode SVM dan NBC. Alur penelitian terkait studi kasus serangan *ransomware* pada data KIP-K dilihat pada Gambar 1



Gambar 1. Kerangka Penelitian

Dataset

Proses *crawling* atau pengumpulan dataset ini menggunakan data *internal* dari media sosial X atau internet untuk menentukan apakah sebuah kalimat termasuk dalam opini positif, netral, atau negatif. Proses ini dikenal sebagai klasifikasi, yang menempatkan dokumen dalam kategori positif, netral, atau negatif. Selain itu, data juga dikumpulkan melalui tinjauan literatur yang menggunakan data sekunder dari media sosial X atau internet (Darwis et al., 2021). Tahapan *crawling* dilihat pada Gambar 2.



Gambar 2. Alur *Crawling* Dataset

Crawling dataset dilakukan dengan menggunakan bahasa pemrograman *pyhton* yang terbaru dengan versi *tweet-harvest@2.6.1*. Pengumpulan data menggunakan *tools google collabboratory* dengan memanfaatkan kode token API media sosial X yang diperlukan untuk proses *crawling* sebagai kunci pengumpulan data yang akan dicari yang dimana token tersebut diperoleh pada setiap akun media sosial X penggunanya (Herlinda et al., 2021).

Labelling

Labelling atau pelabelan adalah proses pemberian label atau kategori pada data mentah untuk mendukung pelatihan model *machine learning*, dengan pembagian seperti sentimen positif, negatif, dan netral (Romadoni et al., 2020). Setiap data dalam dataset diberi tag sesuai dengan karakteristik atau kelas yang relevan, memungkinkan model untuk belajar mengenali pola dan membuat prediksi berdasarkan data tersebut.

Preprocessing

Preprocessing adalah proses mengubah data mentah ke dalam format yang sesuai untuk penambangan data, dan merupakan fase paling krusial dalam proses penambangan data. Pada tahap ini, data yang masih kotor akan dibersihkan (Dirjen et al., 2020). Proses pembersihan dalam penelitian meliputi beberapa langkah:

1. ***Cleaning,*** Proses *cleaning* meliputi mengubah teks menjadi huruf kecil (*case folding*), menghapus karakter non-huruf, menghilangkan username atau mentions (@), menghapus hashtag (#), serta menghapus URL atau link dari setiap komentar.
2. ***Case Folding,*** *"Case folding* adalah proses mengubah setiap kata dalam dataset menjadi huruf kecil menggunakan fungsi *lowercase*."
3. ***Tokenizing,*** *Tokenisasi* adalah proses memecah teks dalam sebuah kalimat menjadi kata-kata yang terpisah. Sehingga setiap kata akan memiliki nilai dan makna tersendiri yang akan membantu pencarian hasil nilai akurasi.
4. ***Stopword Removal,*** *Stopword* berfungsi untuk menghilangkan kata-kata yang tidak penting, seperti "yang", "di" dll. Tahapan ini sangat berpengaruh terhadap klasifikasi metode yang akan di gunakan.
5. ***Stemming,*** Tahap terakhir adalah *stemming*, yaitu proses menghilangkan *prefiks* dan *sufiks* untuk mengubah kata menjadi bentuk dasarnya.

Split Data

Sebelum menerapkan algoritma, kita perlu membagi dataset menjadi data *training* dan data *testing*. Data pelatihan untuk memberi instruksi kepada algoritma dalam mengenali data yang diklasifikasikan sebagai positif atau negatif. Setelah pelatihan selesai, data pengujian untuk mengevaluasi model yang diperoleh dari data pelatihan. Studi ini berupaya menentukan akurasi yang optimal. Kumpulan data dibagi 70% data pelatihan dan 30% data pengujian.

Klasifikasi Metode

Klasifikasi merupakan contoh penambangan data prediktif. Dalam penambangan data prediktif, terdapat prosedur yang dikenal sebagai pemisahan data. Pemisahan data melibatkan pemisahan data menjadi dua segmen: data pelatihan untuk pelatihan sistem dan data pengujian untuk evaluasi sistem. Algoritma NBC dan SVM yang digunakan dalam studi ini termasuk dalam pembelajaran terbimbing, yang menunjukkan bahwa data yang digunakan untuk pelatihan model telah diberi label. Dalam analisis ini, label yang digunakan adalah sentimen yaitu positif dan negatif.

Naïve Bayes Classifier* (NBC*)

Algoritma NBC merupakan teknik yang menggunakan probabilitas dan statistik untuk mengatasi masalah klasifikasi. Metode ini melakukan klasifikasi dengan menghitung probabilitas kondisional P(X|Y) dari kemungkinan kelas X. Penentuan kelas dicapai dengan memilih nilai P(X|Y) terbesar (Febtadianrano Putro et al., 2020). Keunggulan klasifikasi ini

terletak pada persyaratannya untuk data pelatihan minimal guna memperkirakan parameter yang diperlukan. Persamaan selanjutnya adalah algoritma NBC untuk menentukan nilai P(X|Y):

$$P\ (X|Y) = \frac{\text{P (Y|X).(X)}}{\text{P(Y)}} \quad (1)$$

Pada fungsi rumus di atas dapat di jelaskan mulai dari "fungsi P(X|Y) merupakan posterior | *probability* yaitu nilai X berdasarkan kondisi Y. Untuk P(Y|X) merupakan probabilitas Y yang ditentukan X adalah benar, sedangkan P(X) adalah peluang *evidence* penyakit X, dan terakhir P(Y) probabilitas dari nilai Y".

***Support Vector Machine* (SVM)**

SVM merupakan teknik klasifikasi yang menggunakan pembelajaran terbimbing untuk meramalkan kategori berdasarkan pola yang diperoleh dari fase pelatihan. "Klasifikasi dengan hyperplane yang membedakan antara kelas positif dan negatif. Hyperplane yang ideal memaksimalkan jarak ke titik data pelatihan terdekat dari setiap kelas, karena margin yang lebih lebar biasanya mengurangi kesalahan generalisasi." Pemisahan yang optimal dicapai dengan menilai margin hyperplane dan mengidentifikasi titik terbesarnya (Apriyani & Kurniati, 2020).

Berikut fungsi kerja perhitungan dari metode SVM penelitian ini untuk menentukan suatu klasifikasi ataupun prediksi.

$$f(X_d) = \sum_{i=1}^{ns} a_{i,} y_i, x_i, x_d \ + b \quad (2)$$

Fungsi persamaan dari rumus yang telah dijelaskan "dimn *ns*=jumlah *support vector*, *ai*=nilai bobot setiap titik data, *yi*=kelas data *xi*=variabel *support vector*, *xd*=data yang akan dilkasifikasikan, dan *b*=nilai eror atau bias".

Evaluasi Model

Tahap evaluasi bertujuan untuk menilai kinerja model algoritma yang diterapkan. Evaluasi model dilakukan menggunakan matriks kebingungan, yaitu tabel yang menyajikan data perbandingan antara hasil klasifikasi sistem (prediksi) dan hasil klasifikasi aktual. Matriks kebingungan menampilkan jumlah data uji yang diklasifikasikan secara akurat di samping jumlah yang diklasifikasikan secara tidak akurat. Dengan memanfaatkan hasil matriks kebingungan ini, kita dapat menghitung nilai akurasi, presisi, perolehan kembali, dan skor F1 menggunakan persamaan berikut:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Presisi = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP+TN}{TP+FN} \quad (5)$$

$$f1 - score = \frac{2 \times Presisi \times Recall}{Presisi \times Recall} \quad (6)$$

"Pada nilai *true* positif (TP) dan *true* negatif (TN), klasifikasi dilakukan dengan benar. *False* positif (FP) adalah kasus di mana prediksi menunjukkan hasil positif padahal sebenarnya negatif, sedangkan *false* negatif (FN) adalah situasi di mana prediksi menunjukkan hasil negatif padahal seharusnya positif" (Handayanto et al., 2021). Berikut ialah tabel *confusion matrix*:

Tabel 1. *Confusion Matrix*

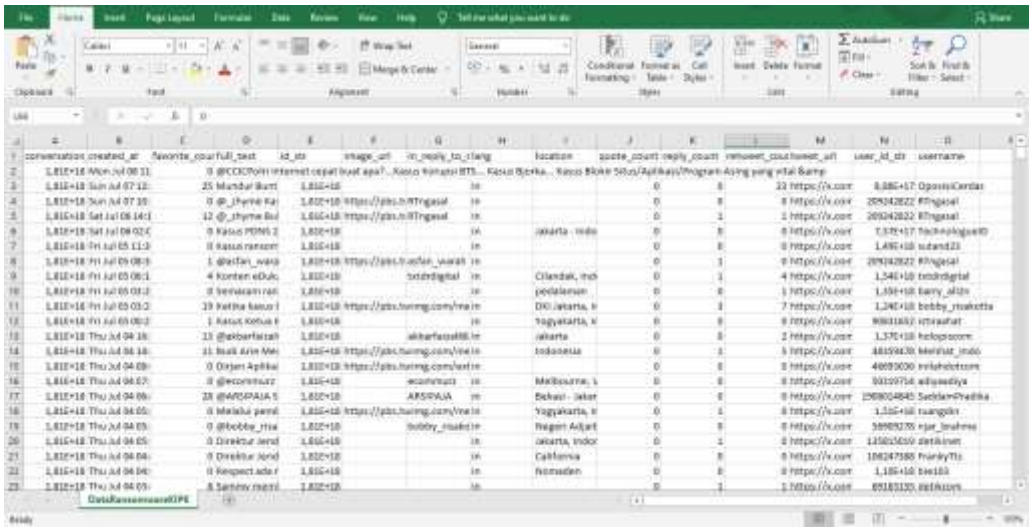
Kelas Prediksi	Kelas Aktual	
	Sentimen Positif	Sentimen Negatif
Sentimen Positif	TP	FN
Sentimen Negatif	FP	TN

HASIL DAN PEMBAHASAN

Pengumpulan Dataset

Pengambilan data untuk analisis sentimen diperoleh dengan cara *scraping*/*crawling* dan menggunakan bahasa pemrograman *python* melalui token API media sosial X. Setiap satu kode token dapat mengumpukan data sebanyak 1.500 setiap harinya. Dataset yang telah di *crawling*/*scraping* menggunakan token API tersebut terkumpul sebanyak 2.648 ,data tersebut masih berupa data mentah yang harus melalui tahap *filtering* dan *preprocessing* terlebih dahulu

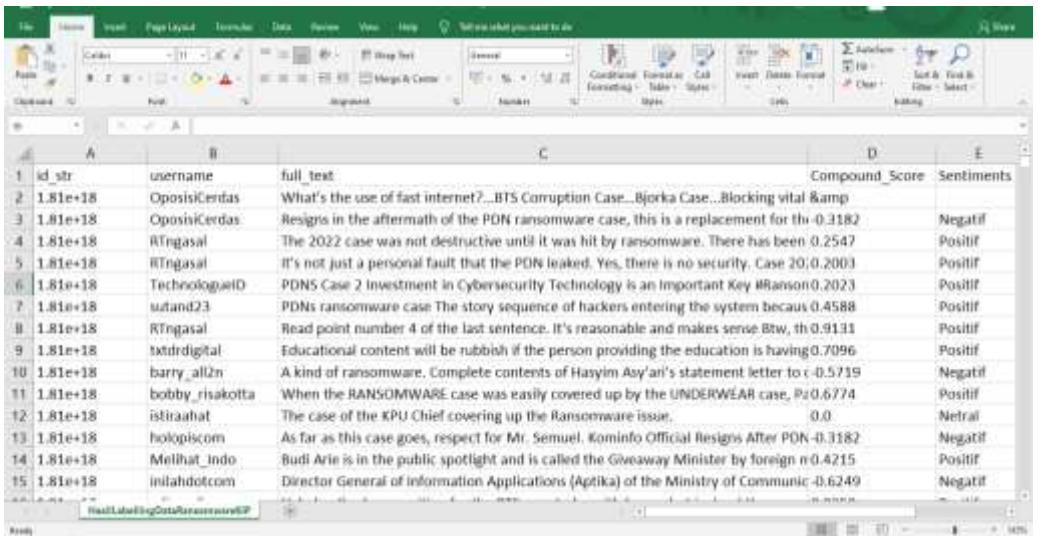
agar data yang dipakai untuk analisis sentimen dapat lebih akurat. Berikut contoh hasil *crawling* dataset pada Gambar 3.



Gambar 3. Hasil *Crawling*

Labelling Dataset

Prosedur ini untuk memastikan kategori emosi yang digunakan untuk menghitung metrik akurasi dan memvisualisasikan data tweet. Sebelum dimulainya proses pelabelan, data harus diterjemahkan ke dalam bahasa Inggris melalui analisis sentimen VADER untuk hasil yang optimal. Hasil pelabelan diklasifikasikan ke dalam tiga kategori: "emosi positif, negatif, dan netral". Meskipun demikian, sentimen netral diabaikan karena dianggap tidak penting atau remeh. hasil proses *labelling* pada Gambar 4.



	A	B	C	D	E
1	id_str	username	full_text	Compound_Score	Sentiments
2	1.81e+18	OposisiCerdas	What's the use of fast internet?...BTS Corruption Case...Bjorka Case...Blocking vital &amp		
3	1.81e+18	OposisiCerdas	Resigns in the aftermath of the PDN ransomware case, this is a replacement for th	-0.3182	Negatif
4	1.81e+18	RTngasal	The 2022 case was not destructive until it was hit by ransomware. There has been	0.2547	Positif
5	1.81e+18	RTngasal	It's not just a personal fault that the PDN leaked. Yes, there is no security. Case 20	0.2003	Positif
6	1.81e+18	TechnologueID	PDNS Case 2 Investment in Cybersecurity Technology is an Important Key #Ranson	0.2023	Positif
7	1.81e+18	sutand23	PDNs ransomware case The story sequence of hackers entering the system becaus	0.4588	Positif
8	1.81e+18	RTngasal	Read point number 4 of the last sentence. It's reasonable and makes sense Btw, th	0.9131	Positif
9	1.81e+18	txtdrdigital	Educational content will be rubbish if the person providing the education is having	0.7096	Positif
10	1.81e+18	barry_all2n	A kind of ransomware. Complete contents of Hasyim Asy'ari's statement letter to c	-0.5719	Negatif
11	1.81e+18	bobby_risakotta	When the RANSOMWARE case was easily covered up by the UNDERWEAR case, P	0.6774	Positif
12	1.81e+18	istiraahat	The case of the KPU Chief covering up the Ransomware issue.	0.0	Netral
13	1.81e+18	holopiscom	As far as this case goes, respect for Mr. Semuel. Kominfo Official Resigns After PDN	-0.3182	Negatif
14	1.81e+18	Melihat_Indo	Budi Arie is in the public spotlight and is called the Giveaway Minister by foreign n	0.4215	Positif
15	1.81e+18	initahdotcom	Director General of Information Applications (Aptika) of the Ministry of Communic	-0.6249	Negatif

Gambar 4. Hasil *Labelling*

***Preprocessing* Dataset**

1. *Cleaning*

Tahap *cleaning* dilakukan untuk membersihkan dokumen dari isi data yang tidak penting atau tidak berguna untuk tahap pengujian berikutnya. Berikut ini merupakan hasil tahap *cleaning* Tabel 2.

Tabel 2. Hasil *Cleaning*

Tweet	Cleaning
Resigns in the aftermath of the PDN ransomware case, this is a replacement for the subordinate of the Minister of Communication and Information, Budi Arie	Resigns in the aftermath of the PDN ransomware case this is replacement for the subordinate of the Minister of Communication and Information Budi Arie

2. *Case Folding*

Tahapan ini berfokus dalam transformasi kalimat yang mana huruf besar dirubah menjadi huruf kecil. Untuk memastikan keseragaman dalam format surat. Hasil dari prosedur pelipatan kasus disajikan pada Tabel 3 di bawah ini.

Tabel 3. Hasil *Case Folding*

Tweet	Case Folding
Resigns in the aftermath of the PDN ransomware case, this is a replacement for the subordinate of the Minister of Communication and Information, Budi Arie	resigns in the aftermath of the pdn ransomware case this is replacement for the subordinate of the minister of communication and information budi arie

3. *Tokenizing*

Tokenisasi adalah proses di mana sebuah kalimat dipecah menjadi beberapa kata yang terpisah. Tahap ini bertujuan mengelompokkan teks menjadi kata-kata yang berbeda dengan menghilangkan karakter pembatas seperti titik, koma, dan spasi. Hasil proses *tokenizing* dilihat pada Tabel 4 berikut.

Tabel 4. Hasil *Tokenizing*

Tweet	Tokenizing
Resigns in the aftermath of the PDN ransomware case, this is a replacement for the subordinate of the Minister of Communication and Information, Budi Arie	['resigns', 'in', 'the', 'aftermath', 'of', 'the', 'pdn', 'ransomware', 'case', 'this', 'is', 'replacement', 'for', 'the', 'subordinate', 'of', 'the', 'minister', 'of', 'communication', 'and', 'information', 'budi', 'arie']

4. *Stopword Removal*

Metode ini menghilangkan kata-kata atau kalimat yang berlebihan atau tidak berarti dari data. Hasil proses *stopword removal* dapat dilihat pada Tabel 5 berikut.

Tabel 5. Hasil *Stopword Removal*

Tweet	Stopword
Resigns in the aftermath of the PDN ransomware case, this is a replacement for the subordinate of the Minister of Communication and Information, Budi Arie	['resigns', 'in', 'the', 'aftermath', 'of', 'the', 'pdn', 'ransomware', 'case', 'this', 'is', 'replacement', 'for', 'the', 'subordinate', 'of', 'the', 'minister', 'of', 'communication', 'and', 'information', 'budi', 'arie']

5. *Stemming*

Tahap ini berguna untuk menghilangkan imbuhan pada kata atau kata dasar. Dilihat pada Tabel 6.

Tabel 6. Hasil *Stemming*

Tweet	Stemming
Resigns in the aftermath of the PDN ransomware case, this is a replacement for the subordinate of the Minister of Communication and Information, Budi Arie	resigns in the aftermath of the pdn ransomware case this is replacement for the subordinate of the mister of communication and information budi arie

Evaluasi dan Klasifikasi Metode

***Naïve Bayes Classifier* (NBC)**

Setelah melakukan tahap pemrosesan data selanjutnya penelitian ini mengeluarkan hasil implementasi dengan metode naïve bayes. "Dengan jumlah data yang di uji sebanyak 681 data yang menghasilkan nilai akurasi, *precision*, *recall*, *support*, dan *f1-score"* yang terlihat pada Gambar 5 berikut dengan klasifikasi menggunakan fungsi *confusion matrix.*

Laporan Klasifikasi :

	precision	recall	f1-score	support
Negatif	0.73	0.75	0.74	383
Positif	0.67	0.65	0.66	298
accuracy			0.70	681
macro avg	0.70	0.70	0.70	681
weighted avg	0.70	0.70	0.70	681

Gambar 5. Hasil Klasifikasi *Naïve Bayes Classifier*

Dari hasil perhitungan diatas dapat diketahui nilai akurasi yang diperoleh dengan data *ransomware* KIP-K sebesar 70% dengan *support* 681 data. Hasi "nilai *precision* negatif 73%, *precision* positif 67%, nilai *recall* negatif 75%, *recall* positif 65%, nilai *f1-score* negatif 74%, *f1-score* positif 66%". Hasil tersebut dapat dikatakan baik dalam pencarian analisis sentimen dengan jumlah data yang cukup besar, karena nilai rata-rata akurasi yang baik diatas 50%.

***Support Vector Machine* (SVM)**

Klasifikasi dilakukan dengan data latih dan data uji yang sama yakni data uji sebesar 30% dari keseluruhan data setelah preprocessing sebanyak 1.738. Hasil perhitungan dengan metode SVM dengan topik *ransomware* KIP-K dan data yang serupa dengan fungsi *confusion matrix* dilihat pada Gambar 6.

	precision	recall	f1-score	support
Negatif	0.88	0.91	0.90	392
Positif	0.88	0.83	0.85	289
accuracy			0.88	681
macro avg	0.88	0.87	0.87	681
weighted avg	0.88	0.88	0.88	681

Gambar 6. Hasil Klasifikasi *Support Vector Machine*

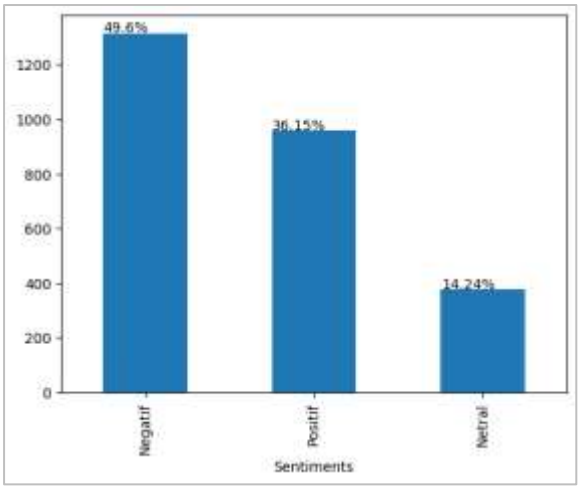
Berdasarkan hasil perhitungan diatas nilai akurasi pada topik *ransomware* KIP-K dengan metode SVM mempunyai nilai lebih besar dibanding metode NBC dengan jumlah dan data yang sama. Nilai akurasi SVM sebesar 88% dengan *precision* negatif 88%, *precision* positif 88%, *recall* negatif 91%, *recall* positif 83%, *f1-score* negatif 90%, *f1-score* positif 85% dengan *support* 681.

Visualisasi

Pada tahap visualisasi akan menampilkan hasil dari *histogram*, *wordcloud*, visualisasi sentimen kata negatif dan positif. Berikut merupakan visualisasi dari metode NBC dan SVM.

Histogram

Analisis sentimen dengan menggunakan pemrograman *python* menampilkan "hasil histogram sentimen positif , negatif, dan netral dari dataset yang sudah melalui proses pelabelan menggunakan *vader sentimen"*. Hasil histogram mendapatkan 49.6% data negatif, 36.15% data positif, dan 14.24% data netral dengan jumlah komentar yang telah melalui proses *preprocesing* sebesar 1.738 data. Berikut merupakan histogram kelas sentimen pada Gambar 7.



Gambar 7. Histogram Kelas Sentimen

Wordcloud

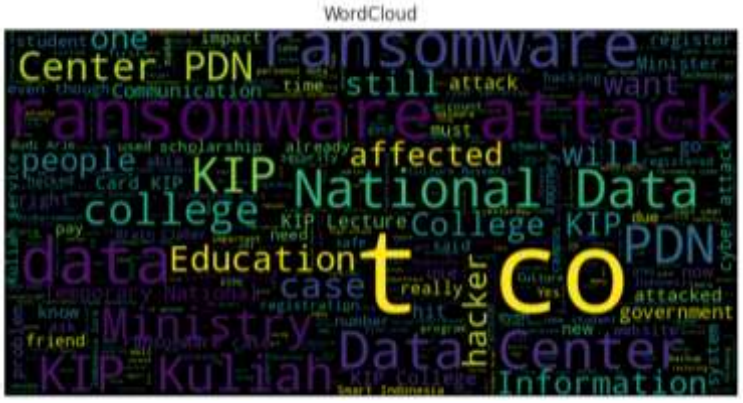
Tahapan *wordcloud* menampilkan hasil berupa kata-kata yang di analisis sentimen berupa bentuk pengeluaran hasil kata negatif dan positif. "Hasil *wordcloud* ini sama dengan histogram memakai bahasa pemrograman *python* dengan bantuan tools *google colab* sebagai pencariannya." Hasil *wordcloud* pada Gambar 8.

Gambar 8. *Wordcloud* Visualisasi Kata Positif

Hasil visualisasi menunjukkan kata-kata positif yang sering muncul seperti *ransomware*, kip kuliah, *ministry*, data, *center, education, information* dan sebagainya. Visualisasi ini diperoleh melalui proses pemrograman *python* yang kompleks, sehingga menghasilkan data yang lebih akurat.. Untuk hasil visulisasi kata negatif dapat dilihat pada Gambar 9.

Gambar 9. *Wordcloud* Visualisasi Kata Negatif

Dari hasil proses *wordcloud* selain menampilkan visualisasi kelas negatif dan positif, proses ini menghasilkan visualisasi dari *worcloud* keseluruhan data yang mana setiap kata pada data tersebut akan tervisualisasikan, hasil proses *wordcloud* keseluruhan data dilihat pada Gambar 10.



Gambar 10. Hasi *Worcloud Dataset*

Dari hasil *wordcloud* pada kasus *ransomware* data kip-k dapat dilihat bahwa *frequensi* kata yang sering muncul ditandai besar dan kecilnya kata yang muncul, seperti "*ransomware*", "*national*", "kip", "pdn", "kuliah", "*attack*", dan lainya.

KESIMPULAN

Dari hasil penelitian bertujuan analisis sentimen terhadap serangan *ransomware* data KIP menggunakan perbandingan metode NBC dan SVM dengan data dari media sosial X menggunakan proses *crawling*. Proses *crawling* menggunakan kode token API media sosial X dan dijalankan dengan bahasa pemrograman *python*. Sebelum proses klasifikasi data akan diproses melalui tahapan *labelling, preprocessing*, dan terakhir klasifikasi metode. "Data terbagi dua yakni data *training* 70% dan data *testing* 30% yang akan digunakan untuk pengujian metode." Proses labelling data menghasilkan 49.6% data negatif, 36.15% data positif, dan 14.24% data netral. Data mentah yang telah terkumpul sebayak 2.648 kemudian masuk tahap *preprocessing* dan hasil data bersih sebanyak 1.738. Hasil proses klasifikasi dengan metode NBC memiliki akurasi sebesar 70% dengan support data 681, nilai *precision* negatif 73%, *precision* positif 67%, nilai *recall* negatif 75%, *recall* positif 65%, nilai *f1-score* negatif 74%, *f1-score* positif 66%. sedangkan klasifikasi pengujian metode SVM nilai akurasi sebesar 88% dengan support data 681 dengan *precision* negatif 88%, *precision* positif 88%, *recall* negatif 91%, *recall* positif 83%, *f1-score* negatif 90%, *f1-score* positif 85%. Setelah melakukan klasifikasi metode selanjutnya data akan ditampilkan dalam visualisasi yang terdiri dari *wordcloud* dan histogram. Sesuai pembahasan diatas ditarik kesimpulan dari penelitian ini. Maka dikatakan bahwa lebih banyak pengguna media sosial X yang memberikan respon negatif terhadap serangan

ransomware data KIP-K ini dengan harapan pemerintah dapat meningkatkan sistem keamanan data agar tidak terjadi lagi hal yang serupa dan tidak menimbulkan keresahan pada masyarakat.

SARAN

Saran bagi penelitian selanjutnya mencoba metode klasifikasi lain seperti metode Decision Tree & Random Forest, Logistic Regression, K-Nearest Neighbors (KNN) dan metode lainnya yang mungkin lebih efisien agar meningkatkan nilai akurasi analisis sentimen. Perbaikan pada tahap *preprocessing* dapat dilakukan kembali untuk penelitian selanjutnya agar hasil akurasi setiap metode dapat lebih akurat.

DAFTAR PUSTAKA

Amilia, A. A., Ans'harikhu, P., Alfan, M., Bimantara, A., Suciani, L., Yanuar, A., & Rahmawati, P. (2020). Gerakan Ayo Kuliah Program Keluarga Harapan untuk Memotivasi Siswa Melanjutkan Pendidikan ke Perguruan Tinggi. *Community Empowerment*, *5*(3), 177–185. https://doi.org/10.31603/CE.3986

Amrullah, A. Z., Sofyan Anas, A., Adrian, M., & Hidayat, J. (2020). Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square. *Jurnal Bumigora Information Technology (BITe)*, *2*(1), 40–44. https://doi.org/10.30812/BITE.V2I1.804

Anjelina, C. D. (2024). *Data 282 Layanan Kementerian/Lembaga Hilang Usai Diserang Ransomware, Ini Kata Ahli*. Kompas.Com. https://www.kompas.com/tren/read/2024/06/28/193000265/data-282-layanan-kementerian-lembaga-hilang-usai-diserang-ransomware-ini?page=all#:~:text=Editor&text=kompas.com - Sepekan berlalu sejak,negara hilang dan sulit dikembalikan.

Apriani, R., & Gustian, D. (2019). Analisis Sentimen Dengan Naïve Bayes Terhadap Komentar Aplikasi Tokopedia. *Jurnal Rekayasa Teknologi Nusa Putra*, *6*(1), 54–62. https://doi.org/10.52005/REKAYASA.V6I1.86

Cindo, M., Rini, D. P., & Ermatita, E. (2019). Literatur Review: Metode Klasifikasi Pada Sentimen Analisis. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, *1*(1), 66–70.

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131–145. https://doi.org/10.33365/JTK.V15I1.744

Dias Saputri, A., Pratama -, A. R., Pradiatiningtyas, D., Bayu Dewa, C., Ayu Safitri, L., Ajeng Kristiyanti, D., Andini Putri, D., Indrayuni, E., Nurhadi, A., & Hairul Umam, A. (2020). E-Wallet Sentiment Analysis Using Naïve Bayes and Support Vector Machine Algorithm. *Journal of Physics: Conference Series*, *1641*(1), 012079. https://doi.org/10.1088/1742-6596/1641/1/012079

Dirjen, S. K., Riset, P., Pengembangan, D., Dikti, R., Nugroho, A., Bimo Gumelar, A., Sooai, A. G., Sarvasti, D., & Tahalele, P. L. (2020). Perbandingan Performansi Algoritma Pengklasifikasian Terpandu Untuk Kasus Penyakit Kardiovaskular. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(5), 998–1006. https://doi.org/10.29207/RESTI.V4I5.2316

Fatmawati, N. (2021). *Pengaruh Positif dan Negatif Media Sosial Terhadap Masyarakat*. Kementrian Keuangan Republik Indonesia. https://www.djkn.kemenkeu.go.id/kpknl-semarang/baca-artikel/14366/Pengaruh-Positif-dan-Negatif-Media-Sosial-Terhadap-Masyarakat.html

Ferdiana, R., Jatmiko, F., Purwanti, D. D., Sekar, A., Ayu, T., & Dicka, W. F. (2019). Dataset Indonesia untuk Analisis Sentimen. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *8*(4), 334–339.

Handayanto, R. T., Herlawati, H., Atika, P. D., Khasanah, F. N., Yusuf, A. Y. P., & Septia, D. Y. (2021). Analisis Sentimen Pada Situs Google Review dengan Naïve Bayes dan Support Vector Machine. *Jurnal Komtika (Komputasi Dan Informatika)*, *5*(2), 153–163. https://doi.org/10.31603/KOMTIKA.V5I2.6280

Herlinda, V., Darwis, D., & Dartono, D. (2021). Analisis clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means. *Jurnal Teknologi Dan Sistem Informasi*, *2*(2), 94–99. https://doi.org/10.33365/JTSI.V2I2.890

Lumbanraja, F. R., Sani, R. S., Kurniawan, D., & Irawati, A. R. (2019). *Implementasi Metode Support Vector Machine Dalam Prediksi Persebaran Demam Berdarah di Kota Bandar Lampung*.

Nopriandi, H., Aprizal, A., & Chairani, S. (2023). Sistem Pendukung Keputusan Seleksi Calon Penerima Beasiswa Kartu Indonesia Pintar Kuliah (Kip-K) Di Universitas Islam Kuantan Singingi. *Jurnal Teknologi Dan Open Source*, 41–54. https://doi.org/10.36378/JTOS.V6I1.2698

Rahat, A. M., Kahir, A., & Masum, A. K. M. (2020). Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset. *Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019*, 266–270. https://doi.org/10.1109/SMART46866.2019.9117512

Romadoni, F., Umaidah, Y., Nurina Sari, B., & Ilmu Komputer, F. (2020). Text Mining Untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma Support Vector Machine. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *9*(2), 247–253. https://doi.org/10.32736/SISFOKOM.V9I2.903

Sharazita Dyah Anggita, & Ikmah. (2020). Algorithm Comparation of Naive Bayes and Support Vector Machine based on Particle Swarm Optimization in Sentiment Analysis of Freight Forwarding Services. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(2), 362–369. https://doi.org/10.29207/resti.v4i2.1840

Syah, H., & Witanti, A. (2022). Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Pada Media Sosial Twetter Menggunakan Algoritma Support Vector Machine (SVM). *Jurnal Sistem Informasi Dan Informatika (Simika)*, *5*(1), 59–67. https://doi.org/10.47080/simika.v5i1.1411

Winston Talakua, M., Tomasouw, B. P., Ilwaru, V. Y. I., Talakua, M. W., & Tomasouw, B. P. (2023). Design Of Kip Kuliah Selection System And Recipient Determination Using Support Vector Machine (SVM). *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *17*(3), 1803–1814. https://doi.org/10.30598/BAREKENGVOL17ISS3PP1803-1814

Zamachsari, F., Vangeran Saragih, G., Susafa'ati, & Gata, W. (2020). Analysis of Sentiment of Moving a National Capital with Feature Selection Naive Bayes Algorithm and Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *4*(3), 504–512. https://doi.org/10.29207/RESTI.V4I3.1942