

# Klasifikasi Opini Pengguna Twitter Terhadap Sekolah Daring dengan Metode Naïve Bayes dan Support Vector Machine

Muhamad Khafidhun Alim Muslim<sup>1</sup>, Yuni Yamasari<sup>2</sup>

<sup>1,2</sup> Jurusan Teknik Informatika, Universitas Negeri Surabaya

<sup>1</sup>[muhamad.17051204063@mhs.unesa.ac.id](mailto:muhamad.17051204063@mhs.unesa.ac.id)

<sup>2</sup>[yuniyamasari@unesa.ac.id](mailto:yuniyamasari@unesa.ac.id)

**Abstrak**— Covid-19 atau *Coronavirus Disease* telah melanda negara Indonesia. Hal ini menyebabkan segala aktifitas terganggu, termasuk dalam dunia pendidikan. Sistem pembelajaran yang semula dilakukan tatap muka menjadi secara daring atau online. Sistem pembelajaran ini mendapatkan banyak tanggapan masyarakat. Fasilitas dan cara belajar mengajar menjadi perdebatan dimasyarakat. Hal ini menjadi landasan untuk melakukan penelitian ini. Penelitian ini memfokuskan pada analisa sentimen tentang sekolah daring dengan metode SVM dan NBC. Terkait dengan pengumpulan data, penelitian ini berasal dari 414 tweet yang terbagi dengan 3 sentimen, secara berturut-turut, yaitu positif, negatif, netral sebanyak 27.7%, 36.9%, 35.2%. Teknik evaluasi menggunakan 2 teknik yaitu split data dan *cross-validation*. Hasil uji menggunakan split data pada metode NBC mendapatkan nilai akurasi tertinggi yaitu pada rasio 4:6 dengan nilai sebesar 68%, untuk nilai presisi 74%, recall 65%, dan f1-score 68%. Sedangkan, penerapan metode SVM dengan kernel=linier mendapatkan kinerja tertinggi yaitu pada rasio 1:9 yaitu level akurasi sebesar 71.4%, presisi 73%, recall 70%, dan f1-score 71%. Untuk hasil uji dengan menggunakan *cross-validation* pada metode NBC, kinerja tertinggi yang dicapai pada k=10 dengan level akurasi sebesar 69.25%, presisi 86%, recall 79%, dan f1-score 81%. Sedangkan pada metode SVM dengan kernel linier, kinerja tertinggi dihasilkan pada k=10 dengan nilai akurasi sebesar 68.87%, presisi 74%, recall 68%, dan f1-score 69%.

**Kata Kunci** : Sekolah daring, *Twitter*, *Naïve Bayes Classifier*, *Support Vector Machine*, *k-Fold Cross Validation*.

## I. PENDAHULUAN

Pada bulan Maret 2020, Presiden Joko Widodo membuat pengumuman bahwa virus Covid-19 yang berasal dari Wuhan menjangkit dua warga Indonesia [1]. *Covid-19* atau *Coronavirus Disease* merupakan virus yang menyerang manusia terutama pada bagian saluran pernapasan.

Setelah awal kasus *Covid-19* ini diumumkan, kasus penyebaran di Indonesia meningkat secara drastis, dilansir dari data BNPB tanggal 7 februari 2021 kasus aktif di Indonesia mencapai 1.157.837[2]. Akhirnya pemerintah mengeluarkan keputusan yaitu menerapkan PSBB, keputusan ini dibuat untuk menekan penyebaran Covid-19 di Indonesia. Presiden Joko Widodo juga memberi instruksi agar seluruh aktifitas mulai dari perkantoran, peribadatan, dan seluruh aktifitas di luar ruangan dilakukan secara *Work From Home* (WFH)[3].

Merebaknya kasus *Covid-19* di Indonesia memberikan dampak tersendiri khususnya dalam dunia Pendidikan. Penyebaran *Covid-19* yang sangat cepat dan masif memberikan kekhawatiran tersendiri baik bagi pelajar dan juga orang tua, mereka takut anak mereka dapat tertular *Covid-19*. Hal ini menyebabkan Nadiem Makarim sebagai Mendikbud memberikan instruksi yaitu mewajibkan lembaga pendidikan untuk melakukan sistem belajar mengajar secara online atau daring[4].

Sistem Pendidikan daring adalah sebuah pembelajaran yang menggunakan sarana media untuk melakukan kegiatan belajar mengajar. Pelaku pendidikan dapat melakukan aktivitas belajar mengajar dengan memanfaatkan media yang ada seperti *Google Meet*, *Zoom*, dan lain-lain. Dibalik kegiatan belajar mengajar secara daring atau online ini masih terdapat beberapa masalah yaitu orang tua siswa yang tidak memiliki handphone, ketersediaan kuota internet yang membutuhkan biaya cukup tinggi, dan lain-lain. Hal ini menjadi masalah yang besar apabila tidak diselesaikan dengan cepat dan tepat karena kegiatan belajar mengajar dapat terganggu [5].

Saiful Mujani *Research and Consulting* (SMRC) telah melakukan survey tentang punya tidaknya akses internet di masyarakat. Hasil survey memperlihatkan 76% masyarakat mempunyai dan 24% masyarakat belum mempunyai akses internet. Dalam survey SMRC terdapat kesenjangan akses internet dalam hal wilayah, antara lain untuk wilayah DKI Jakarta dan Banten 96% memiliki akses internet, wilayah Jawa Barat sebesar 71%, wilayah Jawa Tengah serta DIY sebesar 72%, wilayah Jawa Timur 74%, wilayah Bali serta Nusa Tenggara sebesar 77%, wilayah Sumatera sebesar 76%, wilayah Sulawesi sebesar 71%, wilayah Maluku dan Papua sebesar 77%, wilayah Kalimantan sebesar 79%. Selain itu terdapat kesenjangan bilamana dilihat dari latar belakang pendidikan, antara lain terdapat 50% warga berpendidikan SD yang memiliki akses internet, 69% warga berpendidikan SMP yang memiliki akses internet, 88% warga berpendidikan SMA yang memiliki akses internet, dan 96% warga berpendidikan Perguruan Tinggi yang memiliki akses internet[6].

Berdasarkan data diatas, kondisi ini memunculkan berbagai reaksi dari kalangan masyarakat di media sosial terutama *Twitter*. Terdapat masyarakat yang mendukung maupun keberatan dengan kebijakan perihal sekolah daring. Komentar yang berupa teks ini terdapat pengetahuan yang dapat diekstraksi dengan menggunakan metode *text mining*. Proses yang dilakukan ini menggunakan *sentiment analysis*.

Menurut Liu, *sentiment analysis* adalah studi pembelajaran yang melakukan analisa terhadap opini tentang peristiwa yang sedang terjadi. Ada banyak istilah tentang analisa sentimen, antara lain penambangan opini, ekstraksi opini, analisis subjektivitas, analisis emosi. Analisis sentimen ini memfokuskan terhadap pendapat yang mengarah pada sentimen positif atau negatif [7].

Analisis sentimen dapat melakukan berbagai hal seperti memprediksi hal yang kemungkinan terjadi, melakukan analisis terhadap suasana publik, menggambarkan suasana hati seseorang atau konsumen terhadap masalah yang terjadi hal ini dikarenakan memfokuskan pada analisa dari teks yang ditulis[8]. Analisis ini diterapkan untuk mengetahui pandangan masyarakat mengenai permasalahan sekolah daring. Proses ini penting dilakukan dikarenakan dapat menjadi evaluasi untuk Mendikbud dalam menyelesaikan masalah sekolah daring.

*Naïve Bayes Classifier* adalah sebuah algoritma yang memiliki asumsi bahwa setiap kata terjadi secara sendiri. Metode ini sangat berguna dikarenakan beberapa alasan antara lain sangat mudah untuk dibangun, mudah untuk diinterpretasikan, dan tidak memerlukan skema estimasi parameter iteratif yang rumit.

SVM atau *Support Vector Machine* merupakan teknik pembelajaran yang memiliki tingkat akurasi tertinggi, SVM merupakan algoritma paling populer dibandingkan semua algoritma yang digunakan pada data mining [9].

Metode NBC atau *Naïve Bayes Classifier* dipilih dikarenakan metode NBC atau *Naïve Bayes Classifier* merupakan metode klasifikasi yang sederhana dan merupakan metode klasifikasi yang sering dalam analisis sentimen. Sedangkan metode SVM berfungsi untuk membandingkan tingkat akurasi.

Beberapa penelitian telah dilakukan sebelumnya dalam melakukan perbandingan antara metode *Naïve Bayes* dan *Support Vector Machine*. Pertama yaitu penelitian yang dilakukan oleh [10] untuk menganalisa dengan judul "*An algorithm and method for sentiment analysis using the text and emoticon*". Hasilnya adalah untuk pengujian dengan *text dan emoticon* menggunakan metode *Naïve Bayes* nilai akurasi yaitu 52%, recall sebesar 57%, precision sebesar 59%, dan F1-Score sebesar 60%, sedangkan untuk pengujian dengan *only text* menggunakan metode *Naïve Bayes* nilai akurasi yaitu 64%, recall sebesar 59%, precision sebesar 59%, dan F1-Score sebesar 77%. Untuk pengujian dengan *text dan emoticon* menggunakan metode *Support Vector Machine* yaitu nilai akurasi yaitu 78%, recall sebesar 71%, precision sebesar 73%, dan F1-Score sebesar 74%, sedangkan untuk pengujian dengan *only text* menggunakan metode *Naïve Bayes* nilai akurasi yaitu 78%, recall sebesar 69%, precision sebesar 74%, dan F1-Score sebesar 71%.

Kedua, penelitian yang dilakukan oleh [11] untuk melakukan membandingkan tingkat performansi metode *Support Vector Machine* dan *Naïve Bayes* dalam mengklasifikasi jalur perminatan SMA. Hasilnya adalah untuk hasil pengujian dataset penjurusan ABC didapatkan hasil seperti berikut, metode SVM mendapatkan nilai akurasi

sebesar 96.88%, nilai presisi sebesar 98.62%, nilai recall sebesar 94.95%, dan nilai AUC sebesar (0-1) 0.996. Sedangkan metode *Naïve Bayes* mendapatkan nilai akurasi sebesar 88.87%, nilai presisi sebesar 89.63%, nilai recall sebesar 87.69%, dan nilai AUC sebesar (0-1) 0.964. Untuk hasil pengujian dataset penjurusan XYZ didapatkan hasil seperti berikut, metode SVM mendapatkan nilai akurasi sebesar 97.14%, nilai presisi 99.44%, nilai recall 95.88%, serta nilai AUC (0-1) 0.999. Sedangkan metode *Naïve Bayes* mendapatkan nilai akurasi sebesar 92.86%, nilai presisi sebesar 93.03%, nilai recall sebesar 95.85%, dan nilai AUC sebesar (0-1) 0.987. Untuk hasil pengujian Rata-Rata didapatkan hasil seperti berikut, metode SVM mendapatkan nilai akurasi sebesar 97.01%, nilai presisi sebesar 99.03%, nilai recall sebesar 95.41%, dan nilai AUC sebesar (0-1) 0.997. Sedangkan metode *Naïve Bayes* mendapatkan nilai akurasi sebesar 90.86%, nilai presisi sebesar 91.33%, nilai recall sebesar 91.77%, dan nilai AUC sebesar (0-1) 0.975.

Ketiga, penelitian yang dilakukan oleh [12] untuk melakukan analisa dengan judul "*Sentiment Analysis of Social Media Twitter with Case of AntiLGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm*". Hasilnya adalah Pengujian dengan metode *Naïve Bayes* mendapatkan hasil untuk prediction positif yaitu nilai *true* positif sebesar 102, nilai *true* negatif sebesar 4, nilai *true* netral sebesar 53, nilai precision sebesar 64.15%, nilai recall sebesar 68.46%, nilai F1-measure sebesar 66.23%. Prediction negatif yaitu nilai *true* positif sebesar 25, nilai *true* negatif sebesar 4, nilai *true* netral sebesar 20, nilai precision sebesar 8.16%, nilai recall sebesar 36.36%, nilai F1-measure sebesar 13.33%. Prediction netral yaitu nilai *true* positif sebesar 22, nilai *true* negatif sebesar 3, nilai *true* netral sebesar 703, nilai precision sebesar 96.57%, nilai recall sebesar 90.59%, nilai F1-measure sebesar 93.48%.

Keempat, penelitian yang dilakukan oleh [13] untuk melakukan analisa dengan judul "*The Impact of Features Extraction on the Sentiment Analysis*". Hasilnya adalah Pengujian dengan metode *Naïve Bayes* mendapatkan hasil yaitu nilai accuracy sebesar 53%, nilai precision sebesar 56%, nilai recall sebesar 44%, nilai F-score sebesar 42%. Pengujian dengan metode *Support Vector Machine* mendapatkan hasil yaitu nilai akurasi 46%, nilai presisi 15%, nilai recall 33%, nilai F-score 21%.

Kelima, penelitian yang dilakukan oleh [14] untuk melakukan analisa dengan judul "*Textual Dissection of Live Twitter Reviews Using Naïve Bayes*". Hasilnya adalah Pengujian dengan metode *Naïve Bayes* mendapatkan hasil yaitu untuk prediction *yes* yaitu nilai *true:yes* sebesar 327, nilai *true:no* sebesar 45, nilai *class* precision sebesar 87.90%. untuk nilai *class* recall masing masing yaitu nilai *true:yes* sebesar 93.43%, nilai *true:no* sebesar 92.05%.

Dari beberapa penelitian diatas, metode *Naïve Bayes* mempunyai keunggulan yaitu data yang digunakan kecil, tingkat akurasi yang dihasilkan cukup tinggi. Sedangkan keunggulan dari metode SVM yaitu menghasilkan tingkat akurasi tinggi serta nilai AUC yang baik.

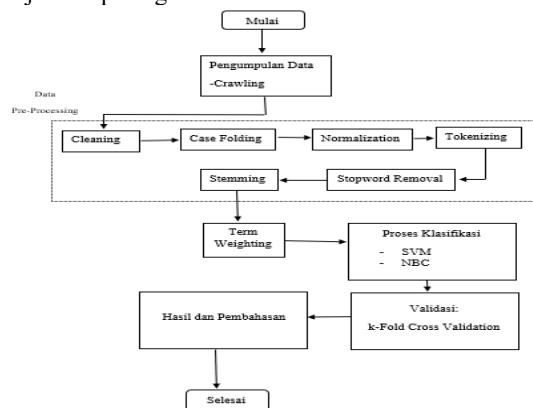
Berdasarkan keunggulan yang telah dijelaskan, peneliti hendak menggunakan metode *Naïve Bayes* dan *Support Vector Machine* untuk membandingkan tingkat akurasi dalam penelitian sekolah daring yang mendapat perhatian dikarenakan masih terdapat beberapa masalah yang belum terselesaikan selama kebijakan sekolah daring dilaksanakan selama situasi pandemi Covid-19. Data yang digunakan bersumber dari pendapat masyarakat di social media terutama Twitter.

Penelitian tentang sekolah daring dilakukan bertujuan untuk mengetahui respon masyarakat dari opini yang mereka sampaikan di sosial media terutama Twitter. Setelah itu opini mereka diklasifikasikan untuk mengetahui apakah respon yang dihasilkan berupa respon positif, respon negatif, atau respon netral terhadap sekolah daring. Oleh karena itu, penelitian ini memfokuskan pada eksplorasi metode *Naïve Bayes* dan *Support Vector Machine* metode *Naïve Bayes* dan *Support Vector Machine* untuk analisa opini publik tentang sekolah daring.

## II. METODOLOGI PENELITIAN

Penelitian ini dilakukan dengan menggunakan metode *Support Vector Machine (SVM)* dan *Naive Bayes Classifier (NBC)*. Studi kasus yang diangkat dalam penelitian ini yaitu tentang sekolah daring. Kedua metode tersebut dipilih hal ini dikarenakan kedua metode tersebut memiliki tingkat akurasi terbaik dalam melakukan analisa sentimen. Berdasarkan penjelasan tersebut diharapkan dalam melakukan pengujian mendapatkan tingkat akurasi yang baik untuk menganalisa pendapat masyarakat di media sosial terutama *Twitter*.

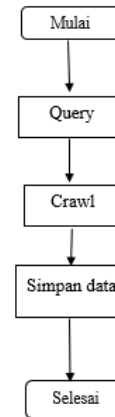
Tujuan dilakukan penelitian ini antara lain melakukan perbandingan dua metode yaitu *Support Vector Machine* dan *Naive Bayes Classifier* dalam melakukan analisa pendapat masyarakat tentang sekolah daring. Ada beberapa tahapan yang dilakukan untuk melakukan analisis sebagaimana yang ditunjukkan pada gbr 1.



Gbr 1. Alur Penelitian Analisis Opini

### A. Pemilihan Data

Proses pemilihan data pada penelitian ini dengan mengambil data dari Twitter yaitu mengambil tweet terdapat #sekolahdaring dan #sekolahonline. Untuk pengambilan data menggunakan cara crawling twitter, data yang didapat yaitu 500 tweet dan data disimpan dalam bentuk excel. Setelah itu data tersebut akan diproses ketahap proses selanjutnya. Alur pengambilan data ditunjukkan pada gbr 2.



Gbar 2. Alur Pengambilan Data

### B. Pre-Processing

Proses *Pre-Processing* pada penelitian ini dilakukan pada dataset, yaitu mengolah data sebelum dataset masuk ke tahap proses dengan menggunakan metode yang telah dipilih. Pada proses ini data akan diubah yang semula data tersebut tidak terstruktur menjadi data terstruktur hal ini dilakukan agar mempermudah pemrosesan data. Tahap pre-processing pada penelitian ini melalui beberapa langkah, yaitu:

1. *Cleaning*  
Proses ini dilakukan untuk membersihkan data yang memiliki noise seperti url, hastag dan lain-lain.
2. *Case Folding*  
Proses ini dilakukan untuk merubah huruf menjadi huruf kecil.
3. *Tokenizing*  
Proses ini merupakan proses pemotongan string input pada tiap kata.
4. *Filtering / stopword removal*  
Proses ini dilakukan untuk menghilangkan kata kata yang tidak terdapat kata penting dari hasil *Tokenizing*.
5. *Normalization*  
Proses ini dilakukan untuk menambahkan kata yang kurang atau tidak lengkap menjadi kata normal.
6. *Stemming*  
Proses ini dilakukan untuk menghilangkan kata yang terdapat imbuhan dan menjadikan kata tersebut menjadi kata dasar.

### C. Term Weighting

Setelah dilakukan *pre-processing* pada dokumen, dimana tahapan *pre-processing* akan menghasilkan kumpulan term atau kata. Selanjutnya proses yang dilakukan pada penelitian ini yaitu Term Weighting dengan cara TF-IDF. Penghitungan bobot setiap term dicari pada setiap dokumen setiap dokumen dengan tujuan untuk mengetahui kemiripan suatu term pada dokumen.

#### D. Support Vector Machine

*Support Vector Machine* merupakan salah satu metode dalam data mining dengan nilai akurasi yang dihasilkan sangat tinggi, SVM adalah algoritma paling populer dibandingkan algoritma yang digunakan pada proses data mining. Metode SVM berfungsi untuk melakukan perbandingan hasil sentimen, SVM akan mengubah dokumen berupa teks menjadi vektor.

Cara kerja *Support Vector Machine* yaitu mencari *hyperlane* dengan memaksimalkan jarak setiap kelas, *hyperlane* merupakan fungsi yang digunakan untuk memisahkan antar setiap kelas. *Support Vector Machine* merupakan sistem klasifikasi secara linier, untuk menyelesaikan masalah non linier metode SVM terlebih dahulu dimodifikasi dengan menggunakan fungsi Kernel. Persamaan kernel menggunakan formula pada tabel dibawah ini[15].

Tabel I  
Formula Kernel

Jenis Kernel	Formula
<i>Polynomial</i>	$K(a_1, a_2) = (a_1, a_2 + 1)^p$ (1)
<i>Gaussian RBF</i>	$K(a_1, a_2) = \exp(-\frac{\ a_1 - a_2\ ^2}{2\sigma^2})$ (2)
<i>Sigmoid</i>	$K(a_1, a_2) = \tanh(\alpha a_1, a_2 + \beta)$ (3)
Linier	$K(a_1, a_2) = a_1^t a_2$ (4)

#### E. Naïve Bayes Classifier

*Naïve Bayes Classifier* merupakan sebuah metode yang beranggapan bahwa setiap kata terjadi secara independen[16]. Metode ini sangat berguna dikarenakan beberapa alasan antara lain sangat mudah untuk dibangun, mudah untuk diinterpretasikan, dan tidak memerlukan skema estimasi parameter iteratif yang rumit. *Naïve Bayes Classifier* salah satu metode klasifikasi dalam data mining dengan probabilitas sederhana berdasarkan Teorema Bayes, *Naïve Bayes Classifier* dapat dilatih secara efisien dalam sistem pembelajaran terawasi (*supervised learning*). Keuntungan menggunakan metode klasifikasi ini yaitu metode ini hanya membutuhkan data pelatihan yang kecil untuk mengetahui parameter (sarana dan varians dari variabel) yang dibutuhkan dalam proses klasifikasi[17].

Pada penelitian ini, penerapan metode *Naïve Bayes Classifier* yaitu melakukan perhitungan nilai probabilitas dari masing-masing label kelas atau variabel dari dataset terhadap masukan yang diberikan, setelah itu nilai probabilitas tersebut akan dibandingkan satu sama lain untuk mencari nilai

probabilitas yang tertinggi. Label kelas atau variabel dengan nilai probabilitas tertinggi akan dijadikan label kelas data masukan. Persamaan *Naïve Bayes Classifier* menggunakan formula dibawah ini[18].

Teorema Bayes

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

#### F. Akurasi, Presisi, dan Recall

Pada proses akhir yaitu menghitung nilai akurasi, presisi, dan recall. Akurasi sendiri yaitu tingkat kedekatan antara nilai prediksi dengan nilai aktual. Presisi yaitu proses menunjukkan nilai ketepatan atau ketelitian dalam proses klasifikasi. Recall berfungsi untuk mengukur proposi positif aktual yang benar diidentifikasi. Pada proses ini memanfaatkan library yang telah disediakan oleh python. Untuk menghitung akurasi, presisi, dan recall menggunakan formula dibawah ini[19].

Akurasi

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Presisi

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

#### G. Hasil dan Pembahasan

Data yang telah melewati berbagai proses seperti proses pemilihan, data *pre-processing*, *term-weighting*, pengklasifikasi dengan metode SVM serta NBC, dan menghitung nilai akurasi, presisi, recall. Maka akan didapatkan hasil dari proses yang telah dilakukan yang akan dibahas pada bab berikutnya.

### III. HASIL DAN PEMBAHASAN

Penelitian ini melakukan analisa sentimen tentang sekolah daring dengan menggunakan dua Metode yaitu *Support Vector Machine (SVM)* dan *Naïve Bayes Classifier (NBC)*. Tujuan dari penelitian ini untuk membandingkan hasil akurasi dari dua algoritma yang digunakan.

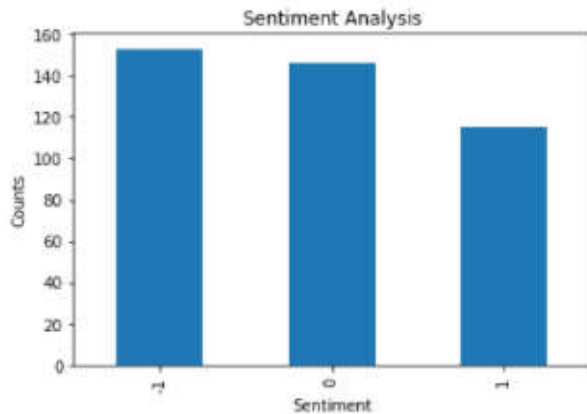
Pada tahap pertama pada analisa sentimen adalah memilih data yang berisi tweet tentang sekolah daring, data tersebut akan digunakan untuk proses pengujian. Pada proses pemilihan data disebut proses crawling. Proses pemilihan data dilakukan dengan mengambil tweet dari *twitter* dengan menggunakan pencarian “#sekolahdaring” dan “#sekolahonline”. Setelah data terkumpul, data tersebut akan memasuki tahap selanjutnya yaitu *pre-processing* dan juga proses selanjutnya seperti yang telah dijelaskan di bab metodologi.

#### A. Pemilihan Data



Proses pemilihan data dari tweet pengguna di twitter tentang sekolah daring yang diambil dengan kata yang telah ditentukan mendapat jumlah tweet sebanyak 414 *tweet*. Setelah data terkumpul kemudian diberi label. Untuk nilai 0 adalah netral, nilai -1 adalah negatif, dan nilai 1 adalah positif.

Setelah data terkumpul dan diberi label didapatkan label 0 atau netral sebanyak 146, label -1 atau negatif sebanyak 153, label 1 atau positif sebanyak 115. Representasi secara visual dapat dilihat pada gbr 3.



Gbr 3. Diagram Jumlah Dataset Sekolah Daring

#### B. Pre-Processing

Pada proses *Pre-Processing* terdapat beberapa tahapan yang dilakukan antara lain:

##### 1. Case Folding

Tahap paling awal yaitu proses *Case Folding*, dimana proses ini dilakukan untuk merubah semua huruf menjadi kecil. Contoh dari proses ini ditunjukkan pada tabel II.

##### 2. Cleaning

Tahap kedua yaitu proses *cleaning*, proses ini membersihkan data yang memiliki noise seperti url, hastag dan lain-lain. Contoh dari proses ini ditunjukkan pada tabel II.

##### 3. Tokenizing

Tahap ketiga yaitu proses *Tokenizing*, dimana pada proses ini akan melakukan pemisahan kata. Contoh dari proses ini ditunjukkan pada tabel II.

##### 4. Normalization

Tahap keempat yaitu proses *Normalization*, dimana pada proses ini akan melakukan normalisasi kata yaitu mengubah kata singkatan menjadi kata yang jelas. Contoh dari proses ini ditunjukkan pada tabel II.

##### 5. Stopword Removal

Tahap kelima yaitu proses *Stopword Removal*, dimana pada proses ini Proses ini dilakukan untuk menghilangkan kata yang tidak mengandung kata penting dari hasil *Tokenizing*. Contoh dari proses ini ditunjukkan pada tabel II.

##### 6. Stemming

Tahap kelima yaitu proses *Stemming*, dimana pada proses ini akan melakukan reduksi kata menjadi kata dasar. Contoh dari proses ini ditunjukkan pada tabel II.

Tabel II  
*Pre-Processing*

Tweet	
Sudah kuduga Makanya aq mending daring sekolah aja	
Text Pre-Processing	Hasil
Case Folding	sudah kuduga makanya aq mending daring sekolah aja
Cleaning	sudah kuduga makanya aq mending daring sekolah aja
tokenizing	[sudah, kuduga, makanya, aq, mending, daring, sekolah, aja]
Normalization	[sudah, kuduga, makanya, aku, mending, daring, sekolah]
Stopword Removal	['sudah', 'duga', 'makanya', 'aku', 'mending', 'daring', 'sekolah']
Stemming	Sudah duga makanya aku mending daring sekolah

#### C. Wordcloud

Hasil proses klasifikasi analisis sentimen digambarkan dengan menggunakan *wordcloud*. *Wordcloud* berguna untuk menunjukkan kata yang sering muncul dari data dalam bentuk visual.

Hasil *crawling* data telah melalui proses *pre-processing* dan pemberian label akan ditampilkan oleh *wordcloud*, dan tampilan tersebut dibagi menjadi tiga tampilan. Dibawah ini adalah tampilan *wordcloud* tentang sekolah daring. Tampilan gbr4 menunjukkan sentimen netral, gbr5 menunjukkan sentimen positif, gbr6 menunjukkan sentimen negatif.



Gbr 4. Wordcloud sentimen netral



Gbr 5. Wordcloud sentimen positif



Gbr 6. Wordcloud sentimen negatif

#### D. Hasil Pengujian Data

Dilakukannya proses pengujian untuk melakukan perbandingan metode dengan tingkat pengujian terbaik antara metode SVM dan NBC. Penelitian ini dilakukan untuk menentukan metode dengan tingkat pengujian terbaik.

Pada proses pengujian pertama yaitu proses dengan menggunakan split data yaitu dengan membagi data menjadi data uji dan data latih. Ratio ditentukan dalam pembagian data dengan beberapa ratio yaitu 1:9, 2:8, 3:7, 4:6 5:5. Pengujian pada penelitian ini mengambil contoh split menggunakan ratio 2:8, yaitu dimana perbandingan data 2 bagian atau 20% data uji dan 8 bagian atau 80% data latih.

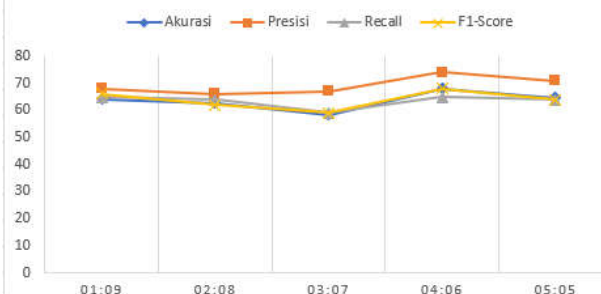
Pada proses pengujian kedua yaitu menggunakan *cross-validation*. Metode *cross-validation* disebut juga dengan *k-fold cross validation*, dimana nilai k yaitu jumlah sub latih. Proses pengujian pada penelitian ini menggunakan 10-fold.

Dibawah ini merupakan hasil akurasi yang berupa tabel pada proses uji data menggunakan metode SVM dan NBC untuk menganalisa sentimen pada penelitian yang dilakukan.

##### 1. Pengujian menggunakan Split Data

Proses pengujian pertama dalam penelitian ini yaitu menggunakan split latih dan uji. pada proses ini data akan diklasifikasi dengan dua metode perbandingan, dengan nilai ratio yang telah ditentukan diatas. Hasil proses klasifikasi dengan menggunakan split data ditunjukkan pada gambar 7 dengan metode NBC dan gambar 8 sampai gambar 11 dengan metode SVM.

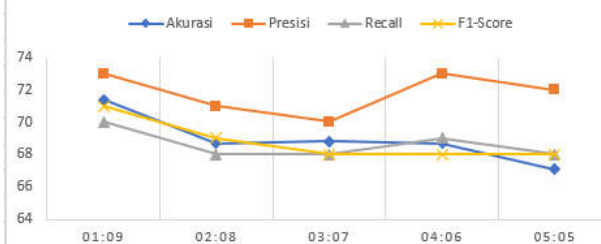
#### HASIL PENGUJIAN SPLIT DATA PERBANDINGAN PADA NBC



Gbr 7. Grafik Hasil Pengujian Split Data Perbandingan pada NBC

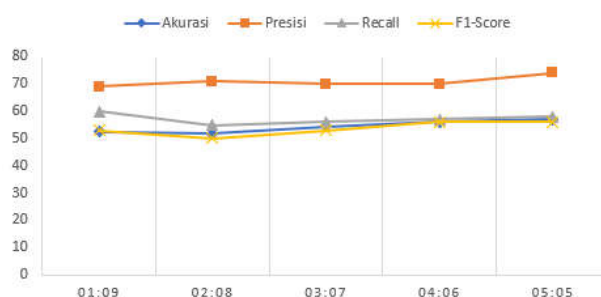
Berdasarkan gambar 7 didapatkan nilai akurasi tertinggi dari data yang telah dibagi menggunakan rasio perbandingan yaitu 4:6 dimana data dibagi menjadi 4 bagian atau 40% sebagai data uji dan 6 bagian atau 60% sebagai data latih. Hasil pengujian dengan menggunakan split data dengan metode NBC menggunakan rasio tertentu didapatkan nilai akurasi tertinggi sebesar 68%.

#### HASIL PENGUJIAN SPLIT DATA PERBANDINGAN PADA SVM (KERNEL = "LINEAR")



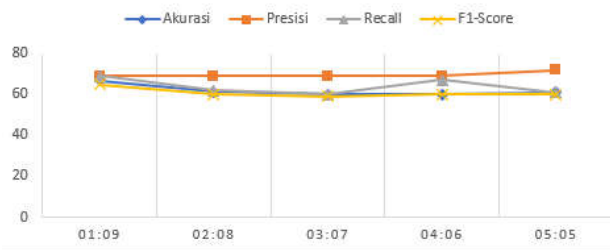
Gbr 8. Grafik Hasil Pengujian Split Data Perbandingan pada SVM (kernel = "linier")

#### HASIL PENGUJIAN SPLIT DATA PERBANDINGAN PADA SVM (KERNEL = "POLY")



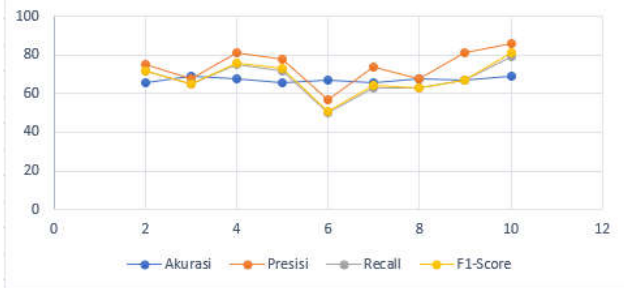
Gbr 9. Grafik Hasil Pengujian Split Data Perbandingan pada SVM (kernel = "poly")

### HASIL PENGUJIAN SPLIT DATA PERBANDINGAN PADA SVM (KERNEL = "RBF")



Gbr 10. Grafik Hasil Pengujian Split Data Perbandingan pada SVM (kernel = "rbf")

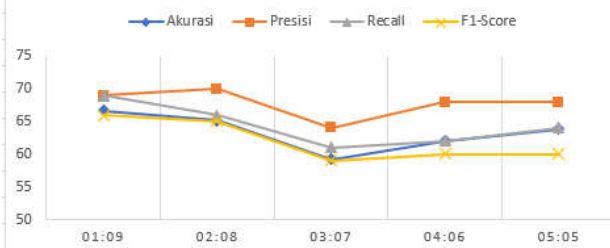
### Hasil k-fold cross validation pada metode NBC



Gbr 12. Grafik Hasil k-fold cross validation pada metode NBC

Berdasarkan tabel 12 hasil pengujian k-fold pada metode NBC didapatkan nilai *accuracy* paling dengan menggunakan nilai *k-fold* = 10 dengan nilai akurasi sebesar 69,25%..

### HASIL PENGUJIAN SPLIT DATA PERBANDINGAN PADA SVM (KERNEL = "SIGMOID")



Gbr 11. Grafik Hasil Pengujian Split Data Perbandingan pada SVM (kernel = "sigmoid")

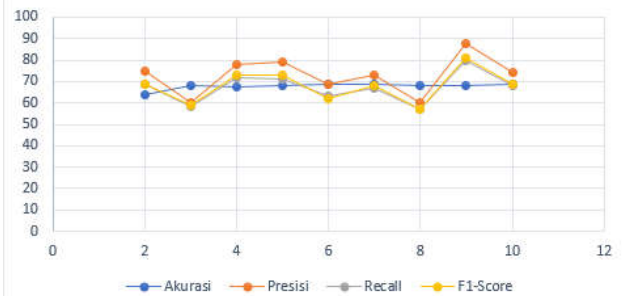
Berdasarkan gambar 13 sampai gambar 16 didapatkan nilai akurasi tertinggi dengan pembagian data menggunakan split data yaitu metode SVM dengan kernel rbf memiliki nilai akurasi sebesar 69,58% dengan nilai *k-fold* = 10. Hal ini menunjukkan bahwa metode SVM kernel rbf lebih optimal dibandingkan dengan Metode SVM menggunakan kernel poly, kernel linier, dan kernel sigmoid.

## 2. Pengujian menggunakan Cross Validation

Pada pengujian kedua yaitu dilakukan dengan menggunakan *cross validation*. Pada pengujian ini nilai *k* ditentukan dengan 2-fold sampai 10-fold untuk mendapatkan nilai *accuracy* dari setiap nilai *k* yang dilakukan pengujian.

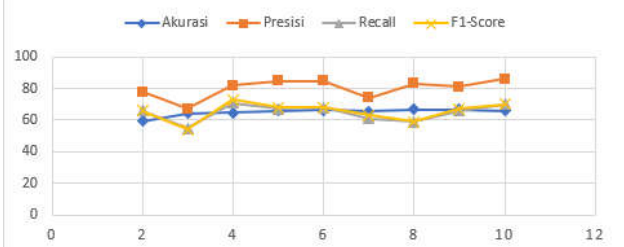
Berikut dibawah ini merupakan . Hasil proses klasifikasi dengan menggunakan *Cross Validation* ditunjukkan pada gambar 12 dengan metode NBC dan gambar 13 sampai gambar 16 dengan metode SVM.

### Hasil k-fold cross validation pada metode SVM (kernel = "linear")

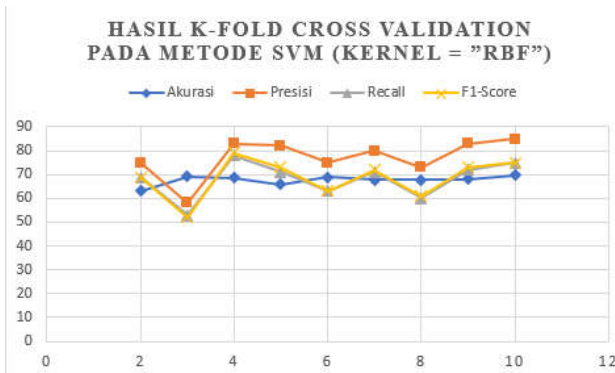


Gbr 13. Grafik Hasil k-fold cross validation pada metode SVM (kernel = "linier")

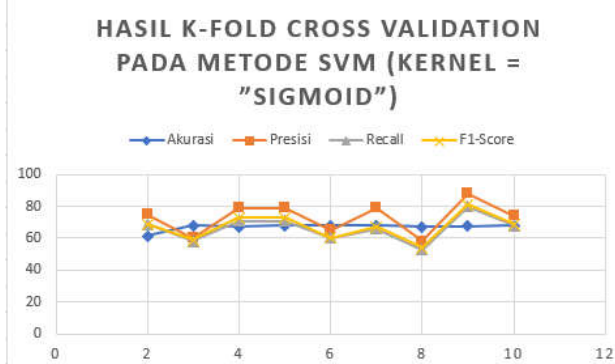
### HASIL K-FOLD CROSS VALIDATION PADA METODE SVM (KERNEL = "POLY")



Gbr 14. Grafik Hasil k-fold cross validation pada metode SVM (kernel = "poly")



Gbr 15. Grafik Hasil *k-fold cross validation* pada metode SVM (kernel = "rbf")



Gbr 16. Grafik Hasil *k-fold cross validation* pada metode SVM (kernel = "sigmoid")

Berdasarkan gambar 13 sampai gambar 16 didapatkan nilai akurasi tertinggi dengan pengujian data menggunakan *k-fold cross validation* yaitu metode SVM dengan kernel linear memiliki nilai akurasi sebesar 69,58% dengan nilai *k-fold* = 10. Hal ini menunjukkan bahwa metode SVM kernel linear lebih optimal dibandingkan dengan Metode SVM menggunakan kernel poly, kernel rbf, dan kernel sigmoid.

#### IV. KESIMPULAN

Penerapan metode dari SVM dan NBC dalam analisa sentimen tentang sekolah daring didapatkan bahwa tingkat accuracy yang didapatkan dengan menggunakan metode SVM lebih tinggi dibanding dengan menggunakan metode NBC dengan menggunakan data sebanyak 414 tweet yang terbagi menjadi 3 sentimen yaitu positif sebesar 27.7%, negatif sebesar 36.9%, dan netral sebesar 35.2%. Hal tersebut ditunjukkan hasil pengujian dengan menggunakan dua pengujian yaitu split data ratio dan cross-validation.

Hasil uji dengan menggunakan split data pada metode NBC mendapatkan nilai accuracy tertinggi yaitu pada ratio 4:6 dengan nilai sebesar 68%, untuk nilai precision 74%, recall 65%, dan f1-score 68%. Sedangkan pada metode SVM dengan kernel=linier mendapatkan nilai accuracy tertinggi yaitu pada ratio 1:9 dengan nilai sebesar 71,4%. untuk nilai precision 73%, recall 70%, dan f1-score 71%.

Untuk hasil uji dengan menggunakan cross-validation pada metode NBC mendapatkan nilai accuracy tertinggi yaitu pada k=10 dengan nilai sebesar 69,25%, untuk nilai precision 86%, recall 79%, dan f1-score 81%. Sedangkan pada metode SVM dengan kernel=linier mendapatkan nilai accuracy tertinggi yaitu pada k=10 dengan nilai sebesar 68,87% untuk nilai precision 74%, recall 68%, dan f1-score 69%.

#### V. SARAN

Pada penelitian sejenis untuk kedepannya dapat dikembangkan dengan menggunakan dataset yang lebih banyak dan juga dapat menambahkan fitur seperti n-gram. Dan juga dapat melakukan perbandingan dengan metode yang lain dalam ruang lingkup data mining seperti KNN, Random Forest, dan Decision Tree.

#### UCAPAN TERIMA KASIH

Dengan mengucap alhamdulillah atas karunia dari Allah SWT yang sudah diberikan pada saya, sehingga penelitian yang saya lakukan dapat diselesaikan dengan baik. Shalawat selalu turcurahan kepada Nabi Muhammad SAW. Terimakasih juga kepada orang tua, dosen pembimbing, serta orang-orang disekitar saya yang selalu memberi semangat dan juga dukungan kepada saya dalam menuntut ilmu.

#### REFERENSI

- [1] (2020) Taniya, N. "kronologi munculnya covid-19 di indonesia hingga terbit keppres darurat kesehatan". <https://www.merdeka.com/trending/kronologi-munculnya-covid-19-di-indonesia-hingga-terbit-keppres-darurat-kesehatan-kl.html?page=all>. Tanggal akses : 15 januari 2021.
- [2] (2020) Satgas Covid-19. "Hub InaCovid-19". <https://bnpb-inacovid19.hub.arcgis.com/>. Tanggal akses : 18 januari 2021.
- [3] Virdita, R. 2020. "instruksi bekerja dari rumah". <https://www.kompas.com/tren/read/2020/03/16/195035165/jokowi-instruksikan-bekerja-dari-rumah-ini-arti-work-from-home?page=all>. Tanggal akses : 15 januari 2021.
- [4] (2020) Joko. S, "belajar secara daring", [Online], <https://republika.co.id/berita/q7klaz318/belajar-secara-daring>. Tanggal akses : 18 januari 2021.
- [5] (2020) Sri, H. "efektivitas pembelajaran dari di masa pandemi covid-19". <https://bdjakarta.kemenag.go.id/berita/efektivitas-pembelajaran-daring-di-masa-pandemi-covid-19>. Tanggal akses : 18 januari 2021.
- [6] (2020) Munjani, S. "kesenjangan akses internet pengaruhi kualitas pendidikan warga". <https://saifulmunjani.com/kesenjangan-akses-internet-pengaruhi-kualitas-pendidikan-warga/>. Tanggal akses : 18 januari 2021.
- [7] L. Bing, "Sentiment Analysis and Opinion Mining". Morgan and Claypool Publishers, 2012.
- [8] T. Henni, I. Ade, "Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode Support Vector Machine dan Naïve Bayes Classifier". Jurnal informatika: Jurnal Pengembangan IT (JPIT), vol. 03, 2018.
- [9] Wu. Xi. Dong, V. Kumar, "Top 10 Algorithms in Data Mining". Knowledge and Informations system, 2009.
- [10] U. M. Aman, M. S. Maliha, B. S. Ara, D. N. Saha, "An algorithm and method for sentiment analysis using the text and emoticon". ICT Express, vol. 6, pp 357 – 360, 2020.
- [11] A. Oki, S. B. Theopilus, "Analisa Perbandingan Tingkat Performansi Metode Support Vector Machine dan Naïve Bayes Classifier Untuk Klasifikasi Jalur Minat SMA", Seminar Nasional Teknologi Informasi dan Multimedia 2018.
- [12] F. A. Venny., A. Rachmadita, H. A. Muhammad, "Sentiment Analysis of Social Media Twitter with Case of AntiLGBT Campaign in



- Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm”. *Procedia Computer Science*, vol. 161, pp 765 - 772, 2019.
- [13] A. Ravinder, C. Aakarsha, K. Shruti, G. Shaurya, A. Pratyush, 2019, “The Impact of Features Extraction on the Sentiment Analysis”, *Procedia Computer Science*, vol. 152, pp 341 – 348, 2019.
- [14] K.Sourav, S. Arijit, V. Aman, T. Vivek, “Textual Dissection of Live Twitter Reviews Using Naïve Bayes”. *Procedia Computer Science*, vol. 132, pp 307 - 313, 2018.
- [15] R. Faisal, S. P. Yoga, F. M. Agtamas, “Implementasi *Twitter Sentiment Analysis* Untuk *Review* Film Menggunakan Algoritma *Support Vector Machine*”. *Jurnal Informatika Polinema*, 2018.
- [16] E. Irish, “An empirical study of the naive Bayes classifier”, *IJCAI-01 Work. Empir. Methods AI*, 2019.
- [17] S. Wahyudi, R. Sofie, “Sistem Pakar Diagnosis Penyakit Mata Menggunakan Naïve Bayes Classifier”. *Conference: Seminar Nasional Sains dan Teknologi*, 2014.
- [18] R. Fajar, “ Implementasi Algoritma *Naive Bayes* Terhadap Analisis Sentimen Opini Film Pada Twitter”. *Jurnal Inovtek Polbeng – Seri Informatika*, vol. 03, no. 1, 2018.
- [19] M. Qurban, A. K. Shakeel, “Data Science Theory, Analysis, and Applications”. *CRC Press Taylor and Francis Group*, 2013.
- [20] Zhang. H, “The Optimality of Naïve Bayes”. *FLAIRS Conference* 2004.