

Prediksi Video Performance Akun Youtube Buzzfeed Menggunakan Metode Naïve Bayes

¹Valerie Leonie, ²Jefri Junifer Pangaribuan
Universitas Pelita Harapan
Medan, Indonesia

¹vleonie.vl@gmail.com, ²jefri.pangaribuan@uph.edu

Abstrak— Selama beberapa tahun terakhir, perkembangan media sosial sangat pesat sehingga mencakup hampir seluruh aspek kehidupan manusia. Dengan berkembangnya media sosial ini, tentu saja hiburan dan teknologi ikut berkembang beriringan bersamanya. Orang-orang sekarang mencari kesenangan bersosialisasi dan berbagi segala sesuatunya melalui internet, salah satunya melalui video dan disebarakan melalui Youtube. Hal ini mendorong banyak orang sekarang menjadikan platform tersebut sebagai media untuk mata pencaharian mereka, salah satunya adalah Buzzfeed. Oleh karena itu diperlukan penelitian mengenai performa video-video tersebut untuk mengetahui bagaimana perkembangan akun Youtube tersebut kedepannya. Dari sekian banyak metode untuk memprediksi performa video-video tersebut, digunakanlah metode *Naive Bayes*. Metode ini digunakan karena merupakan metode yang mudah serta cepat, serta dapat bekerja dengan jumlah data yang sedikit. Pengumpulan data dilakukan dengan teknik dokumentasi, dengan mengumpulkan semua data-data yang ada pada akun Youtube Buzzfeed dan kemudian diolah sesuai keperluan penelitian. Total data yang dikumpulkan selama setahun yaitu pada tahun 2019 mencapai 772 data yang kemudian dikelola menggunakan software Rapid Miner menggunakan metode *Naive Bayes*. Kemudian dari penelitian ini dapat disimpulkan bahwa *Naive Bayes* memiliki akurasi total yang cukup baik yaitu 86.01%, dengan *class recall* untuk kelas rendah sebesar 91.21%, dan kelas sedang dan tinggi sebesar 0%, sedangkan *class precision* untuk prediksi kelas rendah sebesar 94.32% dan kelas sedang dan tinggi sebesar 0%. Hal ini dikarenakan minimnya variasi data dengan kelas sedang dan tinggi sehingga algoritma masih sulit untuk memprediksi dengan tepat kelas dengan nilai sedang dan tinggi tersebut.

Keywords—algoritma prediksi; naïve bayes; *video performance*

I. PENDAHULUAN

Smartphone telah menjadi salah satu bagian yang tidak dapat dipisahkan dalam hidup manusia zaman sekarang, dapat dilihat dari survei yang dilakukan oleh *Pew Research Center* dimana Indonesia sebagai negara urutan keenam dari Sembilan negara berkembang yang mengalami peningkatan pengguna *smartphone* dari 39% hingga 66% dari tahun 2015-2018 untuk pengguna umur 18-34 tahun, dan 13 % untuk pengguna di atas 50 tahun (Alfarizi, 2019).

Media sosial adalah media *online* yang digunakan untuk bersosialisasi dengan orang-orang dari seluruh penjuru dunia secara *online*. Menurut hasil riset *Wearesocial Hootsuite* pada Januari 2019, pengguna media sosial di Indonesia mencapai 15 juta atau 56% dari total populasi Indonesia (Databoks, 2019). Youtube sendiri sebagai salah satu media sosial yang paling sering dipakai di dunia

dengan rekor yang sangat fantastis. Pada tahun 2019 mencapai sekitar dua miliar pengguna per bulannya atau bisa dihitung sekitar 25% dari populasi penduduk dunia dengan sekitar 500 jam konten baru yang diunggah ke Youtube setiap menitnya. Hal ini tentu mendorong nilai jual Youtube semakin tinggi, hingga mencapai sekitar US\$ 140 miliar dengan keuntungan per tahun sekitar US\$ 20 miliar (Gunawan, 2019).

Salah satu *channel* Youtube yang cukup populer di Amerika Serikat adalah Buzzfeed. Akun Youtube Buzzfeed memiliki nama Buzzfeed Video yang dibuat pada tanggal 10 Agustus 2011 dengan total *subscriber* sebanyak 20,1 juta, total 6.709 video, serta jumlah *views* sebesar 16.694.916.605 (per tanggal 16 April 2020 berdasarkan data dari *website SocialBlade*). Dengan banyaknya video dan jumlah penonton yang dimiliki oleh Buzzfeed, dapat kita simpulkan bahwa Buzzfeed terus menerus



berinovasi untuk mencari topik-topik yang dapat mengunggah rasa penasaran publik dan mengetahui apa yang sebenarnya diinginkan oleh para penonton videonya. Oleh karena itu, diperlukan suatu analisa untuk mencari topik-topik apa yang lebih disukai oleh para penonton video Buzzfeed, dan dengan data-data yang dimiliki, kita dapat memperkirakan jumlah *views* dan *likes* yang akan didapatkan per harinya per video yang diunggah oleh Buzzfeed untuk meningkatkan efisiensi biaya serta sumber daya pada konten-konten yang tepat.

Adapula teknik yang akan digunakan untuk melakukan penelitian ini adalah teknik *Naive Bayes*. *Naive bayes* merupakan suatu metode klasifikasi yang menggunakan probabilitas dan statistik untuk memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Informatologi, 2017). Selain itu, *Naive Bayes* merupakan metode yang memanfaatkan perhitungan matematik dasar dengan kuat sehingga efisiensi klasifikasinya stabil dan memiliki tingkat kesalahan yang minim dibanding metode klasifikasi lainnya. Kelebihan dari *Naive Bayes* adalah klasifikasi yang dapat dipersonalisasi sesuai dengan sifat dan kebutuhan masing-masing, serta merupakan metode yang banyak digunakan dalam analisis kategorikal dengan menggunakan teks, oleh sebab itu, metode *Naive Bayes* dianggap cukup efektif dalam penelitian ini.

II. TINJAUAN PUSTAKA

A. Data Mining

Dengan perkembangan teknologi dan informasi yang sangat cepat, saat ini data merupakan aset yang sangat penting. *Data mining* merupakan proses pengolahan data menjadi pengetahuan. Ada beberapa fungsi dari *data minig* yaitu (Kurniawan, 2019):

1. *Classification*, yaitu menyimpulkan definisi-definisi karakteristik sebuah grup
2. *Clustering*, yaitu mengidentifikasi kelompok-kelompok dari atau produk-produk yang mempunyai karakteristik khusus.
3. *Association*, yaitu mengidentifikasi hubungan antara kejadian-kejadian yang terjadi pada suatu waktu
4. *Sequencing*, yaitu mengidentifikasi hubungan-hubungan yang berbeda pada suatu periode waktu tertentu
5. *Forecasting*, yaitu memperkirakan nilai pada masa yang akan datang berdasarkan pola-pola dengan sekumpulan data yang besar

B. Naive Bayes

Metode *Naive Bayes* merupakan metode yang dikemukakan oleh salah satu ilmuwan Inggris bernama Thomas Bayes, yang merupakan salah satu teknik klasifikasi berdasarkan Teorema Bayes dengan asumsi bahwa setiap variabel saling bebas (independen), atau dengan kata lain setiap kejadian walaupun sebenarnya saling bergantung satu sama lain, semua kejadian ini akan dianggap tidak berhubungan terhadap probabilitas yang mungkin akan terjadi. *Naive bayes* pada dasarnya digunakan untuk klasifikasi dokumen, deteksi spam atau filtering spam, dan masalah klasifikasi lainnya (Wajih, 2017).

Dasar dari teorema *Naive Bayes* adalah:

$$P(C|X) = \frac{(P(x|c) \cdot P(c))}{P(x)}$$

Keterangan:

x : data dengan class yang belum diketahui

c : hipotesis data merupakan suatu class spesifik

$P(C|X)$: probabilitas hipotesis berdasarkan kondisi (*posterior probability*)

$P(c)$: probabilitas hipotesis (*prior probability*)

$P(x|c)$: probabilitas berdasar kondisi pada hipotesis

$P(x)$: probabilitas A

Kelebihan dari *Naive Bayes* adalah:

1. Lebih mudah dan cepat untuk memprediksi *class* dari dataset, serta bekerja dengan baik di prediksi *multi class*
2. Ketika asumsi independen terjadi, *Naive Bayes* bekerja lebih baik dibandingkan dengan metode lain seperti *logistic regression* dan lebih sedikit *training data* yang diperlukan
3. Bekerja baik dalam kasus variabel *input* kategori dibandingkan dengan variabel numerik. Untuk variabel numerik, diasumsikan distribusi normal

Sedangkan kekurangan *Naive Bayes* adalah:

1. Apabila variabel kategorial memiliki kategori (di *test dataset*) yang tidak terdeteksi di *training dataset*, maka akan menghasilkan probabilitas 0 dan ketidakmampuan melakukan prediksi. Hal ini sering disebut dengan "*zero frequency*". Untuk mengatasi hal ini, dapat digunakan



smoothing technique. Salah satu *smoothing technique* disebut *Laplace Estimation*.

2. Asumsi prediktor independen karena pada nyatanya hampir tidak mungkin mendapatkan set prediktor yang sama sekali independen.

Teori *Naive Bayes* merupakan salah satu teori populer yang banyak digunakan untuk berbagai hal dan mempunyai tingkat akurasi yang tinggi. Hal ini dapat dilihat dari penelitian yang dilakukan oleh (Praningsi & Budi, 2017) yang memprediksi penyakit kanker serviks dengan membandingkan tiga metode yaitu CART, *Naive Bayes*, dan k-NN, dimana hasil penelitian dapat dilihat dari tabel berikut:

TABEL 1 HASIL PENGUJIAN METODE CART DECISION TREE

Jumlah Data Uji	Accuracy	Recall	Precision	Error Rate
234	88,89%	91,38%	71,62%	11,11%

TABEL 2 HASIL PENGUJIAN METODE NAIVE BAYES

Jumlah Data Uji	Accuracy	Recall	Precision	Error Rate
234	94,44%	81,03%	95,92%	5,56%

TABEL 3 HASIL PENGUJIAN METODE K-NN

Jumlah Data Uji	Nilai k	Accuracy	Recall	Precision	Error Rate
234	1	85,04 %	46,55 %	87,10 %	14,96 %
234	3	85,04 %	44,83 %	89,66 %	14,96 %
234	5	85,04 %	44,83 %	89,66 %	16,67 %
234	7	83,33 %	39,66 %	85,19 %	15,39 %
Rata-rata		84,61 %	43,97 %	87,90 %	15,39 %

III. METODOLOGI PENELITIAN

A. Metode Pengumpulan Data

Data-data yang diperlukan dalam penelitian ini diambil dari data-data selama satu tahun dimulai dari tanggal 1 Januari 2019 hingga 31 Desember 2019 dengan total sebanyak 772 video. Kemudian dari video-video itu diambil lagi data-data seperti jumlah *views* yang dimiliki per videonya, total *thumbs-up* dan *thumbs-down*, komentar, *crew video*, ada atau tidaknya *subtitle*, apakah berupa kolaborasi atau tidak, kategori serta durasi video. Sedangkan data berupa jumlah *subscriber* per bulan diambil dari *website* Social Blade.

Kemudian data-data seperti *thumbs-up*, *thumbs-down* serta *comments* dikalkulasi sesuai dengan *Youtube Engagement Rate Formula (YER)* untuk menghitung rasio interaksi per video yaitu dengan menggunakan rumus:

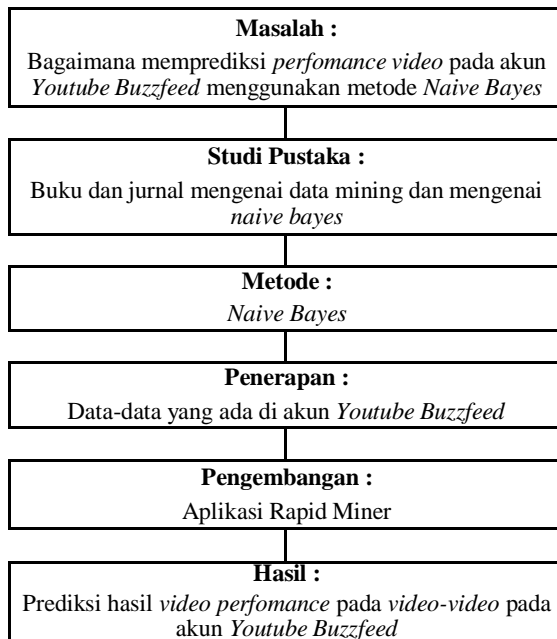
$$YER = \frac{(thumbs - up + thumbs - down + comments)}{subscriber}$$

Kemudian dari data-data tersebut, diambil kesimpulan apakah video tersebut memiliki hasil yang bagus atau kurang. Berdasarkan data yang diambil dari *Youtube Buzzfeed* per 2019, jumlah *subscriber* mereka mencapai angka diatas 18.106.587 *subscriber* pada awal Januari 2019, sedangkan pada akhir bulan Desember 2019, telah mencapai angka 19.800.000. Berdasarkan pembagian ukuran sosial media, akun *Buzzfeed* telah mencapai status ukuran nano dengan *subscriber* diatas 1 juta. Akun *Youtube* dengan ukuran nano memiliki rata-rata *engagement rate* 4.0% untuk dikatakan memiliki *engagement rate* yang bagus (Semeraro, 2019).

Untuk menghitung nilai *performance*, maka dapat dilihat dari berapa jumlah *views* yang dimiliki per video dibandingkan dengan jumlah *subscriber* pada saat itu. Hasil perhitungan ini dinamakan *view rate*. Sebuah video dikatakan memiliki *performance* yang bagus apabila memiliki *viewers* 14% dari jumlah *subscriber* yang dimiliki (Robertson, 2015). Maka berdasarkan hal tersebut, nilai video *performance* dianggap tinggi apabila mencapai angka diatas 14%, dianggap bernilai sedang apabila diatas 4% dan bernilai rendah apabila dibawah angka 4%.

Untuk memudahkan pemahaman atas penelitian ini, disusunlah kerangka berpikir seperti berikut:

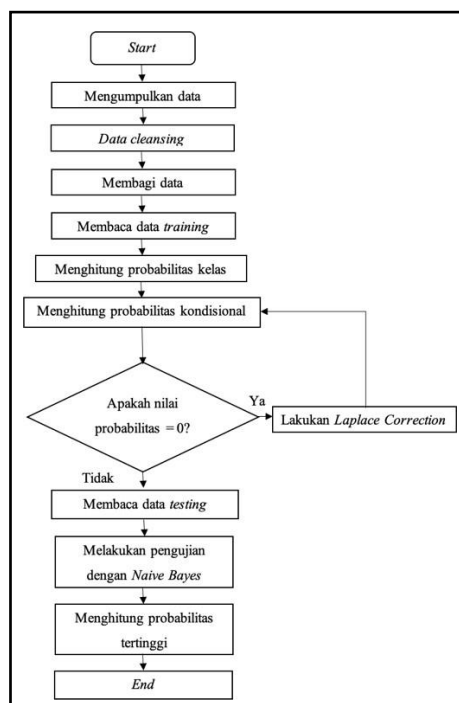




GAMBAR 1 BAGAN KERANGKA BERPIKIR

B. Metode Penyelesaian

Berikut adalah tahapan penggunaan metode *Naive Bayes*:



GAMBAR 2 BAGAN PENYELESAIAN MASALAH

C. Metode Prediksi dengan *Naive Bayes*

Data yang telah dikumpulkan dilakukan tahapan *data cleansing*, dimana pada tahap ini data diubah sedemikian rupa untuk dapat memenuhi kebutuhan penelitian, sehingga data menjadi konsisten dan dapat dimanfaatkan sesuai dengan kebutuhan. Pada tahapan *data cleansing*, data-data berupa data numerik seperti data jumlah *views*, *thumbs-up*, *thumbs-down*, dan *comments* diakumulasi hingga menjadi data non-numerik seperti *engagement result*, *like* or *dislike* serta *video performance*.

- Data *engagement result* diambil dari data *engagement rate* yang apabila dibawah 4% berarti rendah, dan apabila diatas 4% bernilai tinggi
- Like* or *dislike* diambil dari perbandingan jumlah *thumbs-up* dan *thumbs-down*, apabila rasio *thumbs-up* lebih tinggi daripada *thumbs-down*, maka menghasilkan nilai *like*, sedangkan jika rasio *thumbs-down* lebih tinggi, maka akan menghasilkan nilai *dislike*
- Video performance* mempunyai 3 kategori, yaitu tinggi ($view\ rate \geq 14\%$), sedang ($4\% < view\ rate < 14\%$) dan rendah ($view\ rate < 4\%$).

Pada penelitian ini, *data testing* dan *data training* memiliki porposisi masing-masing 75% dan 25% dari semua total data. Pada dasarnya, pada tahapan data training memiliki tiga tahapan yaitu:

- Menghitung probabilitas kelas
 Perhitungan probabilitas kelas dapat menggunakan rumus:

$$P(C_i) = \frac{C_i}{(C_1 + C_2 + \dots + C_i)}$$

Keterangan:

P: Probabilitas

C_i: Kelas ke-i

P(C_i): Probabilitas kelas i

- Menghitung probabilitas kondisional
 Pada tahapan ini terdapat dua macam cara menghitung tergantung dengan jenis data yang dimiliki yaitu:

- Data numerik

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

- Data non-numerik

$$P(X|C_i) = \frac{X}{Total\ Data\ C_i}$$

3. Laplace correction

Tahapan ini dilakukan dengan menambah satu data pada sampel yang memiliki nilai 0 untuk menghindari nilai probabilitas 0. Laplace Correction memiliki persamaan berikut:

$$P_i = \frac{m_i + 1}{n + k}$$

Setelah menghitung probabilitas kelas dan kondisional, tahap selanjutnya adalah menggunakan data testing untuk melakukan pengujian menggunakan metode Naive Bayes. Pengujian ini dilakukan dengan menggunakan nilai probabilitas yang telah didapat dari tahap-tahap sebelumnya yaitu dengan cara menghitung nilai probabilitas setiap atribut berdasarkan masing-masing kelas dengan menggunakan rumus:

$$P(C_i|X) = (P(X1|C_i) * P(X2|C_i) * P(X3|C_i) * \dots * P(Xn|C_i)) * P(C_i)$$

IV. HASIL DAN PEMBAHASAN

A. Hasil Pengumpulan Data

Dengan melakukan metode dokumentasi, data-data dikumpulkan secara bertahap dengan memanfaatkan data-data yang tersedia pada video-video di akun Youtube Buzzfeed. Data-data tersebut merupakan data-data video Buzzfeed selama 1 tahun yaitu pada tahun 2019, dari awal tanggal 1 Januari 2019 hingga video terakhir yang diunggah pada tanggal 31 Desember 2019 dengan total video yang dikumpulkan berjumlah 772 video.

B. Hasil Penelitian

Sebelum dilakukannya perhitungan dengan menggunakan metode Naive Bayes, data-data yang dikumpulkan telah melalui tahapan data cleansing untuk memastikan bahwa data-data yang pada akhirnya akan diproses tetap konsisten dan menghindari kesalahan-kesalahan yang tidak diinginkan.

1) Pembagian data

Dataset yang telah selesai melalui proses data cleansing kemudian melalui proses pembagian data menggunakan operator SplitData yang terdapat pada software RapidMiner. Data dibagi sebesar 75% untuk data training dan 25% untuk data testing sehingga total data yang digunakan untuk data training adalah 579 data sedangkan data yang digunakan untuk data testing berjumlah 193 data.

2) Perhitungan menggunakan Naive Bayes

Setelah dataset dibagi menjadi dua, yaitu data training dan data testing, kemudian dilanjutkan

dengan perhitungan menggunakan metode Naive Bayes sesuai dengan tahapan berikut:

a) Membaca data training

Untuk mengkalkulasi data yang akan dianalisa dengan menggunakan metode Naive Bayes, maka diambil sampel 10 data dari hasil subset data training yang telah dibagi sebelumnya.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	Class
Auriyana Jackson	Patrick Ward	John Gannon	Adriana Gomez	John Gannon	Alex Allsup	rendah	like	People & Blogs	No	Yes	Short	tinggi
Henry Loevner	Michelle Hsieh	John Gannon	Ada Enechi	John Gannon	Henry Loevner	rendah	like	People & Blogs	No	Yes	Short	rendah
Dhruv Vohra	Dhruv Vohra	John Gannon	Inga Lam	John Gannon	Alex Allsup	rendah	like	People & Blogs	Yes	No	Short	sedang
Aria Inthavong	Aria Inthavong	Jessica Massa	Megan Boykoff	Adam Colman	Katherine Hartman	rendah	like	People & Blogs	Yes	Yes	Medium	rendah
Kevin McShane	Eric Talschi	John Gannon	Inga Lam	John Gannon	Alex Allsup	rendah	like	People & Blogs	Yes	No	Short	sedang
Sara Rubin	Caryn Esch	Caitlin Cowie	Inga Lam	Caitlin Cowie	Andrew St. Juste	rendah	like	People & Blogs	Yes	Yes	Short	rendah
Aria Inthavong	Maggie Mel	Jessica Massa	Inga Lam	Aria Inthavong	Alex Allsup	rendah	like	People & Blogs	No	Yes	Short	rendah
Emily Trio	Emily Trio	Elyse O'Neil/oran	Inga Lam	John Gannon	Joyce Louis-Jean	rendah	like	People & Blogs	No	Yes	Short	rendah
Teddy Villa	Teddy Villa	John Gannon	Inga Lam	John Gannon	Carol Rasaphan	rendah	like	People & Blogs	Yes	Yes	Short	rendah
Kevin McShane	Denny Wong	John Gannon	Kyra Kupetsky	John Gannon	Alex Allsup	rendah	like	People & Blogs	Yes	Yes	Medium	rendah

GAMBAR 3 SAMPEL DATASET TRAINING

b) Menghitung probabilitas kelas

Dari gambar 4.3, kita dapat melihat bahwa dataset tersebut memiliki 3 jenis kelas, yaitu kelas rendah, sedang dan tinggi. Dari sampel data tersebut, dapat dihitung probabilitas dari masing-masing kelas dengan membagi jumlah data per kelas dibagi dengan total data pada sampel. Contoh perhitungannya menggunakan sampel tersebut adalah:

- Peluang kelas rendah:

$$P(\text{rendah}) = \frac{\text{jumlah}(\text{rendah})}{\text{jumlah}(\text{rendah}) + \text{jumlah}(\text{sedang}) + \text{jumlah}(\text{tinggi})}$$

$$P(\text{rendah}) = \frac{7}{7+2+1}$$

$$P(\text{rendah}) = \frac{7}{10}$$

$$P(\text{rendah}) = 0.7$$

- Peluang kelas sedang:

$$P(\text{sedang}) = \frac{\text{jumlah}(\text{sedang})}{\text{jumlah}(\text{rendah}) + \text{jumlah}(\text{sedang}) + \text{jumlah}(\text{tinggi})}$$

$$P(\text{sedang}) = \frac{2}{7+2+1}$$

$$P(\text{sedang}) = \frac{2}{10}$$

$$P(\text{sedang}) = 0.2$$



- Peluang kelas tinggi:

$$P(\text{tinggi}) = \frac{\text{jumlah}(\text{tinggi})}{\text{jumlah}(\text{rendah}) + \text{jumlah}(\text{sedang}) + \text{jumlah}(\text{tinggi})}$$

$$P(\text{tinggi}) = \frac{1}{7+2+1}$$

$$P(\text{tinggi}) = \frac{1}{10}$$

$$P(\text{tinggi}) = 0.1$$

c) Menghitung probabilitas kondisional

Tahap selanjutnya adalah menghitung probabilitas kondisional yaitu dengan menghitung setiap probabilitas atribut terhadap nilai kelas. Berikut contoh perhitungan salah satu atribut yang ada yaitu probabilitas atribut *team lead* (A3):

TABEL 4 TABEL DATA TEAM LEAD

	Rendah	Sedang	Tinggi
John Gannon	3	2	1
Jessica Massa	2	0	0
Caitlin Cowie	1	0	0
Ellyse O'Halloran	1	0	0

$$P(A3 = \text{JohnGannon} | \text{class} = \text{rendah}) = \frac{3}{3 + 2 + 1 + 1}$$

$$P(A3 = \text{JohnGannon} | \text{class} = \text{rendah}) = \frac{3}{7} = 0.43$$

d) Laplace Correction

Dengan menerapkan persamaan yang diatas, maka perhitungan probabilitas kondisionalnya sebelum dilakukannya *laplace correction* adalah:

TABEL 5 PROBABILITAS TEAM LEAD SEBELUM LAPLACE CORRECTION

	Rendah	Sedang	Tinggi
John Gannon	3/7 = 0.43	2/2 = 1	1/1 = 1
Jessica Massa	2/7 = 0.29	0/2 = 0	0/1 = 0
Caitlin Cowie	1/7 = 0.14	0/2 = 0	0/1 = 0
Ellyse O'Halloran	1/7 = 0.14	0/2 = 0	0/1 = 0

Dari data diatas, maka kita dapat melihat bahwa ada probabilitas kondisional yang bernilai 0, oleh karena itu akan diterapkan *Laplace Correction*. Setelah diterapkannya *Laplace Correction*, data akan berubah seperti:

TABEL 6 DATA TEAM LEAD SETELAH LAPLACE CORRECTION

	Rendah	Sedang	Tinggi
John Gannon	3	3	2
Jessica Massa	2	1	1
Caitlin Cowie	1	1	1
Ellyse O'Halloran	1	1	1

Dari data tersebut, *laplace correction* dilakukan dengan menambah 1 nilai pada data yang bernilai 0, kemudian dilakukan hal yang sama pada semua atribut pada kelas yang sama.

e) Melakukan Pengujian dengan Naive Bayes

Dengan menerapkan perhitungan probabilitas kelas dan probabilitas kondisional yang dijelaskan pada tahap sebelumnya, maka dari hasil-hasil tersebut, kita dapat melakukan pengujian dengan menggunakan metode *naive bayes* dengan cara memanfaatkan data *testing*.

3) Implementasi Menggunakan Rapid Miner

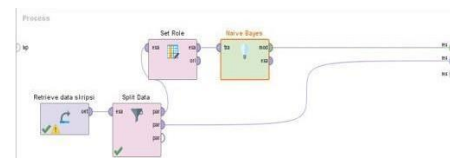
Penggunaan *software Rapid Miner* pada penelitian ini bertujuan untuk menganalisa sebuah *data* untuk kemudian digunakan untuk memprediksi kejadian kedepannya. Adapun tahapan dalam penggunaan *software Rapid Miner* adalah sebagai berikut:

- Hal yang pertama dilakukan dengan data adalah membagi data menjadi dua *dataset* yaitu data *training* dan data *testing* dengan rasio 0,75 dan 0,25 memanfaatkan operator *SplitData* yang ada pada *software Rapid Miner*.



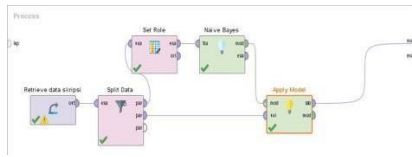
GAMBAR 4 TAHAPAN SPLIT DATA RAPID MINER

- Setelah data dibagi menjadi data *training* dan data *testing*, maka data *training* kemudian digunakan dengan operator *Naive Bayes* untuk dapat mendapatkan algoritma baru serta memanfaatkan operator *SetRole* untuk menetapkan *class video performance* sebagai atribut spesial diantara atribut-atribut lainnya.



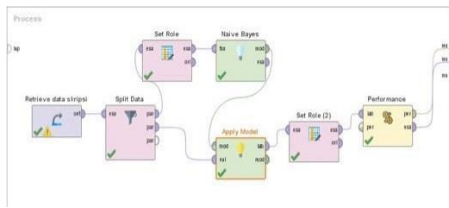
GAMBAR 5 TAHAPAN NAIVE BAYES RAPID MINER

- Kemudian setelah data *training* dilatih menggunakan *Naive Bayes*, hasilnya dimasukkan kedalam operator *Apply Model*, yang kemudian juga mendapatkan input dari *Split Data* sebelumnya berupa data *testing*. Data *testing* itu kemudian digunakan untuk menguji algoritma yang telah didapat dari data *training*.



GAMBAR 6 TAHAPAN APPLY MODEL RAPID MINER

- d. Dan tahapan terakhir, untuk mengevaluasi hasil kinerja dari perhitungan *Naive Bayes* pada data *training*, dilakukan dengan memanfaatkan operator *Performance* dan *Set Role* seperti yang telah digunakan sebelumnya untuk menentukan atribut spesial.



GAMBAR 7 TAHAPAN PERFORMANCE RAPID MINER

C. Pembahasan

Setelah melakukan penerapan metode *naive bayes* pada *software Rapid Miner*, hasil yang didapatkan adalah:

accuracy: 86.01%

	true rendah	true sedang	true tinggi	class precision
pred. rendah	166	10	0	94.32%
pred. sedang	12	0	0	0.00%
pred. tinggi	4	1	0	0.00%
class recall	91.21%	0.00%	0.00%	

GAMBAR 8 HASIL PERFORMANCE

Gambar 8 adalah hasil dari operator *performance* yang menghasilkan tabel diatas yang disebut *confusion matrix*. Dimana tabel ini bertujuan untuk membantu mengecek prediksi yang benar dan salah. Dari hasil *confusion matrix* tersebut, kita dapat melihat bahwa total akurasi mencapai angka 86.01%. Dari gambar tersebut juga, kita dapat menarik kesimpulan bahwa dari data *set*, ada 182 data *video performance* yang bernilai rendah namun hasil prediksi hanya berhasil memprediksi 166 secara benar, 12 data lainnya diprediksi bernilai sedang dan 4 lainnya bernilai tinggi, sehingga menghasilkan *class recall* sebesar 91,21%. Sedangkan untuk data *video performance* yang bersifat sedang dan tinggi, algoritma belum berhasil memprediksi kedua nilai *class* tersebut, sehingga menghasilkan nilai *class recall* 0% dikarenakan berdasarkan data, kebanyakan hasil data *video performance* menghasilkan nilai rendah mengakibatkan minimnya data dengan *video performance* bernilai sedang dan tinggi.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Dari hasil penelitian dan analisa yang dilakukan pada akun *Youtube Buzzfeed* menggunakan metode *naive bayes*, maka dapat diambil kesimpulan sebagai berikut:

1. Kesetaraan data *set* yang dimiliki sangat berpengaruh pada akurasi penelitian, karena pada kasus penelitian kali ini, berdasarkan data yang ada, *video performance* yang bernilai sedang dan tinggi cukup sedikit sehingga hal ini membuat prediksi untuk *class* bernilai *video performance* sedang dan tinggi tidak dapat mencapai hasil yang optimal.
2. Dari hasil prediksi yang dilakukan, hal ini menunjukkan bahwa kebanyakan hasil dari *video performance* akun *Buzzfeed* masih bernilai rendah.
3. Dikarenakan minimnya data dengan nilai *class video performance* sedang dan tinggi, hasil perhitungan algoritma belum sanggup menarik kesimpulan mengenai faktor-faktor yang mempengaruhi *video performance* dari akun *Youtube Buzzfeed*. Karena apabila terdapat lebih banyak data yang lebih variatif, dapat dengan lebih mudah mencari perbedaan dan persamaan dari setiap nilai *class video performance* yang bernilai rendah, sedang dan tinggi.

B. Saran

Berdasarkan penelitian yang telah dilakukan, ada beberapa hal yang dapat dilakukan untuk mendapatkan hasil yang lebih optimal, yaitu:

1. Menambah jumlah variasi data *set* sehingga dapat menambah akurasi hasil prediksi
2. Dapat membandingkan metode *Naive Bayes* dengan metode-metode lainnya untuk dapat mencapai hasil yang lebih spesifik dan akurat



VI. DAFTAR PUSTAKA

- ALFARIZI, M. K. (2019, MARET 4). *TEKNO SAINS*.
RETRIEVED FROM TEMPO:
<HTTPS://TEKNO.TEMPO.CO/READ/1181645/SURVEI-KEPEMILIKAN-SMARTPHONE-INDONESIA-PERINGKAT-KE-24/FULL&VIEW=OK>
- Databoks. (2019, Februari 8). *Demografi*. Retrieved from
Katadata:
<https://databoks.katadata.co.id/datapublish/2019/02/08/berapa-pengguna-media-sosial-indonesia>
- Gunawan, A. (2019, Agustus 31). *Fun Fact*.
Retrieved from IDN Times Jabar:
<https://jabar.idntimes.com/hype/fun-fact/arifgunawan/statistik-mencengangkan-soal-youtube-regional-jabar/full>
- Informatologi. (2017, April 8). *Algoritma Naive Bayes*. Retrieved from Informatologi:
<https://informatikalogi.com/algoritma-naive-bayes/>
- Kurniawan, A. (2019, Mei 26). Retrieved from Guru Pendidikan:
<https://www.gurupendidikan.co.id/data-mining/>
- Wajih, F. (2017, Oktober 1). *Apa saja kelebihan dan kekurangan metode naive bayes*. Retrieved from Dictio Id: <https://www.dictio.id/t/apa-saja-kelebihan-dan-kekurangan-metode-naive-bayes/12476/2>
- Praningki, T., & Budi, I. (2017). Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes dan k-NN. *Cite c Journal, Vol 4 No 2*, 83-92.
- Semeraro, E. D. (2019, Juli 9). *What Actually Counts As Good Engagement For Influencer Marketing*. Retrieved from Tubefilter:
<https://www.tubefilter.com/2019/07/09/influencer-marketing-good-engagement-rates-data/>
- Robertson, M. (2015, September 18). *Strategy*. Retrieved from TubularInsights:
<https://tubularinsights.com/3-metrics-youtube-success/>

