# Analisis Klasifikasi Tweet Suatu Akun Film Production Dengan Kontent-Based Dan Time-Based Menggunakan Metode Naïve Bayes

1st Rizki Luthfan Azhari
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
rizkiluthfan@students.telkomuniversity.ac.id

2nd Jondri
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
jondri@telkomuniversity.ac.id

3rd Kemas Muslim Lhaksamana
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
kemasmuslim@telkomuniversity.ac.id

**Abstrak**
Pada era digital yang serba modern ini, media sosial menjadi sarana atau platform untuk menyebarkan berbagai macam informasi secara mudah. Twitter merupakan salah satunya, twitter sendiri adalah sebuahmedia sosial yang bisa menyebarkan suatu informasi melalui tweet (kata kata yang diunggah oleh pengguna). Tweet bisa mengandung berbagai macam informasi, pembahasan, video, gambar maupun tautan ke suatu website. Suatu tweet akan disebarkan dari suatu pengguna ke pengguna lainnya dengan cara me-meretweetnya. Pada penelitian ini bertujuan untuk menganalisa apakah suatu tweet akan di retweet oleh pengguna lainnya dengan menggunakan fitur kontent-based dan time-based dengan metode klasifikasi naïve bayes serta menggunakan k-fold cross validation dengan nilai k=5 untuk melakukan splitdata. Hasil performansi yang didapatkan dengan menerapkan metode tersebut berupa nilai rata-rata akurasi 61,36%, rata-rata precision yang didapatkan sebesar 65,06%, rata-rata untuk recall sebesar 55,61%, lalu rata-rata untuk f1-score sebesar 50,49%.

***Abstract***
*In this modern digital era, social media has become a means or platform to easily disseminate various kindsof information. Twitter is one of them, twitter itself is a social media that can spread information through tweets (words uploaded by users). Tweets can contain various kinds of information, discussions, videos, images or links to a website. A tweet will be spread from one userto another by retweeting it. This study aims to analyze whether a tweet will be retweeted by other users using content-based and time-based features with the nave Bayes classification method andusing k-fold cross validation with a value of k=5 to split the data. The performance results obtained by applying this method are in the form of an average accuracy value of 61.36%, an average precision obtained of 65.06%, an average for recall of 55.61%, then an average for f1- score. by 50.49%.*

## I. PENDAHULUAN

Latar Belakang

Twitter adalah sebuah *microblogging* dari Amerika dan sebuah media sosial yang dimana user bisa mem-posting dan berinteraksi satu sama lain menggunakan pesan bernama "*Tweets*". Dalam *tweets* ini sendiri bisa berupa kata-kata atau teks, foto, video, maupun suatu tautan yang menghubungkan dengan media sosial lainnya.

Suatu perusahaan dalam dunia perfilman biasanya memiliki caranya tersendiri dalam memasarkan produk atau filmnya. Banyak *platform* yang bisa digunakan untuk melakukan advertising contohnya seperti iklan pada Youtube sesaat sebelum video yang dipilih oleh *user* dimulai, koran, majalah, radio dan sosial media (Instagram,Twitter, Facebook, etc). Pada sosial media biasanya perusahaan tersebut akan membuat suatu account yang akan memunculkan informasi mengenai film yang mereka produksi ke masyarakat luas. Informasi tersebut bisa berupa jadwal tayang, *trailer*, pemeran, dan sebagainya. *Account* yang mereka buat untuk pertama kali ini tidak ada bedanya dengan *account* biasa pada umumnya. Tapi seiring berjalannya waktu *account* tersebut akan memiliki banyak pengikut dikarenakan informasi yang mereka berikan lebih akurat dan lebih cepat jika dibandingkan dengan *account* lainnya. Oleh karena itu, Twitter *user* akan percaya bahwa *account* tersebut memang dikelola oleh perusahaan produsen film resmi, hal tersebut akan menghasilkan pengikut yang banyak untuk *account* perusahaantersebut sehingga perusahaan tersebut akan bisa mengajukan ke Twitter untuk menjadi *account* resmi atau yang biasa dibilang dengan *official account.*

Difusi sendiri merupakan suatu proses yang dimana suatu informasi akan disebarkan atau diteruskan melalui suatu interaksi. Berdasarkan penelitian yang dilakukan oleh Liangjie Hong, Ovidiu Dan, dan Brian D.Davidson yang berjudul *"Predicting Popular Message in Twitter"*, informasi dari suatu *tweet* akan disebarkan ke user lain atau orang lain dengan cara me-*retweetnya*[1], sehingga informasi yang berada didalam *tweet* tersebut akan denganmudah tersalurkan ke orang atau *user* lain. Pada penelitian[2] penulis mendapatkan hasil bahwa penelitiannya mendapatkan peningkatan F-measure sebesar 5% dibanginkan dengan *state of art* (secara statistik) dengan menggunakan 3 fitur yaitu keseluruhan dari berbasis pengguna, berbasis waktu, dan berbasis dan bebasis konten dengan menggunakan metode *random forest*. Lalu, pada penelitian[3] mengatakan bahwa metode random forest memiliki akurasi berkisar diantara 40%-90% dan bisa dioptimalkan menggunakan

algoritma genetika dan akan menunjukan hasil yang lebih baik yaitu berkisar diantara 91%-95%.

Untuk memastikan bahwa berita atau informasi mengenai produk yang ingin mereka pasarkan, suatu akun produksi film akan mempost hal-hal yang terkait dengan produk terbaru yang akan mereka pasarkan. Oleh karena itu sangat diperlukan pengetahuan untuk menganalisis faktor-faktor apa saja yang mempengaruhi jumlah retweetdan performansi dari model yang ingin dibangun dengan metode klasifikasi naive bayes.

Hal tersebut akan dilakukan dengan menggunakan metode klasifikasi naive bayes yang dimana akan memunculkan performansi dari penelitian ini. Naive bayes sendiri merupakan metode yang digunakan berdasarkan probabilitas yang dirancang sedemikian rupa yang dimana antar kelasnya memiliki ketergantungan (independen)[4].

A. Topik dan Batasannya

Topik dan batasan masalah dalam penilitian ini untuk mengetahui performansi berdasarkan skenario pengujian. Dalam penilitian ini batasan masalahnya adalah analisis performansi dari klasifikasi dataset yang diambil dari *tweet* suatu akun twitter produksi film dengan jumlah 950 data yang labelnya akan dibagi 2 berdasarkan *tweet* yang di-retweet sebanyak lebih dari 100 kali dan *tweet* yang di-retweet kurang dari 100 kali. Dalam dataset tersebut *tweet* akan dibagi menjadi 3 jenis *tweet* yang informasinya berisikan suatu film (jadwal tayang, tempat tayang, dsb.), aktor atau aktris yang mendapatkan penghargaan, dan *tweet* lainnya.

B. Tujuan

Tujuan dari penilitian ini adalah untuk memprediksi atau menganalisis suatu tweet apakah akan di retweet berdasarkan fitur yg digunakan dan mengetahui performansi dari model yang ingin dibangun dengan metode klasifikasi naive bayes menggunakan dataset yang menggunakan tweet dari akun twitter produksi film yaitu DisneyStudios.

C. Organisasi Tulisan

Bab 2 pada penelitian ini akan membahas tentang studi atau penelitian terkait, lalu bab 3 akan membahas tentang metode dan sistem penelitian yang dilakukan, bab 4 akan membahas tentang evaluasi model, dan bab5 akan membahas kesimpulan dari penelitian yang dilakukan.

## II. KAJIAN TEORI

Pada tahun 2019, Fera Fanesya, Randy Cahya Wihandika dan Indriati melakukan penelitian[5] dengan judul "Deteksi Emosi Pada Twitter Menggunakan Metode Naive Bayes Dan Kombinasi Fitur" yang betujuan untuk menentukan keadaan emosi seseorang dari isi tweetnya dengan menggunakan metode klasifikasi Naive Bayes dan kombinasi fitur. Penulis menggunakan fitur linguistik, fitur ortografik dan kombinasi fitur N-gram. Pada penelitian ini penulis mendapatkan hasil akurasi terbaik yang didapatan dari kombinasi tunggal dari fitur N-gram dengan nilai 0,555. Sedangkan penulis mendaparkan akurasi sebesar 0,5317 yang didapatkan dari kombinasi fitur yang meliputi fitur linguistik, fitur ortografik, dan N-gram. Yang dimana dapat disimpulkan bahwa kombinasi fitur seperti fitur linguistik, fitur ortografik, dan N-gram dapat menutupi kelemahan masing-masing fitur yang akan meningkatkan akurasi walaupun peningkatannya tidak terlalu signifikan.

Pada tahun 2015, Maresha Caroline Wijanto melakukan penelitian[6] yang bertujuan untuk membuat sistem dimana akan mendeteksi apakah pengirim atau pembuat suatu *tweet* merupakan *user* dari si pemilik akun atau bukan. Karena telah ditemukan banyak kasus penipuan dengan menggunakan SMS ataupun sosial media seperti twitter. Banyaj yang tertipu dengan sebuah *tweet* yang dikirim oleh orang tidak dikenal atau bukan *user* asli dari pemilik akun. Data yang diambil pada penelitian ini berasal dari token yang dipilih berdasarkan dua buah model yaitu jumlah kejadian n-waktu minimum dan jumlah kejadian tertinggi ke-n. Pada penelitian ini penulis mendapatkan hasil akurasi sebesar 82,145%

Di tahun 2018, Sigit Suryono, Ema Utami dan Emha Taufiq Luthfi melakukan penelian[7] yang berjudul "Klasifikasi Sentimen Pada Twitter Dengan Naive Bayes Classifier" yang dimana pada penelitiannya bertujuan untuk mencari tahu popularitas dari sentiment positif, negatif dan netral. Hal tersebut dapat diperoleh dari *Crawling Data* yang didapatkan menggunakan *tools* API Twitter. Penulis menggunakan metode klasifikasi untuk percobaan sebanyak 3 kali, dan didapatkan akurasi sebesar 64,95% untuk yang pertama, 66,36% untuk percobaan kedua, dan 66,79% untuk yang ketiga.

Penelitian yang dilakukan oleh Adhi Viky Sudiantoro dan Eri Zuliarso yang berujudul "Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier" pada tahun 2018[8], penelitian ini bertujuan untuk mengetahui reaksi masyarakat terhadap pilkada Jawa Barat apakah reaksi masyarakat tersebut positif atau negatif. Data yang digunakan sebanyak 300 *tweet* dan akan dibagi 2 menjadi 200 data latih dan 100

data uji, yang dimana 100 data uji yang diklasifikasi menghasilkan 32 data dengan nilai sentimen positif dan 68 data dengan nilai sentimen negatif. Setelah itu penulis mendapatkan akurasi dengan menggunakan algoritma Naive Bayes Classifier dan mendapatkan akurasi sebesar 84%.

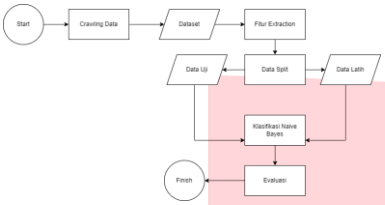
Syarli, dan Asrul Ashari Muin melakukan penelitian pada tahun 2016 dengan judul "Metode Naive Bayes Untuk Prediksi Kelulusan[9]. Penelitian ini bertujuan untuk melakukan klasifikasi terhadap kelulusan mahasiswa baru. Penulis menggunakan metode naive bayes karena memiliki kemampuan yang cukup tinggi untuk memprediksi peluang terhadap data yang akan diolah karena berdasarkan

penglaman atau data yang terdapat pada masa lalu. Kesimpulan yang didapat oleh penulis yaitu akurasi yang didapat dengan menunjukan keefektifan dataset yang diterapkan menggunakan metode algoritma naive bayes yaitu mencapai 94%.

## III. METODE

### A. Flowchart

Pada bagian ini berupa ilustrasi model penelitian yang akan dibangun menggunakan algoritma klasifikasi naive bayes. Alur penelitiannya diilustrasikan seperti pada gambar 1:

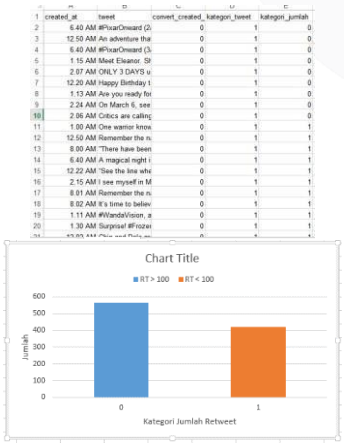



Gambar 1. Flowchart Sistem

Berdasarkan gambar diatas penelitian ini diawali dengan *Crawling Data* menggunakan package Rtweet untuk mendapatkan dataset berupa tweet dan atribut lainnya dari akun twitter produksi film DisneyStudios. Setelah itu akan dilakukannya fitur selection untuk meyeleksi fitur yang akan digunakan, dan melabeli serta mengubah tweet dan jumlahretweet yang didapatkan menjadi beberapa kategori. Setelah mendapatkan dataset yang dibutuhkan langkah berikutnya adalah menyaring beberapa fitur yang dibutuhkan, dan membuang yang tidak diperlukan. Untuk penelitian ini fitur yang digunakan hanya menggunakan fitur time-based dan content-based yang dimana didalamnya terdapat beberapa sub fiturlagi, untuk sub fitur ini juga tidak digunakan semua, hanya beberapa saja. Lalu setelah itu melakukan split data dengan menggunakan k-fold cross validation dengan k=5 untuk membagi data train dengan data test agar bisa diproses menggunakan klasifikasi naïve bayes pada step berikutnya. Dan diakhir akan dimunculkan nilai performansinya untuk dijadikan sebagai sevaluasi.

### B. Dataset

Dataset yang digunakan untuk penelitian ini adalah data hasil *Crawling* tweet dari produksi film DisneyStudios menggunakan *Rtweet*. Lalu data yang digunakan diseleksi dengan memilah fitur apa sajayang digunakan, setelah itu dilabeli dan dibagi menjadi beberapa kategori.




Gambar 2. Label Kelas dan Dataset

### C. Crawling Data

Crawling Data merupakan salah satu teknik atau cara untuk mengumpulkan data pada sebuah *platform*, contohnya seperti twitter, dengan memasukan *Uniform Resource Locator* (URL). Hasil *crawling data* ini adalah sebuah data yang berasal dari server twitter yang berisikan isi tweet, nama akun yang melakukan tweet tersebut serta atribut-atribut lainnya[10]. Melakukan crawling data ke twitter juga bisa dilakukan dengan menggunakan *Application Programming Interface* (API) twitter. API sendiri merupakan program yang sudah disediakan oleh pihak twitter untuk membantu developer dalam melakukan penelitian dan membantu developer untuk mendapatkan beberapa dataset yang terdapat pada twitter[10]. Untuk menggunakan API *user* atau *developer* harus mendaftar di laman https://developer.twitter.com untuk mendapatkan consumer key, consumer access, access token dan access token secret yang akan digunakan sebagai syarat authentifikasiuntuk dapat mengakses data yang terdapat pada twitter.

### D. Klasifikasi Naïve Bayes

Klasifikasi Naïve Bayes atau Naïve Bayes Classifier adalah suatu metode klasifikasi yang berdasarkan atau mengacu kepada teorema bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuan inggris Thomas Bayes, yitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal dengan teorema Bayes[11]. Naïve Bayes Classifier ini memiliki ciri utama yaitu bersifat asumsi yang kuat (naif) dan independensi dari masing masingkondisi atau kejadian.

Menurut[12] Naive Bayes memiliki beberapa kelebihan seperti modelnya mudah dibuat, memiliki perthitungan yang cepat dan efisien, tidak terlalu membutuhkan data yang besar atau banyak, dan bisa dilakukan dengan data training yang sedikit, dsb. Dengan menggunakan data yang kecil metode klasifikasi ini akan bisa mendapatkan nilai evaluasi yang cukup besar[12], yang dimana memiliki persamaan(1) [12].

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

Dimana :

X: Data dengan class yang belum diketahui
Y: Hipotesis data merupakan suatu class spesifik
P(Y|X) : Probabilitas hipotesis Y berdasarkan kondisi X (posteriori probabilitas)
P(Y): Probabilitas hipotesis Y (prior probabilitas)
P(X|Y) : Probabilitas X berdasarkan kondisi YP(X) : Probabilitas X

Persamaan diatas baru menunjukan teorema Bayes, oleh karena itu metode Naïve bayes dijelaskan dengan cara perlunya dikaetahui kelas mana saja yang sesuai dengan sampel yang akan dianalisis dalam klasifikasi.

Berdasarkan [12], persamaan diatas akan di sesuaikan dengan:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

Yang dimana C merupakan kelas, dan variabel F1 ... Fn merupakan karakteristik petunjuk yang diperlukan untuk melakukan proses klasifikasi. Setelah itu untuk melakukan proses klasifikasi dengan tipe data kontinyu rumus yang digunakan adalah rumus klasifikasi Densitas Gauss[12], yaitu:

$$P(Xi = xi\,|Y = yi) = \frac{1}{\sqrt{2\pi\sigma ij}} e - \frac{(x-\mu)^2}{2\sigma^2 ij} \quad (3)$$

Dimana:
Xi : Atribut ke-i
xi : Nilai atribut ke-i
Y : Kelas yang dicari
yi : Sub kelas Y yang dicari
$\mu$ : mean, melambangkan nilai rata-rata dari seluruh atribut
$\sigma$ : Standar deviasi, melambangkan varian atau jenis dari seluruh atribut

E. Konten-based

Berdasarkan[2] Kontent_based merupakan atribut atau fitur yang bertipekan boolean yang mana fiturnya berisikan isi suatu tweet seperti apakah tweet tersebut mengandung contain_location, contain_video, contain_picutre, contain_number, contain_url, dsb.

F. Time-based

Menurut penelitian[2] Time_based adalah fitur atau atribut yang rata-rata bertipekan boolean yang dimana fitur tersebut mengacu ke waktu ataupun tanggal tweet tersebut dibuat. Seperti apakah tweet tersebutdi buat pada malam, siang ataupun pagi hari, apakah saat liburan, dsb.

G. Evaluasi Sistem

Untuk melakukan perhitungan performansi atau evaluasi sistem dari metode yang sudah ditentukan, untuk mendapatkan ukuran ketepatan dari data yang digunakan akan digunakan F-score, lalu dari ketepatan data tersebut akan digunakan untuk menentukan keseimbangan nilai antara Precision Recall[13] yang akan ditunjukan dengan confusion matrix sebagai berikut :

Tabel 1. Confusion Matrix

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | T P | FN |
| Actual Negative | F P | TN |

Yang dimana True Positive (TP) merupakan nilai positif yang diprediksi dengan benar, False Negative (FN) merupakan nilai positif yang dianggap nilai negative, False Positive (FP) merupakan nilai negative yangdianggap sebagai nilai positif dan yang terahir True Negative (TN) yang merupakan nilai negative yang diprediksi dengan benar. Selain itu berdasarkan confusion matriks yang sudah dibuat dapat juga diperoleh perhitungan performansi model berupa :

a. Akurasi

Akurasi merupakan penggambaran dari suatu pengklasifikasian atau singkatnya seberapa akurat model dapat mengklasifikasikan dengan benar. Akurasi juga merupakan rasio benar dari keseluruhan data yang dimana akurasi merupakan tingkat kedekatan nilai prediksi dengan nilai actual. Akurasi dapat diperoleh dengan persamaan :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

b. Precision

Precision menggambarkan tingkat keakuratan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Bisa dibilang juga bahwa precision merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Nilai dari precision ini dapat diperoleh dengan persamaan :

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

c. Recall

Recall menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi, yang dimana recall merupakan rasio prediksi benar positif yang dibandingkan dengan keseluruhan data yang benar positif. Nilai dari recall dapat diperoleh dengan persamaan :

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

d. F-score

F-Score merupakan nilai rata-rata dari Precision dan Recall untuk mendapatkan nilai Precisiondan Recall yang seimbang. Nilai atau bobot F-score bisa didapatkan dari persamaan :

$$F - Score = \frac{2(Precision\ x\ Recall)}{Precision + Recall} \quad (5)$$

## IV. HASIL DAN PEMBAHASAN

### H. Hasil Pengujian

Penelitian ini menggunakan K-fold cross validation untuk membagi data menjadi data latih dan data uji dengan k=5. Hal ini dilakukan agar mengetahui performansi yang berupa precision, recall, f1-score dan akurasi dari metode klasifikasi Naïve Bayes. Berikut ini berupakan tabel hasil pengujian yang sudah dilakukan:

Tabel 2. Hasil Pengujian Dengan K=5

| K-Fold | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| K-Fold 1 | 73,15% | 54,84% | 49,43% | 67% |
| K-Fold 2 | 49,19% | 49,66% | 42,81% | 53,57% |
| K-Fold 3 | 73,52% | 59,17% | 56,61% | 68,36% |
| K-Fold 4 | 64,37% | 58,37% | 55,56% | 62,24% |
| K-Fold 5 | 65,07% | 56% | 47,99% | 55,61% |
| Rata-rata | 65,06% | 55,61% | 50,49% | 61,36% |

Dari table diatas dapat diketahui bahwa hasil yang didapatkan tidak begitu bagus dengan rata-rata akurasi 61,36%, rata-rata precision yang didapatkan sebesar 65,06%, rata-rata untuk recall sebesar 55,61%, lalu rata-rata untuk f1-score sebesar 50,49%. Hal tersebut menunjukan bahwa data yang digunakan pada penelitian ini tidak *balance* atau tidak seimbang, maka dari itu akan dilakukan perbandingan dengan metode Oversampling dan Undersampling.

### I. Analisis Hasil Pengujian

Tabel 3. Hasil Pengujian Dengan Undersampling

| K-Fold | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| K-Fold 1 | 76,70% | 57,15% | 55,36% | 73,80% |
| K-Fold 2 | 44,36% | 47,81% | 36,67% | 44,31% |
| K-Fold 3 | 70,62% | 58,46% | 57,05% | 70,65% |
| K-Fold 4 | 68,41% | 57,15% | 51,56% | 61,07% |
| K-Fold 5 | 46,49% | 44,29% | 26,69% | 27,54% |
| Rata-rata | 61,32% | 52,97% | 45,46% | 55,48% |

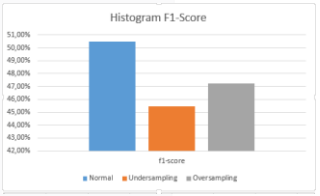
Tabel 4. Hasil Pengujian Dengan Oversampling

| K-Fold | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| K-Fold 1 | 72,82% | 56,39% | 48,65% | 59,73% |
| K-Fold 2 | 50,03% | 50,01% | 42,40% | 53,33% |
| K-Fold 3 | 72% | 58,98% | 55,53% | 65,77% |
| K-Fold 4 | 56,88% | 55,78% | 49,90% | 56,88% |
| K-Fold 5 | 59,35% | 55,46% | 39,70% | 41,33% |
| Rata-rata | 63,20% | 55,32% | 47,24% | 55,41% |

Pada table 3 dengan menggunakan metode klasifikasi naïve bayes dan split data dengan k-fold cross validation mendapatkan performansi dengan nilai rata-rata akurasi 61,36%, rata-rata precision yang didapatkan sebesar 65,06%, rata-rata untuk recall sebesar 55,61%, lalu rata-rata untuk f1-score sebesar 50,49%. Hal tersebut menunjukan bahwa data yang digunakan pada penelitian ini tidak *balance* atau tidak seimbang, maka dari itu akan dilakukan Oversampling dan Undersampling.

Undersampling merupakan sebuah proses untuk membuat dataset dalam suatu penelitian menjadi seimbang dengan cara membuang data dari label atau kelas yang lebih banyak (mayoritas) dibandingkan datayang lebih sedikit (minoritas). Setelah dilakukannya undersampling, seperti yang ada pada table 4, hasil performansi yang didapatkan cenderung mendurun dengan rata-rata nilai akurasi 55,48%, precision 61,32%, recall 52,97%, dan f1-score sebesar 45,46%. Pada table 5, didapatkan hasil performansi setelah melakukan oversampling. Oversampling sendiri merupakan metode untuk membuat dataset menjadi seimbang dengan cara menambahkan data yang minoritas dengan memilih data yang mayoritas secara acak agar jumlah kelas datanya menjadi sama banyak. Setelah dilakukannya metode tersebut didapatkan nilai rata-rata akurasi sebesar 55,41%, precision 63,20%, recall 55,32% dan f1-score sebesar 47,24%. Oleh karena itu dapat disimpulkan bahwa dengan melakukan metode oversampling dan undersampling tidak membuat perfomansi menjadi lebih baik melainkan membuatnya menjadi menurun jika dibandingkan dengan performansklasfiksai tanpa menggunakan kedua metode tersebut. Hal ini juga bisa disebabkan oleh minimnya fitur yang tersedia atau unik karena dalam satu akun twitter hanya sedikit fitur yang berbeda dengan yang lainnya.

Gambar 3. Histogram F1-Score




## V. KESIMPULAN

Dari penelitian yang sudah dilakukan dapat disimpulkan bahwa performansi yang didapatkan dengan menggunakan klasifikasi naïve bayes dan k-fold cross validation dengan k=5 cukup bagus dengan nilai rata-rata yang terdapat pada tabel 3 didapatkan rata-rata nilai f1-score sebesar 50,49%, rata rata akueasi sebesar 61, 36%, rata rata recall sebesar 55,61% dan rata rata precision sebesar 65,06% dengan menggunakan fitur kontent-based (berbasis konten), yang berupa isi dari tweet, dan time-based yang berupa waktu dibuatnya tweet tersebut.

Untuk penelitian selanjutnya penulis menyarankan untuk menggunakan dataset yang lebih variatif seperti akun twitter yang satu dengan yang lainnya dengan jumlah user yang lebih banyak dan menggunakan metodeklasifikasi yang lain.

## REFERENSI

[1] R. E. Putri, Suparti, and R. Rahmawati, “Perbandingan Metode Klasifikasi Naãve

Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012," *J. Gaussian*, vol. 3, no. 4,pp. 831–838, 2014.

[2] L. Hong and B. D. Davison, "[2011WWW]Predicting popular messages in Twitter.pdf," pp. 57–58,2011.

[3] T. B. N. Hoang and J. Mothe, "Predicting information diffusion on Twitter – Analysis of predictivefeatures," *J. Comput. Sci.*, vol. 28, pp. 257–264, 2018, doi: 10.1016/j.jocs.2017.10.010.

[4] L. Binarwati, I. Mukhlash, and S. Soetrisno, "Implementasi Algoritma Genetika untuk Optimalisasi Random Forest dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru : Studi Kasus PT.XYZ," *J.Sains dan Seni ITS*, vol. 6, no. 2, pp. 2–6, 2017, doi: 10.12962/j23373520.v6i2.26887.

[5] M. Athaillah, Y. Azhar, and Y. Munarko, "Perbandingan Metode Klasifikasi Berita Hoaks Berbahasa Indonesia Berbasis Pembelajaran Mesin," *J. Repos.*, vol. 2, no. 5, p. 675, 2020, doi: 10.22219/repositor.v2i5.692.

[6] F. Fanesya, R. C. Wihandika, and Indriati, "Deteksi Emosi Pada Twitter Menggunakan Metode NaïveBayes Dan Kombinasi Fitur," vol. 3, no. 7, pp. 6678–6686, 2019.

[7] M. C. Wijanto, "Sistem Pendeteksi Pengirim Tweet dengan Metode Klasifikasi Naïve Bayes," *J.Tek.Inform. dan Sist. Inf.*, vol. 1, no. 2, pp. 172-182,2015, doi: 10.28932/jutisi.vli2.378.

[8] S. Suryono, E. Utami, and E. T. Luthfi, "Klasifikasi Sentimen Pada Twitter Dengan Naive Bayes.

[9] S. Syarli and A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data MahasiswaBaru Perguruan Tinggi)," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2016.

[10] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, "Data Crawling Otomatis pada Twitter,"no. August, pp. 11–16, 2016, doi: 10.21108/indosc.2016.111.

[11] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Creat. Inf. Technol. J.*, vol. 2, no. 3, pp. 207–217, 2015.

[12] Alvina Felicia Watratan, Arwini Puspita. B, and Dikwan Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *J. Appl. Comput. Sci. Technol.*, vol. 1,no. 1, pp. 7–14, 2020, doi: 10.52158/jacost.v1i1.9.

[13] I. Maulida, A. Suyatno, H. Rahmania Hatta, and U. Mulawarman, "Seleksi Fitur Pada Dokumen AbstrakTeks Bahasa Indonesia Menggunakan Metode Information Gain," *Oktober 2016 Ijccs*, vol. 17, no. 2, pp.1–5, 2016.

, 1993, pp. 123-35