Prediksi *Retweet* Berbasis Konten dan Pengguna dengan Metode Klasifikasi *Naïve Bayes*

Tasya Maulasirri Sutisna[1], Jondri[2], Kemas Muslim Lhaksamana[3]

[1,2,3] Universitas Telkom, Bandung
[1]tasyamaulasrrs@students.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id, [3]kemasmuslim@telkomuniversity.ac.id

Abstrak

Twitter merupakan salah satu media sosial yang paling banyak diminati oleh berbagai kalangan untuk berinteraksi antar pengguna. Selain itu, Twitter dimanfaatkan untuk mendapatkan informasi terbaru mengenai suatu produk, isu sosial-politik, selebriti, dan sebagainya. Pengguna Twitter dapat memposting kembali informasi tersebut dengan cara meretweet *tweet* sehingga informasinya dapat tersebar lagi ke pengguna lain. Penelitian ini bertujuan untuk membangun sistem prediksi *retweet* dan melihat bagaimana performansi serta akurasi dari metode klasifikasi Naïve Bayes dengan dua model fitur yaitu berbasis konten dan berbasis pengguna. Pembagian dataset menggunakan *k-fold cross validation* dengan nilai k = 10. Hasil yang didapatkan pada penelitian ini cukup baik yaitu dengan tingkat akurasi 76,41%, tingkat *precision* 76,72%, tingkat *recall* 99,41% , dan *F1-score* 86,42%.

Kata kunci : Twitter, tweet, retweet, Naïve Bayes, k-fold cross validation, klasifikasi

Abstract

Twitter is one of the most popular social media to interact with users. In addition, Twitter is used to get the latest information about a product, socio-political issues, celebrities, and so on. Twitter users can repost the information by retweeting a tweet so that the information can be spread again to other users. This study aims to build a retweet prediction system and see how the performance and accuracy of the Naïve Bayes classification method are based on two feature models, content-based and user-based. Dataset splitting uses k-fold cross validation with a value of k = 10. The results obtained in this study are good enough with an accuracy rate of 76.41%, precision rate of 76.72%, recall rate of 99.41% , and f1-score of 86.42%.

Keywords: Twitter, tweet, retweet, Naïve Bayes, k-fold cross validation, classification

1. Pendahuluan

Latar Belakang

Di era digital ini, perkembangan teknologi informasi dan komunikasi yang pesat telah memberikan banyak manfaat di semua aspek sosial. Salah satu perkembangan teknologi informasi dan komunikasi yaitu dengan adanya media sosial yang digunakan untuk bertukar informasi terkini dan memberikan kesempatan kepada pengguna untuk menghasilkan sebuah konten [1]. Salah satu media sosial yang banyak diminati oleh berbagai kalangan saat ini yaitu Twitter.

Twitter adalah layanan *microblog* popular yang menarik lebih dari 500 juta pengguna dan menghasilkan lebih dari 340 juta *tweet* setiap harinya [2]. *Tweet* yang diposting dapat berupa tulisan, gambar atau video yang setiap *tweet* nya dibatasi hanya 280 karakter saja [3]. Tidak heran jika Twitter banyak dimanfaatkan untuk berbagai keperluan seperti, mendapatkan informasi terbaru tentang suatu produk, isu sosial, politik, selebriti dan banyak lagi [3]. Berbagai manfaat yang diberikan oleh Twitter telah menjadikannya media sosial yang efektif dan efisien dalam menyampaikan informasi dengan cepat [4]. Selain itu, pengguna Twitter dapat mempublikasikan kembali informasi tersebut dengan cara me-*retweet* atau memposting ulang *tweet*, menyukai postingan *tweet*, mengutip *tweet* dan juga dapat menjadi pengikut dari pengguna lain atau diikuti oleh banyak pengguna [3]. Sehingga informasi tersebut dapat tersebar lagi di Twitter. Pemodelan difusi informasi yang berkembang ini sangat penting agar dapat dipahami penyebarannya dan dapat mengendalikannya dengan baik [5].

Pada penelitian sebelumnya, menerapkan pembelajaran mesin menggunakan metode pengklasifikasian *random forest* dengan tiga jenis fitur: berbasis pengguna, berbasis waktu dan berbasis konten [5]. Penelitian serupa mengenai prediksi *retweet* dengan fitur berbasis sosial, berbasis konten, berbasis *tweet* dan berbasis pengguna dengan membandingkan metode klasifikasi *decision tree* J48, SVM dan regresi logistik [6]. Berdasarkan pernyataan diatas penulis tertarik untuk melakukan penelitian tentang bagaimana prediksi pengguna Twitter yang tertarik untuk melakukan *retweet* suatu konten isu yang sedang berkembang. Penelitian penulis berfokus pada penyebaran informasi di media sosial atau lebih tepatnya untuk memprediksi apakah suatu *tweet* akan mendapatkan *retweet* atau tidak. Fitur yang dipakai pada penelitian tugas akhir ini menggunakan dua model

yaitu fitur pengguna dan konten. Metode yang digunakan adalah Naïve Bayes yang memiliki kesalahan paling sedikit dibandingkan dengan algoritma klasifikasi lainnya [7]. Metode ini memiliki perhitungan matematik dasar yang sangat kuat dan dalam efisiensi klasifikasinya juga stabil serta dapat mengolah data dalam jumlah yang besar dengan menghasilkan akurasi yang tinggi [7].

Topik dan Batasan

Berdasarkan latar belakang yang telah disampaikan sebelumnya, maka rumusan masalah yang dapat diambil yaitu memprediksi apakah suatu *tweet* akan mendapatkan *retweet* atau tidak dengan fitur berbasis konten dan berbasis pengguna. Serta melihat performansi dan akurasi dari metode Naïve Bayes dalam prediksi *retweet*. Sedangkan batasan masalah dari penelitian ini adalah penelitian ini menggunakan data *tweet* pengguna Twitter yang diambil pada bulan April hingga Mei 2021.

Tujuan

Tujuan dari penelitian tugas akhir ini adalah membangun sistem prediksi *retweet* berbasis konten dan berbasis pengguna menggunakan metode klasifikasi Naïve Bayes.

Organisasi Tulisan

Selanjutnya akan dijelaskan mengenai studi terkait penelitian yang serupa dan hasil penelitian tersebut. Pada bagian tiga akan dijelaskan sistem yang dibangun pada penelitian, yang menjelaskan mengenai gambaran sistem yang diterapkan beserta teori-teori yang terkait. Pada bagian empat evaluasi, akan dijelaskan mengenai dataset yang digunakan, hasil pengujian dengan perhitungan akurasi, *precision*, *recall* dan *F1-score*, kemudian analisis dari hasil pengujian. Dan pada bagian terakhir adalah kesimpulan dari hasil penelitian ini serta saran untuk penelitian selanjutnya.

2. Studi Terkait

Penelitian ini berdasarkan beberapa penelitian terkait sebagai bahan perbandingan. Berdasarkan penelitian serupa sebelumnya, penelitian yang bertujuan untuk memprediksi apakah sebuah postingan *tweet* akan di *retweet* atau tidak. Selain itu, untuk memprediksi seberapa luas informasi tersebut akan tersebar. Menggunakan tiga jenis fitur: berbasis pengguna, berbasis waktu, dan berbasis konten menggunakan metode klasifikasi *Random Forest.* Hasil penelitian ini meningkat sekitar 5% *F-measure* secara statistik dibandingkan dengan keadaan yang signifikan dengan total 16 juta *tweet*. Model *F-measure* yang digunakan antara 70% dan 82%. Fitur pengguna menjadi fitur yang paling penting pada penelitian tersebut [5].

Selanjutnya, penelitian serupa tentang analisis perilaku pengguna *retweet* di Twitter yang bertujuan untuk memperkirakan apakah *tweet* akan *diretweet* oleh pengguna tertentu dengan memanfaatkan empat jenis fitur yang berbeda yaitu: berbasis sosial, berbasis konten, berbasis *tweet* dan berbasis pengguna dengan membandingkan metode klasifikasi *Decision Tree* J48, SVM dan regresi logistik. Untuk mengetahui faktor-faktor apa saja yang yang penting dalam memprediksi perilaku *retweet* individu, penelitian ini melakukan perbandingan *"leave one feature out"*. Hasilnya mengungkapkan fitur yang terkait dengan pengguna cukup penting dalam memprediksi *retweet* terutama fitur berbasis sosial [6].

Penelitian-penelitian tersebut menggunakan metode yang berbeda-beda dan mendapatkan hasil yang cukup baik. Pada penelitian ini penulis akan mengimplementasikan metode klasifikasi Naïve Bayes untuk memprediksi *retweet* menggunakan fitur berbasis konten dan berbasis pengguna.

2.1. Twitter

Media sosial Twitter berisi pesan singkat yang diposting pengguna melalui situs *microblogging* [4]. Isi teks pada Twitter dapat terdiri dari angka, huruf ataupun simbol dengan struktur kalimat yang bebas sesuai dengan keinginan pengguna [4]. Pengikut dari pengguna Twitter dapat melihat postingan yang disampaikan pengguna Twitter kepada pengguna lainnya yang disebut dengan istilah *tweet* [4]. Empat bagian postingan yang paling umum digunakan di Twitter antara lain *hashtag*, *mention*, *retweet* dan *replies* [8]. *Hashtag* # digunakan untuk menyoroti wacana orang lain, *mention* (@ diikuti nama pengguna) menandakan pesan kepada orang lain dan membagikan informasi kepada orang lain [8], *retweet* adalah pesan dari satu pengguna yang diteruskan oleh pengguna lain dengan cara menekan fitur *retweet* dan *replies* merupakan percakapan antar pengguna dengan cara membalas *tweet* dari pengguna lain [1].

Twitter menyediakan cara untuk mengakses dan memperoleh data *tweet* melalui Twitter API [4]. Metode ini bersifat terbuka dan dapat diakses oleh publik dengan syarat dan ketentuan yang ditetapkan oleh Twitter, seperti batasan jumlah *tweet* yang dapat diambil, rentang waktu data yang dapat diambil dan sebagainya [4]. Twitter API menyediakan akses ke data *tweet* untuk rentang waktu tertentu, pengguna tertentu, kata kunci tertentu atau area geografis tertentu, tetapi tidak menyediakan fungsi mengekstraksi struktur dari *tweet*, serta

tidak menyediakan data agregat Twitter yang merangkum berbagai topik (contoh, frekuensi *tweet* pada topik tertentu selama periode waktu tertentu) [4].

2.2. Naïve bayes

Naïve Bayes adalah sebuah metode klasifikasi menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas Bayes. Algoritma Naïve Bayes memprediksi peluang dimasa depan berdasarkan pengalaman di waktu sebelumnya sehingga dikenal dengan Teorema Bayes [9]. Ciri utama dari Naïve Bayes ini adalah asumsinya yang kuat (naif) tidak bergantung pada setiap kondisi [10]. Menurut Olson Delen (2008) probabilitas adalah vektor informasi dari suatu objek tertentu dengan syarat bahwa kelas keputusannya benar [10]. Algoritma Naïve Bayes mengasumsikan bahwa atribut obyeknya tidak saling ketergantungan atau independen yang diberikan oleh nilai variabel kelas [9]. Probabilitas yang terlibat dalam prediksi akhir dihitung sebagai jumlah frekuensi dalam tabel keputusan utama. Jika dibandingkan dengan pengklasifikasian yang lain, Naïve Bayes berkinerja sangat baik [10]. Keuntungan menggunakan metode Naïve Bayes adalah metode ini hanya membutuhkan sedikit data latih untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi [9].

Persamaan Teorema Bayes ditunjukan sebagai berikut (1) [9].

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

Keterangan:
X : data dengan kelas yang belum diketahui
C : hipotesis data merupaan suatu kelas spesifik
P(*C*|*X*) : probabilitas hipotesis *C* berdasar kondisi *X* (*posterior* probabilitas)
P(*X*|*C*) : probabilitas *posterior X* berdasarkan kondisi pada hipotesis *C*
P(*C*) : probabilitas *prior* hipotesis *C* (*prior* probabilitas)
P(*X*) : probabilitas *X*

Rumus diatas menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (*posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, disebut *prior*) dikali dengan peluang kemunculan karakteristik-klarakteristik sampel pada kelas C (disebut *likelihood*), dibagi dengan peluang kemuculan karakteristik secara global (disebut *evidence*) [9]. Dapat ditulis dengan rumus:

$$posterior = \frac{prior\ x\ likelihood}{evidence} \quad (2)$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *posterior* tersebut nantinya akan dibandingkan dengan nilai-nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan [9]. Disini diasumsikan indenpendensi yang sangat tinggi, yaitu setiap petunjuk saling bebas satu sama lain [9]. Dengan asumsi tersebut:

$$P(X_i|X_j) = \frac{P(X_i \cap X_j)}{P(X_j)} = \frac{P(X_i)P(X_j)}{P(X_j)} = P(X_i) \quad (3)$$

Untuk i ≠ j , sebagai berikut

$$P(X_i|C, X_j) = P(X_i|C) \quad (4)$$

Persamaan diatas digunakan dalam proses klasifikasi yang merupakan model dari teorema Naïve Bayes. Untuk pengklasifikasian dengan atribut bertipe kontinyu akan digunakan rumus *Densitas Gauss* yang ditunjukan dalam persamaan (5) [9].

$$P(X_i = x_i \mid C = c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma ij^2}} \quad (5)$$

Keterangan:
P : peluang
X_i : atribut ke *i*
x_i : nilai atribut ke *i*
C : kelas data

C_j: kelas ke j
μ_{ij} : *mean*, rata-rata atribut ke i pada kelas j
σ_{ij} : deviasi standar, atribut ke i pada kelas j

2.3. Metode *Resampling*

Metode *resampling* bekerja untuk menyeimbangkan data sehingga data yang akan digunakan sama besar jumlahnya [11]. Metode ini digunakan untuk menangani ketidakseimbangan data di dalam kelas. Dengan menerapkan sampling pada data, tingkat ketidakseimbangannya semakin kecil, dan pengklasifikasian dapat dilakukan dengan baik [11]. Terdapat dua cara sampling yaitu *Random Oversampling* dan *Random Undersampling*.

Pada metode *random oversampling*, dilakukan penyeimbangan jumlah data dengan cara menambah jumlah data pada kelas minoritas. Langkah yang dilakukan *random oversampli*ng, pertama menghitung selisih data antara kelas mayoritas dan minoritas. Kemudian melakukan perulangan sebanyak hasil selisih dan menambahkan data secara acak pada kelas minoritas sampai jumlah data sama dengan kelas mayoritas. [11].

Pada metode *random undersampling*, dilakukan dengan menyeimbangkan data dengan cara mengurangi jumlah data pada kelas mayoritas. Langkah pertama yang dilakukan, menghitung selisih antara kelas mayoritas dan minoritas, jika terdapat selisih pada data kelas maka jumlah data kelas mayoritas akan dihapus secara acak sampai jumlahnya sama dengan jumlah kelas minoritas [12] .

2.4. Penggunaan Fitur

Untuk melakukan prediksi *retweet* diperlukan fitur yang dapat berpengaruh pada *tweet* yang akan di*retweet.* Fitur yang digunakan pada penelitian ini fitur berbasis pengguna dan berbasis konten. Fitur ini yang paling berdampak pada *tweet* untuk mendapatkan *retweet* [3]. Fitur berbasis pengguna merupakan fitur utama dalam meprediksi *retweet* dari sudut pandang global [5]. Pengguna juga merupakan orang yang berinteraksi dan akan mendapatkan perhatian dari pengguna lain di Twitter melalui *tweetnya* [6]. Berikut fitur berbasis pengguna yang digunakan:

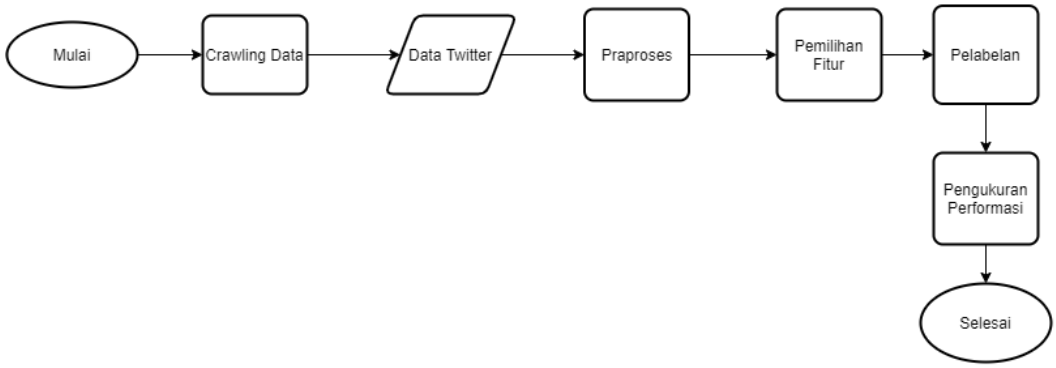
- Jumlah *followers*: jumlah orang yang mengikuti pengguna. Semakin banyak pengikut yang pengguna miliki maka semakin tinggi peluang *tweet* untuk di *retweet* [3].
- Jumlah *following*: jumlah orang yang diikuti oleh pengguna. Semakin banyak pengguna yang diikuti maka semakin tinggi peluang pengguna me-*retweet* suatu *tweet*.
- Total *tweet*: total *tweet* yang diposting oleh pengguna.
- Akun *verified*: akun pengguna terverifikasi atau tidak. Akun terverifikasi merupakan akun yang berpengaruh di Twitter.
- Akun *protected*: akun pengguna dikunci atau tidak.

Fitur berbasis konten merupakan informasi yang terkandung pada suatu *tweet* yang dapat menarik perhatian pengguna lain [6]. Pada berbasis konten fitur URL dan *hashtag* memiliki hubungan yang kuat dalam mendapatkan *retweet* [2]. Berikut fitur berbasis konten yang digunakan:

- *Hashtag*: *tweet* ini mengandung tagar tertentu. *Hashtag* ini adalah kata kunci yang digunakan untuk menandai *tweet* sehingga lebih mudah dikategorikan dan ditemukan oleh pengguna [3].
- URL: *tweet* ini berisi URL tertentu. URL ini biasanya dicantumkan untuk membagikan informasi penting seperti berita.
- Media: pengguna *tweet* sering melampirkan media agar *tweet*nya menjadi lebih menarik, media yang dicantumkan seperti gambar atau video.
- *Mention*: *tweet* yang dibuat pengguna menyebutkan nama pengguna lain.
- *Quote*: *tweet* yang dibuat pengguna di kutip kembali oleh pengguna lain.
- Jumlah *like*: jumlah pengguna yang menyukai *tweet* ini.

3. Sistem yang Dibangun

Pada bagian ini, akan dijelaskan mengenai gambaran sistem yang diterapkan pada penelitian tentang prediksi *retweet* dengan menggunakan metode klasifikasi Naïve Bayes. Data *tweet* pengguna diperoleh dari Twitter API.



Gambar 1. Alur sistem yang dibangun

1. Crawling data
   Tahap crawling data merupakan pengumpulan data dari suatu database [13]. Pengumpulan data didapatkan dengan cara mengunduh secara otomatis dari Twitter API. Hasil dari tahap ini yaitu data dari pengguna dan data pada *tweetnya*.
2. Praproses
   Tahap praproses ini dilakukan sebelum proses klasifikasi ataupun prediksi. Pada penelitian tugas akhir ini tahap praproses yang dilakukan yaitu menyiapkan data agar diproses dengan cara menghilangkan duplikat data.
3. Pemilihan fitur
   Tahap ini dilakukan pemilihan fitur yang penting atau relevan terhadap data dan menghilangkan yang tidak relevan. Hal ini dapat meningkatkan kinerja algoritma klasifikasi. Fitur yang digunakan pada penelitian ini yaitu *username*, *tweet*, *hashtag*, *mention*, url, media, *tweet* yang di *quote* oleh pengguna lain, jumlah pengikut, jumlah akun yang diikuti, verifikasi akun, akun *protected*, total *tweet*, jumlah *like*, dan *retweet*.
4. Pelabelan
   Pada tahap ini dilakukan pemberian label kelas *retweet* dengan *tweet* yang memiliki *retweet* kelas 1 dan yang tidak memiliki *retweet* kelas 0.
5. Pengukuran performasi
   Tahap ini merupakan tahapan terakhir, yaitu akan dilakukan perhitungan akurasi, *recall*, *precision* dan *F1-score*. Perhitungan ini digunakan untuk mengukur akurasi dan performasi dari model klasifikasi yang telah dibangun. Penjelasan mengacu pada tabel *confusion matrix* berikut

Tabel 1. *confusion matrix*

Kelas Sebenarnya	Kelas Prediksi	
	True	*False*
True	*True Positive* (TP)	*False Positive* (FP)
False	*True Negative* (TN)	*False Negative* (FN)

True Positive (TP) adalah model memprediksi data ada di kelas positif dan data memang ada di kelas positif. *True Negative* (TN) adalah model memprediksi data ada di kelas negatif dan memang ada di kelas negative. *False Positive* (FP) adalah model memprediksi data ada di kelas postif tetapi sebenarnya data ada di kelas negative. *False Negative* (FN) model memprediksi data ada di kelas negatif tetapi sebenarnya ada di kelas positif.

- Akurasi

Akurasi didefinisikan sebagai tingkat kedekatan nilai prediksi dengan nilai aslinya. Semakin besar nilai akurasi maka semakin baik performansinya. Berikut persamaan akurasi :

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

- *Recall*
  Recall merupakan rasio dari jumlah total positif yang diklasifikasikan dibagi dengan jumlah total positif [14]. Berikut persamaan *recall* :

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

- *Precision*
  Precision adalah rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total prediksi yang diklasifikasikan ke dalam kelas. *Precision* membagi jumlah total positif yang diklasifikasikan dengan jumlah total positif [14]. Berikut persamaan *precision* :

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

- F1-Score
  F1-Score atau F-Measure adalah penyetaraan nilai *precision* dan *recall* menjadi sebuah rumus tunggal. Berikut persamaan F1-Score :

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (10)$$

4. Evaluasi

Bagian ini berisi hasil pengujian dan analisis dari hasil pengujian sistem prediksi *retweet* yang telah dibangun. Pengujian dan analisis yang dilakukan selaras dengan tujuan yang dinyatakan pada bagian pendahuluan.

4.1 Dataset

Dataset yang digunakan untuk penelitian ini diambil dengan cara crawling data *tweet* pengguna menggunakan dua cara yaitu *by keyword* dan *by timeline* dengan total 7163 *tweet* pada bulan April hingga Mei 2021. Data *tweet* yang diambil, yaitu *username* dan *tweet* sebagai informasi, kemudian jumlah *followers*, jumlah *following*, jumlah *like*, jumlah *retweet*, total *tweet* pengguna, akun terverifikasi, akun di *protect*, media, URL, *hashtag*, *mention* pengguna, dan *tweet* dikutip sebagai fiturnya. Dalam dataset ini atribut *retweet* yang menjadi kelas klasifikasi. Berikut atribut yang digunakan

Tabel 2. Atribut pada dataset

	Atribut	Keterangan	Tipe data
Berbasis pengguna	*Followers*	Jumlah orang yang mengikuti pengguna	Numerik
	Following	Jumlah orang yang pengguna ikuti	Numerik
	Total Tweet	Jumlah tweet pengguna	Numerik
	Verified	Akun yang terverifikasi	Boolean
	Protected	Akun yang dikunci	Boolean
Berbasis konten	Jumlah like	Jumlah like dari suatu tweet	Numerik
	Hashtag	*Tweet* ini berisi *hashtag* tertentu	Boolean
	Media	*Tweet* ini berisi gambar dan video	Boolean

	URL	*Tweet* ini berisi URL	Boolean
	Is_quote_status	*Tweet* ini dikutip oleh pengguna lain	Boolean
	Mention	*Tweet* ini berisi mention pengguna lain	Boolean

4.2 Hasil Pengujian

Pengujian dari penelitian ini menggunakan metode *k-fold cross validation* dimana satu dataset utuh dibagi sejumlah bagian k. Pada bagian tersebut 1 bagian sebagai data test dan bagian yang lain sebagai data train. Nilai k yang digunakan sebanyak 10, yang bertujuan untuk mengetahui hasil dari metode klasifikasi Naïve Bayes yang diterapkan pada prediksi *retweet*. Menggunakan k sebanyak 10 tujuannya untuk memvalidasi kinerja algoritma Naïve Bayes agar lebih teruji dan jumlah *fold* ini yang terbaik untuk uji validitas [15]. Evaluasi dihitung dengan mengukur akurasi, *precision*, *recall* dan *F1-score*. Berikut merupakan hasil pengujian dengan nilai *k-fold* sebesar 10:

Tabel 3. akurasi dari setiap fold pada k = 10

Fold	F1-Score	Precision	Recall	Akurasi
Fold 1	88,75%	79,77%	100%	79,77%
Fold 2	89,09%	80,55%	99,65%	80,33%
Fold 3	90,61%	83,19%	99,49%	82,84%
Fold 4	86,43%	76,22%	99,81%	76,11%
Fold 5	91,10%	83,77%	99,83%	83,65%
Fold 6	90,01%	81,95%	99,82%	81,84%
Fold 7	86,34%	76,19%	99,63%	75,97%
Fold 8	86,79%	77,43%	98,74%	76,67%
Fold 9	81,65%	69,67%	98,60%	68,99%
Fold 10	73,38%	58,45%	98,57%	57,96%
Rata-rata	86,42%	76,72%	99,41%	76,41%

Hasil prediksi *retweet* menggunakan klasifikasi Naïve Bayes dengan nilai k = 10 seperti yang ditunjukan pada tabel 3, memberikan hasil rata-rata akurasi 76,41% , *recall* 99,41% , *precision* 76,72%, dan F1-*score* 86,42%.

4.3 Analisis Hasil Pengujian

Berdasarkan hasil pengujian yang telah dilakukan untuk prediksi *retweet* menggunakan klasifikasi Naïve Bayes menghasilkan performansi yang cukup baik dengan pengukuran akurasi mendapatkan hasil rata-rata sebesar 76,41%, Selanjutnya, pengujian dengan *precision* mendapatkan hasil rata-rata 76,72%, Pengujian dengan *recall* mendapatkan hasil rata-rata 99,41%, dan terakhir dengan pengukuran *F1-score* mendapatkan hasil rata-rata 86,42%. Nilai dari *precision* dan *recall* hasilnya timpang dikarenakan kecenderungan pada model, lebih banyak kelas 1 (*tweet* yang di *retweet*) dibandingkan kelas 0 (*tweet* yang tidak di *retweet*).

Pada dataset yang digunakan, jumlah data pada kelas klasifikasi terbagi secara tidak merata karena dipengaruhi oleh pengambilan data *tweet* yang secara acak. 77% untuk kelas *tweet* yang mendapatkan *retweet* dan 23% untuk kelas *tweet* yang tidak mendapatkan *retweet*. Untuk menangani ketidakseimbangan data tersebut, dilakukan metode *resampling* menggunakan dua metode, yaitu *random oversampling* dan *random undersampling.*

Metode *random oversampling* ini dimana data train menambahkan kelas minoritas agar sama dengan kelas mayoritas dengan cara menduplikasi data secara acak pada kelas minoritas. Berikut merupakan hasil dari pengujian *oversampling*:

Tabel 4. Hasil Pengujian *Oversampling*

Fold	F1-Score	Precision	Recall	Akurasi
Fold 1	65,23%	52,33%	86,55%	54,49%
Fold 2	66,39%	53,82%	86,60%	55,40%
Fold 3	66,57%	54,23%	86,18%	55,04%
Fold 4	67,52%	54,46%	88,80%	56,31%
Fold 5	65,57%	53,10%	85,68%	54,31%
Fold 6	63,93%	51,11%	85,34%	52,86%

Fold 7	65,70%	52,57%	87,59%	54,49%
Fold 8	65,53%	52,39%	87,47%	53,95%
Fold 9	63,87%	50,94%	85,60%	52,90%
Fold 10	64,21%	50,99%	86,70%	53,09%
Rata-rata	65,45%	52,59%	86,65%	54,28%

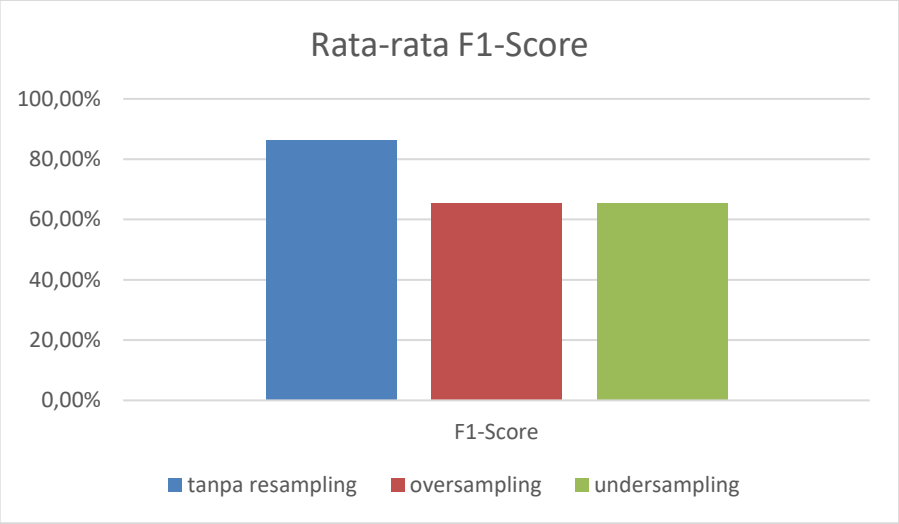
Pada tabel 4 menunjukan rata-rata akurasi yang didapat 54,28%, *recall* 86,65%, selanjutnya pengujian dengan *precision* mendapatkan rata-rata 52,59% dan rata-rata pada hasil *F1-score* 65,45%. Hasil dari *oversampling* ini mengalami penurunan dibandingkan tanpa *oversampling.*

Selanjutnya, melakukan pengujian dengan menerapkan *random undersampling*, dimana mengurangi beberapa data pada kelas mayoritas sehingga jumlahnya sama dengan jumlah data minoritas. Berikut hasil dari pengujian klasifikasi Naïve Bayes setelah dilakukan *undersampling*:

Tabel 5. Hasil Pengujian *Undersampling*

Fold	F1-Score	Precision	Recall	Akurasi
Fold 1	66,35%	52%	91,66%	56,32%
Fold 2	64,09%	50,17%	88,67%	52,40%
Fold 3	63,72%	51,11%	84,56%	53,01%
Fold 4	65,56%	53,46%	84,75%	56,02%
Fold 5	66,06%	53,09%	87,42%	54,81%
Fold 6	64,71%	52,55%	84,21%	52,71%
Fold 7	69,93%	57,74%	88,64%	57,53%
Fold 8	62,29%	49,62%	83,64%	51,50%
Fold 9	63,90%	50,72%	86,33%	52,56%
Fold 10	68,60%	56,45%	87,42%	57,70%
Rata-rata	65,52%	52,69%	86,73%	54,46%

Pada tabel 5 menunjukkan hasil rata-rata akurasi 54,46%, hasil rata-rata pada *precision* 52,69%, selanjutnya hasil rata-rata *recall* 86,73% dan hasil *F1-score* 65,52%. Pada pengujian dengan menerapkan *undersampling* tidak menunjukan hasil yang lebih baik dari pengujian *oversampling*.



Gambar 2. Rata-rata F1-Score pada setiap skenario

Gambar 2 menunjukan rata-rata hasil *F1-score* dari tiga skenario yang telah dilakukan, yaitu pengujian tanpa *resampling*, menerapkan *oversampling* dan *undersampling*. Hasilnya, pengujian tanpa *resampling* menunjukan nilai yang lebih baik yaitu dengan *F1-score* 86,42%. Sedangkan hasil dari *oversampling* dan *undersampling* menunjukkan nilai yang hampir sama yaitu 65,45% untuk hasil *oversampling* dan 65,52% untuk *undersampling*. Namun, hasil tersebut mengalami penurunan dari hasil klasifikasi tanpa menerapkan resampling.

5. Kesimpulan

Kesimpulan dari penelitian yang bertujuan untuk membangun sistem prediksi *retweet* apakah *tweet* yang diberikan akan mendapatkan *retweet* atau tidak. Serta hasil dari performansi dan akurasi yang didapatkan dengan menerapkan model pembelajaran mesin pengklasifikasian Naïve Bayes dengan fitur yang digunakan pada penelitian yaitu jumlah *followers*, jumlah *following*, jumlah *tweet*, jumlah *like*, *verified*, *protected*, *mention*, media, URL, *hashtag*, *quote* status. Metode klasifikasi Naïve Bayes yang digunakan pada penelitian mampu mencapai kinerja yang cukup baik dalam prediksi *retweet*. Hal ini bisa dilihat dari hasil akurasi yang didapat cukup baik dengan menggunakan 10 *fold cross validation*. Hasil rata-rata yang didapat dari akurasi 76,41%, *precision* 76,72%, *recall* 99,41%, dan *F1-score* 86,42%.

Untuk menangani ketidakseimbangan data pada kelas klasifikasi dilakukan pengujian dengan menerapkan *random oversampling* dan *random undersampling*, hasil klasifikasinya mengalami penurunan dari klasifikasi tanpa menerapkan *resampling*. Pada pengujian ini, menerapkan metode *resampling* tidak efektif dalam menyelesaikan ketidakseimbangan data kelas pada prediksi *retweet* dengan metode klasifikasi Naïve Bayes.

Penulis memberikan saran untuk penelitian selanjutnya menggunakan dataset dengan pembagian label kelas yang tidak cukup jauh antara kelas yang mendapatkan *retweet* dan tidak. Dan juga, dapat mengembangkan fitur baru seperti fitur berbasis waktu atau sosial dengan membandingkan hasil performansinya menggunakan metode klasifikasi lainnya.

REFERENSI

[1] H. Becker and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter (Tech Report).pdf," pp. 438–441, [Online]. Available: https://pdfs.semanticscholar.org/2573/060fb7b47e1a69933a28118fc9fd60c393ff.pdf.

[2] Z. Luo, M. Osborne, J. Tang, and T. Wang, "Who will retweet me? Finding retweeters in twitter," *SIGIR 2013 - Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, no. May 2015, pp. 869–872, 2013, doi: 10.1145/2484028.2484158.

[3] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," *WWW 2013 Companion - Proc. 22nd Int. Conf. World Wide Web*, pp. 657–664, 2013, doi: 10.1145/2487788.2488017.

[4] A. Wibowo and E. Winarko, "Paper Review: Data Mining Twitter," *Maint. Cult. Herit. Through Inf. Technol. a Smart Futur.*, no. November 2014, pp. 1–10, 2014, [Online]. Available: https://www.researchgate.net/publication/329207488_Paper_Review_Data_Mining_Twitter.

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

[6] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," *Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012*, pp. 46–50, 2012, doi: 10.1109/ASONAM.2012.18.

[7] Y. S. Nugroho, U. M. Surakarta, and C. Channels, "Prediksi Rating Film Menggunakan Metode Naive Bayes," *J. Tek. Elektro Unnes*, vol. 8, no. 2, pp. 60–63, 2016, doi: 10.15294/jte.v8i2.7764.

[8] A. Black, C. Mascaro, M. Gallagher, and S. P. Goggins, "Twitter zombie," p. 229, 2012, doi: 10.1145/2389176.2389211.

[9] 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.

[10] A. F. B. Watratan, A. Puspita, and D. Moeis, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *J. Appl. Comput. Sci. Technol. (Jacost)*, vol. 1, no. 1, pp. 7–14, 2020.

[11] S. S. Utomo, T. A. Cahyanto, and B. H. Prakoso, "Penggunaan Algoritma Random Over Sampling Untuk Mengatasi Masalah Imbalance Data Pada Klasifikasi Gizi," pp. 1–9, 2007.

[12] E. Irawan, "Penggunaan Random Under Sampling untuk Penanganan Ketidakseimbangan Kelas pada Prediksi Cacat Software Berbasis Neural Network," *J. Softw. Eng.*, vol. 1, no. 2, pp. 92–100, 2015.

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

[14] Y. Umar, Hanafi, S. Mardi, Nugroho, Susiki, and R. F. Rachmadi, "Deteksi Penggunaan Helm Pada Pengendara Bermotor Berbasis Deep Learning," 2020.

[15] U. S. Utara, "Universitas Sumatera Utara 4," pp. 4–16, 2003.