Prediksi Retweet Menggunakan Metode Bernoulli dan Gaussian Naive Bayes di Media Sosial Twitter Dengan Topik Vaksinasi Covid-19

Ika Puspita Dewi[1], Jondri[2], Kemas Muslim Lhaksmana[3]

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

Abstrak

Media sosial twitter adalah media sosial internasional yang mengizinkan pengguna untuk berbagi pesan atau biasa disebut *tweet* dengan maksimal 280 karakter per-*tweet*, dapat dilakukan secara publik maupun pribadi dengan pengguna lain. Twitter menyediakan berbagai informasi yang diperlukan mulai dari informasi kesehatan, pendidikan, olahraga, politik, makanan, dan keuangan, disediakan pula aktivitas *retweet* untuk menyebarkan kembali *tweet* orang lain sehingga penyebaran informasi menjadi lebih luas. Tujuan penelitian yaitu membangun sistem yang dapat memprediksi penyebaran informasi di twitter menggunakan metode Bernulli dan Gaussian Naive Bayes yang menerapkan beberapa fitur seperti Network Feature, Content Similarity, dan Content Based Feature. Hasil penelitian yang didapat dengan menggunakan *k-fold cross validation* 10 yaitu menunjukkan Bernoulli Naïve bayes lebih unggul dibanding metode Gaussian Naïve Bayes dengan perolehan rata-rata f1-score Benoulli Naïve Bayes yaitu untuk skenario pertama sebesar 60.06% (*f1-score*), skenario kedua sebesar 60.08% (*f1-score*), dan skenario ketiga sebesar 60.09% (*f1-score*).

Kata kunci : Penyebaran Informasi, Twitter, Content Similarity, Naïve Bayes

Abstract

Twitter is an international social media that allows users to share messages or so-called tweets with a maximum of 280 characters per tweet, which can be done publicly or privately with other users. Twitter provides various necessary information ranging from health, education, sports, political, food, and financial information, as well as retweeting activities to redistribute other people's tweets so that the dissemination of information becomes wider. The purpose of this study is to build a system that can predict the dissemination of information on Twitter using the Bernulli and Gaussian Naive Bayes methods that implement several features such as Network Feature, Content Similarity, and Content Based Feature. The results obtained using k-fold cross validation 10 show that Bernoulli Naïve Bayes is superior to the Gaussian Nave Bayes method with an average Benoulli Naïve Bayes f1-score, namely for the first scenario of 60.06% (f1-score), the second scenario by 60.08% (f1-score), and the third scenario by 60.09% (f1- score).

Keywords: Diffusion Information, Twitter, Content Similarity, Naïve Bayes

1. Pendahuluan

Latar Belakang

Seiring berkembang pesatnya teknologi informasi, yang memudahkan dalam mendapatkan informasi melalui media sosial, dimana media sosial telah mendapat banyak perhatian dari masyarakat. Media sosial tidak hanya menyediakan beberapa informasi berita tetapi media sosial seperti Twitter menyediakan berbagai informasi yang diperlukan mulai dari informasi Kesehatan, Pendidikan, Olahraga, Politik, Makanan, dan Keuangan. Berdasarkan fakta yang ada dengan menggunakan situs media sosial masyarakat dapat berbagi informasi maupun perasaan yang mereka miliki untuk mengekspresikan kesukaan dan minat mereka terhadap suatu topik yang mendukung berbagai interaksi sosial dan penyebaran informasi diantara pengguna.

Banyaknya informasi yang tersebar melalui media sosial twitter menjadi sebuah tantangan bagi peneliti sehingga informasi yang tersedia dapat memaksimalkan pengaruh penyebarannya sehingga dapat dimanfaatkan secara lebih efektif. Tidak hanya membantu dalam mempelajarinya saja, tetapi juga dapat dimanfaatkan untuk memecahkan masalah penyebaran informasi yang ada pada media sosial Twitter. Berikut terdapat penelitian mengenai sistem penyebaran informasi pada media sosial yang sudah lebih dahulu dilakukan yaitu Sebagian besar sistem yang dihasilkan dapat mempelajari peyebaran informasi melalui media sosial berdasarkan model probabilitasnya. Pada penelitian [1] fitur yang digunakan berupa *Topic Information, User Activities, Network Connection, Similarity Between User,* dan *Content-Based Similarity*. Menurut penelitian [2] mereka mengusulkan metode yang berfokus pada model linear threshold (LT) untuk menghasilkan probabilitas antara pengguna. Mereka juga menunjukkan Teknik untuk memprediksi waktu yang diinginkan oleh pengguna sesuai dengan waktu yang diinginkan. Pada penelitian [3] dimana fitur yang digunakan yaitu user-based, content-based

dan content similarity dan menghasilkan pemodelan yang. Pada penelitian [4] mengenai penyebaran informasi ditwitter meneliti mengenai relasi re*tweet* dari *user* pada sekumpulan data yang besar, dan menemukan bahwa *user* dengan *followers* yang besar memiliki peran yang besar.

Sistem difusi informasi twitter merupakan sistem penyebarluasan informasi pada media sosial twitter yang dilakukan secara meluas. Sistem ini ditujukan untuk meningkatkan pengetahuan dari segala bidang yang mana dalam penelitian ini lebih khusus membahas bidang kesehatan, antara lembaga dengan pengguna twitter. Sasaran yang hendak dicapai dari sistem ini adalah dapat memprediksi suatu informasi akan disebarkan atau tidak dengan penambahan penggunaan fitur *network*, kesamaan antara konten dan berbasis konten, serta melihat seberapa berpengaruh penambahan fitur tersebut terhadap prediksi sistem difusi informasi.

Sistem digunakan metode Bernoulli dan Gaussian *Naïve Bayes*. Metode klasifikasi ini dipilih karena dianggap memiliki tingkat akurasi yang lebih baik, efisien dan stabil [5] dibanding model klasifikasi yang lain serta menghasilkan nilai untuk melakukan prediksi berdasarkan fitur yang akan digunakan yaitu *Network Feature*, *Content Similarity* dan *Content Based Feature*. Menggunakan data dari media sosial twitter berbahasa indonesia sebagai datasetnya.

Topik dan Batasan

Berdasarkan latar belakang yang telah dibuat, didapatkan rumusan masalah yang ada yaitu Bagaimana cara mendapatkan fitur *Network Feature, Content Similarity* dan *Content Based*, serta bagaimana cara menerapkan metode Bernoulli dan Gaussian Naïve Bayes dalam memprediksi suatu tweet akan disebarluaskan. Batasan masalah dari tugas akhir ini yaitu Fitur yang akan digunakan berupa *Network Feature, Content Similarity* dan *Content Based*. Metode yang digunakan hanya metode Bernoulli dan Gaussian Naïve Bayes. Serta data yang digunakan dari Twitter dengan mengambil topik tentang vaksinasi covid-19.

Tujuan

Tujuan dari penelitian ini adalah membangun sistem yang dapat mempredisksi penyebaran informasi di twitter menggunakan metode Bernoulli dan Gaussian Naive Bayes yang menerapkan beberapa fitur seperti *Network Feature, Content Similarity,* dan *Content Based Feature*.

Organisasi Tulisan

Organisasi tulisan yang terdapat pada proposal ini yaitu pendahuluan, studi terkait, system yang dibangun, evaluasi, kesimpulan dan daftar Pustaka. Pada pendahuluan merupakan penjelasan lebih detail dari abstrak lebih utama lagi mendeskrisikan mengenai latar belakang penjelasan atau mengidentifikasi topik atau masalah erta batasan, adapula tujuan, serta metode penelitian. Untuk studi terkait berisikan teori atau studi literatur dimana berkaitan sesuai topik tugas akhir yang dilakukan. Sistem yang dibangun dapat menjelaskan dan mendeskripsikan rancangan dan system atau produk yang dihasilkan. Evaluasi yaitu berupa evaluasi pengujian yang didapat dan analisis dari pengujian yang dilakukan. Kesimpulan yaitu menjelaskan hasil akhir yang didapat dari hasil pengujian yang dilakukan serta analisa hasil. Daftar Pustaka berisi literatur yang membantu dalam pengerjaan.

2. Studi Terkait

Pada studi terkait yang berkaitan dengan tugas akhir, terdapat penelitian untuk dijadikan sebagai dasar kajian Pustaka dalam pembuatan tugas akhir. Berikut merupakan penjelasan singkat berupa judul, tahun dan deskripsi dari penelitian yang terkait.

Tabel 1 Penelitian Studi Terkait

No	Judul	Tahun	Deskripsi
1.	*Predicting Information Diffusion in Social Networks using Content and User's Profiles*	2013	Penelitian ini menggunakan pendekatan kesamaan antara *user-based* dan *user profile* untuk menghasilkan predikasi penyebaran informasi pada media sosial. Dengan menggunakan pendekatan ini diharapkan dapat mendapatkan nilai probabilitas yang baik
2.	*Prediction of Popular Tweets Using Similarity Learning*	2013	Penelitian ini digunakan rumus cosine similarity dimana terdapat multi-classification berdasarkan fitur re*tweet*.
3.	*Discovering Similar Users on Twitter*	2013	Penelitian ini mempelajari masalah menemukan pengguna "similar" di Twitter, di mana kami mendefinisikan dua pengguna serupa jika mereka menghasilkan konten yang mirip satu sama lain.
4.	*Finding Similar Tweets and Similar Users by Applying Document Similarity to Twitter*	2013	Pada penelitian tersebut merancang skema berbasis konten yang membandingkan kesamaan antara pengguna Twitter dengan mencocokkan *tweet* mereka satu sama lain yang berguna untuk memperkirakan penyebaran informasi di Twitter.

	Streaming Data		
5.	*A tweets classifer based on cosine similarity*	2017	Pada penelitian ini mengusulkan penggunaan *cosine similarity* dengan dua fitur: Bag-of-Words dan word2vec menggunakan dataset lokakarya *Microblog Cultural Contextualization* 2017 untuk menentukan relevansi *tweet* menurut setiap acara dari empat festival Eropa

2.1. Twitter

Twitter merupakan salah satu media sosial internasional dimana mengizinkan pengguna berbagi *tweet* atau biasa disebut *tweet* dengan maksimal 280 karakter per-*tweet*, bisa dilakukan secara publik maupun pribadi dengan pengguna lain. Terhitung pada tahun 2013 twitter memiliki kurang lebih 500 juta, dimana 302 juta diantarannya merupakan pengguna yang aktif. Untuk per-tiap harinya terdapat lebih dari 500 juta *tweet* yang diunggah oleh pengguna Twitter dan menangani 1.6 miliar untuk mesin pencarian dalam per-harinya. Twitter juga menyediakan total 35 Bahasa [6]. Twitter biasanya digunakan pada aktivitas sehari-hari, baik menggunakan komputer maupun ponsel, pengguna juga biasanya membagikan hal-hal yang terjadi pada keseharian mereka seperti membaca, yang sedang dipikirkan, serta yang di alami atau rasakan. Pengguna Twitter untuk menunjukkan minat mereka dengan melakukan following atau mengikuti. Selain *tweet* adapula *retweet* artinya pengguna dapat meneruskan suatu *tweet* dari pengguna lain [7].

2.2. Word2Vec

Word2vec merupakan salah satu library yang disediakan untuk mengolah sebuah kata yang terdapat dalam suatu dataset dimana datanya cukup banyak dan diolah dalam waktu yang cukup singkat dengan hasil akurasinya yang cukup baik diantara metode klasifikasi lainnya. Cara kerja dari word2vec adalah dengan cara menginputkan teks atau korpus [7], sehingga hasil yang didapat berupa representasi vektor dari tiap kata yang terdapat di suatu teks sebagai outputan. Vektor kata tersebut juga bisa dilakukan sebagai pengukur jarak kedekatan antar vector kata yang lain.

2.3. Metode Naïve Bayes

Algoritma Naive Bayes yaitu satu dari banyaknya algoritma yang digunakan untuk pengklasifikasian atau pengelompokkan data dan pengambilan keputusan dan dilakukan sebagai prediksi probabilitas terhadap kelas [8]. Algoritma teorema Bayes menginterpretasikan bahwa setiap atribut yaitu independen yang tidak bergantung pada atribut lainnya. Naives bayes selalu menghasilkan model yang cukup baik dibandingkan model klasifikasi lainnya. Berdasarkan sebuah penelitian oleh Xhemali dkk [9] pada jurnalnya dikatakan metode Naive Bayes menghasilkan akurasi cukup baik dari pada metode klasifikasi yang lain. Kelebihan dari algoritma Naives Bayes yaitu membutuhkan jumlah data sedikit untuk penelitian dimana yang diibutuhkan dalam melaukan pengklasifikasian. Dibawah ini rumus dari teorema Bayes adalah :

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

Dimana :

X : Kelas data yang tidak diketahui
C : Atribut kelas yang sudah diketahui
P(C|X) : *Probability* Atribut C berdasar kondisi X(posteriori probabilitas)
P(C) : *Probability* Atribut C (prior probabilitas)
P(X|C) : *Probability* X berdasar kondisi di C
P(X) : *Probability* X

Dari persamaan teorema bayes diatas di dapatkan algoritma dari metode naïve bayes yaitu :

$$P(\text{C}|\text{X}_1, \dots, \text{X}_n) = \frac{P(\text{C})\prod_{i=1}^{n} P(\text{X}_i|\text{C})}{P(\text{X}_1,\dots,\text{X}_n)} \quad (2)$$

Untuk variabel C merepresetasikan kelas, dan untuk $x_1,\dots,x_n$ merepresetasikan atribut atau fitur yang digunakan. Persamaan diatas merupakan algoritma Naïve Bayes. Untuk klasifikasi menggunakan biner (0 dan 1) [11] digunakan rumus *Bernoulli*, ditunjukan dalam persamaan 3 dan 4 [11].

$$P(\text{X}_i|\text{C}) = \prod_{i=1}^{n} P(\text{X}_i|\text{C})\ (1 - P(\text{X}_i|\text{C}))^{(1-\text{X}_i)} \quad \text{X}_i \in (0, 1) \quad (3)$$

Keterangan :

$\hat{\phi}\hat{\phi}(\hat{\bullet}_i|\hat{\phi}\hat{\phi})$: *Probability* kata yang ada di *class* C

$(1-\hat{\phi}\hat{\phi}(\hat{\bullet}_i|\hat{\phi}\hat{\phi}))$: *Probability* kata yang tidak ada di *class* C

b_{in} : *Probability* kata pada w_t yang ada pada dokumen, jika ada di dokumen b_{it} =1 dan *probability* yaitu $\hat{\phi}\hat{\phi}(\hat{\bullet}_{\hat{\phi}\hat{\phi}}|\hat{\phi}\hat{\phi})$

$(1-b_{it})$: *Probability* kata tidak ada di dokumen b_{it} = 0 dan probability yaitu $1-\hat{\phi}\hat{\phi}(\hat{\bullet}_{\hat{\phi}\hat{\phi}}|\hat{\phi}\hat{\phi})$

$$P(\hat{\bullet}\ |\hat{\bullet}) = \frac{\hat{\bullet}\hat{\bullet}\ \hat{\bullet}\hat{\phi}\ \hat{\bullet}+\hat{\bullet}}{\hat{\phi}\hat{\bullet}\hat{\bullet}+\hat{\bullet}} \quad 4$$

Dimana :

$d\hat{\phi}\hat{\phi}_{\hat{\bullet}\hat{\phi}\hat{\phi},C}$: Total dataset pelatihan yang didalamnya terdapat fitur w_t dan *class* C

$d\hat{\phi}\hat{\phi}_C$: Total dataset pelatihan yang ada di *class* C

+1 dan +2 : Parameter dari *Laplace Smoothing*

Data kontinyu pada dataset digunakan rumus Gaussian Naïve Bayes, dibawah ini rumus yang terdapat pada persamaan 5

$$P(\hat{\bullet} = \hat{\bullet}\ |\hat{\bullet} = \hat{\bullet}) = \frac{\hat{\bullet}}{\hat{\phi}\ \hat{\phi}\ \hat{\phi}\ \sqrt{\hat{\bullet}\pi\hat{\phi}\hat{\phi}\hat{\phi}\hat{\phi}}}\ \hat{\bullet}^{\frac{(\hat{\bullet}_{\hat{\phi}}-\hat{\phi}\hat{\phi}\hat{\phi}\hat{\phi})^{\hat{\bullet}}}{\hat{\bullet}\sigma^{\hat{\bullet}}\hat{\phi}\hat{\phi}}} \quad 5$$

Dimana:

P : Probabilty

X_i : Atribut pada *i*

$\hat{\bullet}_{\hat{\phi}}$: Nilai atribut pada *i*

Y : *Class* ingin didapatkan

$\hat{\bullet}_{\hat{\phi}}$: Sub *class* Y ingin didapatkan

μ : Mean keseluruhan atribut

σ : Standar deviasi.

2.4. Confussion Matrix

Perhitungan matrik digunakan untuk mengetahui seberapa baik hasil dari penelitian yang dilakukan. Digunakan presisi, *recall* dan f1-score [10] untuk melihat akurasi yang didapat berikut penjelasan dari presisi, *recall* dan f1-score :

- Presisi

  Presisi yaitu untuk menghitung keakuratan yang didapat dari informasi yang diinginkan pada suatu pengguna dengan jawaban yang dihasilkan pada sistem. Ditunjukkan rumus precision pada persamaan 6.

$$P\hat{\bullet}\hat{\bullet}\hat{\bullet}\hat{\bullet}\hat{\bullet}\hat{\bullet}\hat{\bullet}\hat{\bullet} = \frac{TP}{TP+FP} \quad 6$$

  dengan :

  TP = Mengasilkan prediksi benar dan target juga benar.

  FP = Mengasilkan prediksi benar dan target salah.

- Recall

  Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Ditunjukkan rumus precision pada persamaan 7.

$$R\hat{\bullet}\hat{\bullet}\hat{\bullet}\hat{\phi}\hat{\phi} = \frac{TP}{TP+\hat{\phi}\hat{\phi}N} \quad 7$$

  dengan :

  TP = Mengasilkan prediksi benar dan target juga benar.

  FN = Mengasilkan prediksi salah dan targetnya benar.

- F1-Score

  F1-Score yaitu sebuah perhitungan yang gabungan antara *recall* dan presisi. Dimana *recall* dan presisi menghasilkan pembobotan yang berbeda satu sama lain . Parameter yang memperlihatkan hubungan timbal balik diantara *Recall* dan Presis yaitu F1-Score yang merepresentasikan pembobotan harmonik rata-rata dan *recall* dan presisi. Ditunjukkan rumus precision pada persamaan 8.

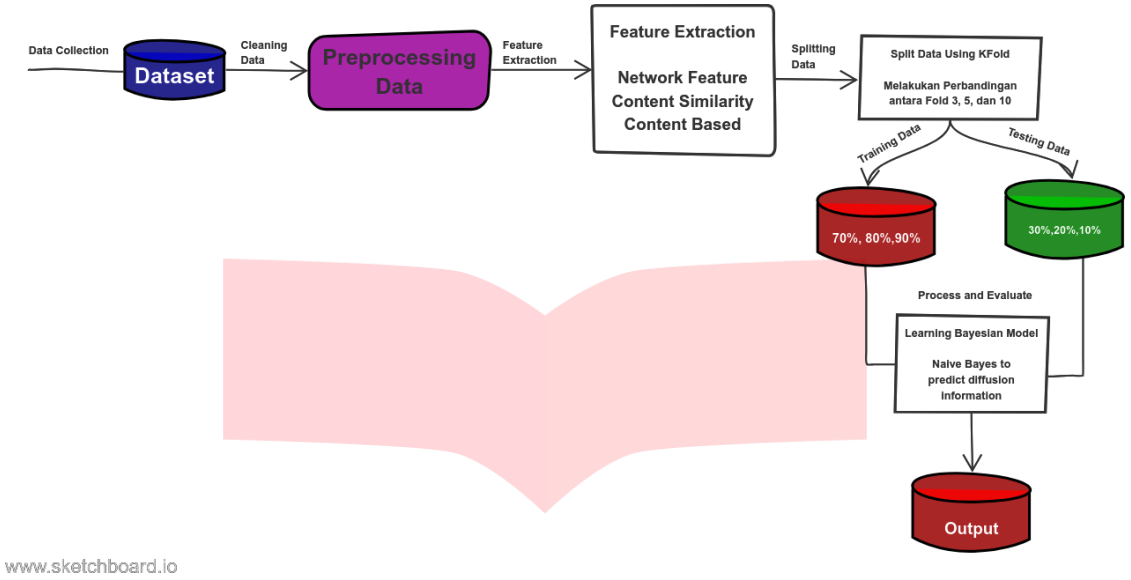
$$F\text{❓} - \text{❓❓❓❓❓} = \text{❓} \times \frac{P\text{❓❓ ❓❓ ❓❓ ❓❓ .❓❓❓ ❓❓ ❓ ❓}}{P\text{❓❓❓❓❓❓❓❓} + \text{❓❓ ❓❓❓❓}}$$

8

Setelah didapat *precision*, *recall* dan *f-measure*-nya, dihitung tingkat akurasinya menggunakan metode Bernoulli dan Gaussian *Naïve bayes.*

3. Sistem yang Dibangun

Deskripsi umum pada sistem yang peneliti bangun digambarkan yang terdapat di Gambar 1. Sistem yang dibangun menggunakan metode bayes untuk memprediksi informasi nilai yang akan muncul. Jika dilihat pada Gambar 3.1 terdapat dataset yang di ektraksi berdasarkan fitur-fitur yang ada untuk menghasilkan prediksi nilai pada sebuah topik yang terkait.



Gambar 3 1. Alur Sistem yang Dibangun

3.1 Dataset

Dataset diambil menggunakan website netlytic yang dimana bisa langsung divisualisasikan datanya sehingga terlihat networknya, pada pengambilan dataset menggunakan web netlytic disini diambil berdasarkan kata kunci vaksinansi covid-19 pada tahun 2021 bulan januari. Total data dalam penelitian ini diperoleh 18468 data pengguna dan 56247 data *tweet*. Namun tidak seluruh data digunakan, data dalam penelitian ini hanya mengambil 600 data pengguna dan 1200 data *tweet* secara random. Atribut yang didapat dari proses scrapping pada website netlytic yaitu id, *tweet* id, guid, link, author, tittle, description, pubdate, source, favorite_count, re*tweet*_count, lang, quoted_text, *tweet*_type, reply_to_screenname, reply_to_*user*_id, reply_to_status_id, *retweeted*_screen_name, *retweeted*_userid, *retweeted*_status_id, *user*_id, profile_image_url, *user*_status_count, *user*_friend_count, *user*_follower_count, *user*_created_at, dan *user*_location. Namun dalam penelitian ini hanya menggunakan beberapa atribut dari keseluruhan atribut yang tersedia.

3.2. Pre-processing Data

Pre-processing data merupakan langkah selanjutnya yang dilakukan setelah pengumpulan data, dimana pada pre-processing akan dilakukan data *cleaning*, *tokenizing*, *normalisasi*, *stop word* dan *stemming* yang dapat meningkatkan kualitas data, sehingga membantu meningkatkan akurasi dan efisiensi proses mining [2]. Dibawah ini merupakan langkah-langkah pre-processing, antara lain:

a. Case Folding

Dimana untuk proses untuk merubah abjad kedalam abjad kecil “a” atau abjad besar “A” berdasarkan kebutuhannya. Karakter pada *tweet* dan *retweet* selain berupa huruf akan dihilangkan. Penggunaan *case folding* digunakan supaya inputan yang dihasilkan menjadi sama.

b. Data Cleaning

Dimana untuk prosedur yang dilakukan untuk menganalisis *null* dan *noise*, menghilangkan *noise*, menghapus simbol, angka dan atribut dan menyelesaikan inkonsistensi dalam data yang akan digunakan yang terdapat pada *tweet*. Karena data yang tersebar pada media sosial cenderung memiliki *null* dan *noise*.

c. Tokenizing

Prosedur untuk memisahkan kalimat kedalam potongan kata yang biasa dikatakan seperti token yang setelahnya dilakukan Analisis. Tokenizing ini dilakukan pada data yang ada terdapat di *tweet*.

d. **Normalisasi (Convert Slangword)**

Memperbaiki kesalahan penulisan pada suatu kata atau mengubah kata yang berupa kata gaul menjadi kata yang sesuai dengan standar KBBI. Contohnya terdapat kata “gue” maka setelah dilakukan normalisasi kata “gue” berubah menjadi “saya”, karena kata “gue” tidak ada dalam KBBI yang terdapat pada *tweet*.

e. **Stemming**

Suatu pemrosesan untuk mengubah suatu kalimat menjadi bentuk dasarnya dari kalimat tersebut. Seperti kata “menjadi” setelah dilakukan stemming yaitu “jadi”.

f. **Stop Words**

Stop words untuk membuang kata pada *tweet* dan *retweet* yang kurang penting dalam pendekatan *bag-of-word* biasannya untuk kata-kata konjungsi, seperti contohnya “dan”, “di” dan “dari” yang terdapat pada *tweet*.

3.3. Feature Extraction

Feature extraction ini memproses hasil dari pre-processing data yang diperoleh kemudian diekstrak dengan fitur-fitur yang digunakan yaitu :

3.3.1. Network Feature

Penting suatu penyebaran informasi mendapatkan informasi dari Twitter yang didapat dari penyebaran *tweet-tweet* dan *retweet* yang ada, untuk mendapatkan informasi data yang terkait yaitu perlu suatu *network* fitur untuk mendapatkan data tersebut dalam bentuk online. Untuk mempelajari difusi informasi, selain fitur berdasarkan isi *tweet*, perlu mempertimbangkan fitur yang memberikan informasi tentang dinamika jejaring sosial. Dinamika jejaring sosial yang terkait dengan pengguna yang terlibat dalam tautan direpresentasikan menggunakan fitur berikut. Menurut penelitian [9] bahwa fitur yang terkait dengan jejaring sosial seperti jumlah teman, jumlah pengikut, dan jumlah sebutan pengguna merupakan indikator yang baik dari 'kemampuan re*tweet*'. Berdasarkan penelitian [11] menyertakan fitur-fitur seperti jumlah pengikut, jumlah penggemar, dan jumlah total keseluruhan *tweet* yang telah diposting.

3.3.2. Content Similarity

Pada penelitian [12] berdasarkan hasil yang didapatkan bahwa probabilitas dari penyebaran informasi yang dilakukan menggunakan fitur *content similarity* terbukti dapat menghitung nilai probabilitas kesamaan antar konten yang terkait dan kertertarikan suatu pengguna pada topik yang terkait dengan hasil probabilitas yang tinggi. Untuk menemukan kesamaan antar konten dan keterkaitan suatu pengguna pada topik, dihitung dengan menggunakan jarak kosinus antara distribusi dari topik konten dan distribusi minat pengguna. *Content similarity* digunakan rumus *cosine similarity*:

Dimana: 9

A = *Vector* A, perlu dibandingkan *similarity*-nya
B = *Vector* B, perlu dibandingkan *similarity*-nya
A • B = *Dot product vector* A dan *vector* B
| A | = Panjang *vector* A
| B | = Panjang *vector* B
| A | | B | = *cross product* antara | A | dan | B |

Pada *Content similarity* untuk mendapatkan kesamaan antara *tweet* dilakukan dengan membandingkan *tweet* antar baris *user*A dan *user*B, yang sebelumnya telah dilakukan pembobotan menggunakan word2vec.

3.3.3. Content Based

Content based merupakan fitur yang dianggap bisa memprediksi *tweet* yang akan di re*tweet* kembali atau tidak. *Content based* merupakan fitur yang memiliki atribut bertipe numerik dan boolean. Atribut yang terdapat pada *content based* dapat berupa nama entitas, sentiment, mengandung media, peningkatan konten, ukuran konten dan sebagainnya [12]. Atribut yang terdapat pada twitter yang seperti contain hashtag, re*tweet* count, length text, url, contain media, contain mention dan sebagainnya.

3.4. Labelling Data

Labelling data ini akan dilihat dari apakah userA melakukan re*tweet* pada userB. Jika userA memiliki aktivitas melaukan *retweet* pada userB maka dilakukan labelling dengan diberi nilai 1, jika userA tidak melakukan aktivitas *retweet* terhadap userB maka dilakukan labelling dengan diberi nilai 0.

3.5. Split Data

Setelah dilakukan *feature extraction*, selanjutnya dipersiapkan metode menggunakan skema K-fold *Cross-Validation* Jumlah *fold* yang akan diuji yaitu 3 , 5, dan 10. Dimana fold 3 memiliki proporsi data 30% data *test* dan 70% data *train*, fold 5 memiliki proporsi data sebanyak 20% data test dan 80% data *train*, sedangkan fold 10 memiliki proporsi data sebanyak 10% data *test* dan 90% data *train*.

3.6. Learning Naïve Bayes Model

Model Bayesian merupakan tahap akhir, yaitu dengan melakukan perhitungan akurasi, *seperti recall, precision*, dan *f1-score*. Perhitungan ini dilakukan untuk mengukur akurasi pada sistem yang peneliti bangun, menggunakan atribut-atribut yang telah digunakan.

4. Evaluasi

Pada evalusasi ini berisikan hasil penelitian dan analisis dari hasil pengujian yang dilakukan berupa sistem yang memprediksi penyebaran informasi apakah suatu *tweet* dapat di*retweet* kembali atau tidak. Pada evaluasi hanya menggunakan *recall, precision, f1-score*, dan akurasi sebagai ukuran kinerja pengklasifikasi.

4.1 Hasil Skenario 1

Berikut merupakan hasil yang didapat menggunakan *k-fold cross validation* dengan total *fold*-nya yaitu 10 dan rasio untuk pembagian datanya yaitu 90:10. Didapatkan perhitungan *confusion matrix* untuk kelas tidak *retweet* dan kelas *retweet* yang terdapat pada Tabel 4.1.

Tabel 4.1 Hasil Confusion Matrix Naïve Bayes Skenario 1

Fold	Prediksi / Aktual	Bernoulli Naïve Bayes		Gaussian Naïve	
		Tidak Retweet	Retweet	Tidak Retweet	Retweet
1	Tidak Retweet	29	31	41	19
	Retweet	9	51	31	29
2	Tidak Retweet	29	31	38	22
	Retweet	8	52	41	19
3	Tidak Retweet	25	35	37	23
	Retweet	13	47	35	25
4	Tidak Retweet	27	33	42	18
	Retweet	10	50	27	33
5	Tidak Retweet	27	33	32	28
	Retweet	7	53	25	35
6	Tidak Retweet	31	29	38	22
	Retweet	9	51	16	44
7	Tidak Retweet	26	34	37	23
	Retweet	18	42	16	44
8	Tidak Retweet	22	38	29	31
	Retweet	15	45	32	28
9	Tidak Retweet	23	37	27	33
	Retweet	24	36	19	41
10	Tidak Retweet	29	31	32	28
	Retweet	21	39	18	42

Berdasarkan Tabel 4.1 Confusion Matrix Skenario 1, kelas “Retweet” menghasilkan nilai paling tinggi di nilai True Positive (TP), dan nilai paling rendah terdapat di nilai False Negative (FN). Nilai pada True Positive (TP) yang tinggi membuktikan bahwa kelas yang bernilai “Retweet” banyak dan menghasilkan klasifikasi dengan benar sebagai “Retweet”. Sementara itu, False Negative (FN) yang rendah membuktikan bahwa sistem masih terdapat kesalahan dalam mengklasifikasikan data, untuk kelas, “Retweet” diklasifikasikan sebagai “Tidak Retweet” oleh model klasifikasi. Kemudian dilakukan perhitungan *precision*, *recall*, *f1-score* serta akurasi sesuai dengan nilai True Positive (TP), False Positive (FP), True Negative (TN), dan False Negative (FN). Nilai pada presisi, *recall*, *f1-score* dan akurasi dihitung berdasarkan persamaan yang telah dijelaskan sebelumnya. Pada Tabel 4.2 merupakan tampilan hasil *classification* pada Skenario 1 dalam memprediksi Retweet dan Tidak Retweet.

Tabel 4.2 Hasil Pengujian Bernoulli Naïve Bayes Skenario 1

Fold	Bernoulli Naïve Bayes						Gaussian Naïve Bayes					
	Retweet			Tidak Retweet			Retweet			Tidak Retweet		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	.622	.850	.718	.763	.483	.592	.604	.483	.537	.569	.683	.621
2	.627	.867	.727	.784	.483	.598	.463	.316	.376	.481	.633	.546
3	.573	.783	.662	.658	.417	.510	.520	.416	.462	.513	.616	.560
4	.602	.833	.699	.730	.450	.557	.647	.550	.594	.608	.700	.651
5	.616	.883	.726	.794	.450	.574	.555	.583	.569	.561	.533	.547
6	.637	.850	.729	.775	.517	.620	.666	.733	.698	.703	.633	.666
7	.553	.700	.618	.591	.433	.500	.656	.733	.692	.698	.616	.654
8	.542	.750	.629	.595	.367	.454	.475	.483	.479	.474	.466	.470
9	.493	.600	.541	.489	.383	.430	.586	.450	.509	.554	.683	.611
10	.557	.650	.600	.580	.483	.527	.640	.533	.581	.600	.700	.646
Macro Average Precision (%)						**62.91**	**Macro Average Precision (%)**					**57.91**
Macro Average Recall (%)						**61.17**	**Macro Average Recall (%)**					**57.75**
Macro Average F1 (%)						**60.06**	**Macro Average F1 (%)**					**57.39**

Berdasarkan pada Tabel 4.2, nilai akurasi akan merepresentasikan tingkat pengklasifikasian pada dataset, berupa nilai untuk kelas ”Retweet” dan diklasifikasi sebagai ”Retweet” dan kelas ”Tidak Retweet” diklasifikasi sebagai ”Tidak Retweet”. Nilai precision adalah nilai yang menunjukkan rasio data yang dilabeli sebagai “Retweet” memang benar bernilai sebagai “Retweet”. *Macro average precision* merupakan hasil rata-rata perhitungan precision “Retweet” dan “Tidak Retweet”. Pada metode Bernoulli Naïve Bayes menghasilkan *macro average precision* dengan nilai 62.91%. Hal itu dikarenakan nilai *False Positive* (FP) lebih rendah dibanding dengan nilai *True Positive* (TP) sehingga menghasilkan nilai presisi yang cukup baik. Nilai *recall* adalah nilai yang menunjukkan rasio dari data yang diklasifikasi secara tepat. *Macro average recall* merupakan hasil rata-rata perhitungan *recall* “Retweet” dan “Tidak Retweet”, yang mana pada metode Bernoulli Naïve Bayes diperoleh hasil dengan nilai 61.17%. Nilai *recall* mendapatkan hasil yang cukup baik karena memiliki *False Negative* (FN) yang lebih kecil. Sedangkan untuk metode Gaussian Naïve Bayes menghasilkan *macro average precision* dengan nilai 57.91%. Hal tersebut dikarenakan nilai *False Positive* (FP) lebih tinggi dibanding dengan nilai *True Positive* (TP). Untuk *macro average recall* pada metode Gaussian Naïve Bayes memiliki hasil yang sama rendahnya dengan *macro average precision* yaitu dengan nilai 57.75%. Nilai *recall* mendapatkan hasil yang rendah karena memiliki *False Negatif* (FN) yang paling besar. Nilai f1-*score* digunakan untuk mengevaluasi hasil rata-rata *precision* dan *recall* hasil klasifikasi.

4.2 Analisis Hasil Pengujian

Pengujian dalam peneilitian ini dilakukan dengan melakukan beberapa skenario berdasarkan pembagian proporsi data menggunakan metode *k-fold cross validation*. Berdasarkan hasil Tabel 4.3 pada metode Bernoulli Naïve Bayes dengan rasio 90:10 didapatkan total data *train* sebanyak 1080 data dan data *test* sebanyak 120 data dengan perolehan rata-rata *precision* 62.91%, *recall* 61.17% dan *f1-score* 60.06%. Pada rasio 80:20 didapatkan total data *train* sebanyak 960 data dan data *test* sebanyak 240 data dengan perolehan rata-rata *precision* 62.89%, *recall* 61.17% dan *f1-score* 60.08%. Pada rasio 70:30 didapatkan total data *train* sebanyak 800 data dan data *test* sebanyak 400 data dengan perolehan rata-rata *precision* 62.86%, *recall* 61.17% dan *f1-score* 60.09%. Pada metode Gaussian Naïve Bayes dengan rasio 90:10 diperoleh rata-rata *precision* 57.91%, *recall* 57.75 %, dan *f1-score* 57.39%. Pada rasio 80:20 Gaussian Naïve Bayes memperoleh hasil rata-rata *precision* 57.75%, *recall* 57.59%, dan *f1-score* 57.23%. Pada rasio 70:30 Gaussian Naïve Bayes memperoleh hasil rata-rata *precision* 58.41%, *recall* 58.17%, dan *f1-score* 57.55%. Dari ketiga skenario dan kedua metode tersebut menunjukkan bahwa semakin bertambahnya jumlah data *training*, maka tingkat akurasi akan cenderung semakin meningkat.

Tabel 4.4 Hasil Pengujian Bernoulli dan Gaussian Naïve Bayes Berdasarkan Proporsi Data

Metode	Skenario	Rasio (%)	Data *Training*	Data *Testing*	Macro *Average Precision*	Macro *Average Recall*	Macro *Average F1-Score*
Bernoulli NB	1	90:10	1080	120	62.91%	61.17%	60.06%
	2	80:20	960	240	62.89%	61.17%	60.08%
	3	70:30	800	400	62.86%	61.17%	60.09%
Gaussian NB	1	90:10	1080	120	57.91%	57.75%	57.39%

	2	80:20	960	240	57.75%	57.58%	57.23%
	3	70:30	800	400	58.41%	58.17%	57.55%

5. Kesimpulan

Dalam penelitian ini, kami menggunakan beberapa fitur informatif seperti fitur *network* fitur *content similarity*, dan fitur *content based*) dari platform media sosial Twitter, lalu menyelesaikan masalah prediksi penyebaran informasi menggunakan metode pembelajaran terawasi (yaitu, Naïve Bayes). Hasil menunjukkan bahwa hasil dari tiga skenario yang di uji terhadap dua metode Naïve Bayes yang digunakan, menunjukkan Bernoulli Naïve bayes lebih unggul dibanding metode Gaussian Naïve Bayes, dikarenakan data yang digunakan terdiri dari beberapa atribut yang memiliki nilai biner. Hal ini karena metode Bernoulli Naïve Bayes dapat bekerja baik dengan data yang memiliki nilai biner. Pengujian yang telah dilakukan menggunakan metode Bernoulli Naïve Bayes, untuk proses klasifikasi didapatkan nilai f1-*score* dari evaluasi dengan *confusion matrix*, yaitu skenario pertama diperoleh nilai *f1-score* 60.06%. Skenario kedua diperoleh nilai *f1-score* 60.08%. Skenario ketiga diperoleh nilai *f1-score* 60.09%. Penulis memberi saran pada penelitian yang akan dating untuk menggunakan dataset yang besar. Dan juga, dapat mengembangkan fitur baru seperti fitur *user-based* atau *user-similarity* dan juga dapat menerapkan metode klasifikasi seperti SVM, *Decision Tree* dan lainnya.

Referensi

[1] S. K. V. G. Devesh Varshneya, "Predicting information diffusion probabilities in social networks: A Bayesian networks based approach," *Elsevier*, pp. 1-11, 2017.

[2] F. B. L. V. S. L. Amit Goyal, "Learning Influence Probabilities In Social Networks," *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, p. 241–250, 2010.

[3] A. S. R. S. Nidhi Singh, "Predicting Information Cascade on Twitter Using Random Walk," *Elsevier*, pp. 201-209, 2020.

[4] N. P. F. T. H. T. Cazabet Remy, "Information Diffusion on Twitter: everyone has its chance, but all chances are not equal," *IEEE*, 2013.

[5] Y. S. N. Riszki Wijayatun Pratiwi, "Prediksi Rating Film Menggunakan Metode Naïve Bayes," *Jurnal Teknik Elektro*, vol. 8 , pp. 60-63, 2016.

[6] T. Inc., "About Twitter," 2014. [Online].

[7] T.-T. K. S.-D. L. San-Chuan Hung, "Novel Topic Diffusion Prediction using Latent Semantic and User Behavior," *Proceedings of the ASE BigData & Social Informatics 2015*, pp. 1-6, 2015.

[8] I. S. K. C. G. C. J. D. Tomas Mikolov, "Distributed Representations of Words and Phrases and their Compositionality," *arXiv*, 2013.

[9] C. J. H. a. R. G. Daniela XHEMALI, "Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *IJCSI*, vol. 4, pp. 16-23, 2009.

[10] H. Annur, "KLASIFIKASI MASYARAKAT MISKIN MENGGUNAKAN METODE NAÏVE BAYES," *ILKOM Jurnal Ilmiah*, vol. 10, pp. 160-165, 2018.

[11] Y. X. Siyao Han, "Link Prediction in Microblog Network Using Supervised Learning with Multiple Features," *Journal of Computer*, vol. 11, pp. 72-82, 2015.

[12] E. F. a. M. H. Ian Witten, Data Mining: Practical Machine Learning Tools and Techniques, 2016.

[13] M. A. F. Y. A. S. d. E. D. L. S. Fachrul Rozy Saputra Rangkuti, "Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dengan Ensemble Feature dan Seleksi Fitur Pearson Correlation Coefficient," *JPTIIK*, vol. 2, 2018.