# Analisis dan implemenasi algoritma C4.5 dan pembobotan TF-IDF untuk menentukan trending topik pada media sosial twitter

**Ridwan Rafif [1], Erwin Budi Setiawan [2], Isman Kurniawan[3]**
[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]ridwanrafif@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id,
[3]ismankrn@telkomuniversity.ac.id,

Abstrak
*Twitter* merupakan media sosial *microblogging* dengan perkembangan tercepat diantara semua media sosial. Pesan yang disematkan pada *twitter* merupakan kejadian yang sedang terjadi. Dengan banyaknya penggunaan *twitter* saat ini, banyaknya jumlah *tweet* yang disematkan setiap harinya dapat dikelompokan menjadi sebuah trending topik yang meggambarkan berita yang banyak dibicarakan pada saat ini. Permasalahannya adalah bagaimana *tweet* yang disematkan ini dapat manjadi sebuah berita yang dibicarakan banyak penggunanya dengan *mapping* atau pembagian kelas untuk setiap *tweet* yang disematkan dan bagaimana sebuah *tweet* dapat dijadikan kedalam satu kelas tertentu. Pada penelitian tugas akhir ini, penulis membangun sistem untuk mengklasifikasi trending topik dan menganalisa trending topik *twitter* apa yang kemungkinan muncul menggunakan algoritma C4.5 dengan metode pembobotan TF-IDF untuk memperoleh hasil yang maksimal.

**Kata Kunci: TF-IDF, *Twitter,* C4.5, Trending Topik,**

*Abstract*
*Twitter is the fastest growing microblogging social media among all social media. Messages embedded in Twitter are events that are happening. With the current number of Twitter uses, the number of tweets embedded with each payment can be recognized as a trending topic that illustrates the news that is widely discussed at the moment. The problem is how these embedded tweets can be the news that many users talk about by mapping or class divisions for each tweet embedded and how tweets can be used for one specific class. In this final project research, the author makes a system to classify trending topics and analyze what Twitter trending topics appear using the C4.5 algorithm with the TF-IDF weighting method to obtain maximum results.*

***Keywords:* TF-IDF, *Twitter,* C4.5, Trending Topic**

## 1. Pendahuluan

Teknologi informasi yang dengan cepat berkembang membuat media sosial menjadi konsumsi yang banyak digunakan pengguna internet seperti *Facebook*, *Twitter*, *Instagram*, *path*, dan lainnya. *Twitter* saat ini menjadi media sosial dengan pengguna terbanyak yaitu sebanyak 140 juta pengguna dan memproduksi 400 juta *tweet* setiap harinya[1]. Berdasarkan dari data yang dirilis oleh *world of tweets dot com* pada tahun 2010, Indonesia menjadi peringkat 3 dunia dengan *tweet* sebanyak 13.39% dari total pengguna *twitter* dunia dan menjadi yang terbanyak di asia dengan aktifitas *twitter* 59.97%[2].

*Twitter* sangat cocok untuk penerbitan dan penyebaran informasi berita. Teks *tweet* pendek, sehingga mudah bagi pengguna untuk memposting *tweet* mereka dengan ponsel atau alat lainnya. Akibatnya, pengguna dapat mempublikasikan apa yang mereka lihat secara *real-time*, terutama untuk informasi berita. Fungsi "*following*" membuatnya sangat tepat untuk difusi informasi berita. Saat pengguna berpartisipasi dalam diskusi topik berita, *tweet* terkait akan didorong ke pengikutnya. Jika para pengikutnya juga tertarik dengan topik tersebut, mereka dapat mengambil bagian dalam diskusi juga. Dengan demikian, topik berita menyebar dengan sangat cepat.[3]

Trending topik adalah fitur yang unik pada media sosial twitter yang dapat dijadikan parameter mengenai berita apa yang sedang terjadi saat ini. Penggunaan fitur unik twitter ini menjadi alasan penggunaan data trending untuk melihat apakah data yang berupa trending topik ini dapat memberikan sebuah informasi.

Pada penelitian sebelumnya [4], masalah yang diangkat, tweet yang menjadi trending topik harus merujuk dari beberapa tweet sebelumnya, pada kasus *boone logan*, harus ada beberapa tweet yang merujuk kepada pitcher *boone logan* sehingga kasus tersebut menjadi trending, penentuan trending topik hanya dilakukan menggunakan kata kunci. klasifikasi *tweet* dapat dibagi menjadi beberapa *class* diantaranya opini, kejadian, berita, Algoritma untuk mengklasifikasikan teks yang dapat digunakan pada penelitian ini terdapat pada *data mining*. *Data mining* adalah proses menemukan data dan mengetahui *big data* [5],. dengan

menggunakan algoritma C4.5 memiliki nilai akurasi yang tinggi dari rata-rata hasil metode penelitian yaitu diatas 50% dan lebih tinggi dari Naïve bayes [6] dari data yang didapat, metode *decision tree induction* memiliki nilai akurasi tinggi.

Dalam penelitian ini penulis melakukan analisis trending topik menggunakan pembobotan *Frequency Inverse Document Frequency* (TF-IDF) yang merupakan metode yang digunakan dalam melakukan pembobotan terhadap kemunculan kata dalam suatu dokumen [7], dan algoritma C4.5 akan melihat keakuratan penggunaan metode algoritma C4.5 dalam menganalisa trending topik serta seberapa banyak pengaruh dari metode pembobotan tersebut.

## 2. Studi Terkait

### 2.1 Twitter

Twitter adalah media sosial yang memungkinkan penggunanya untuk berinteraksi antar sesama penggunanya, memungkinkan penggunanya untuk memposting *tweet, update*, yang tidak lebih dari 140 karakter kedalam jaringan yang disebut dengan *followers*, Sementara beberapa pengguna menganggap bahwa 140 karakter dianggap sebagai informasi singkat dan mudah untuk disebarkan [1]. Beberapa fitur dari *twitter* :

| ***No*** | ***Fitur*** | ***Keterangan*** |
|---|---|---|
| *1* | Halaman Utama (*Home*) | Pada fitur ini kita dapat melihat *tweets* dari orang-orang yang berteman dengan kita atau dapat melihat *tweets* dari seseorang yang kita *follow*. |
| *2* | Profil (*Profile*) | Di halaman profil berisi data diri kita berupa tempat tanggal lahir, bio, serta kumpulan *tweets* yang pernah kita buat. |
| *3* | Pengikut (*Followers*) | *Followers* adalah pengguna lain yang ingin menjadikan kita sebagai temannya. Seorang follower akan menerima *tweets* dari seseorang yang di *follow*-nya di halaman utamanya. |
| *4* | Mengikuti (*Following*) | Kebalikannya dari *followers*, *following* adalah mengikuti akun pengguna lain sehingga *tweets* orang tersebut muncul di halaman utama orang yang mengikutinnya. |
| *5* | *Mentions* | Fitur ini digunakan untuk membalas *tweets* dari seseorang dengan cara menandai orang tersebut di *tweet* yang kita buat. |
| *6* | *Favorite* | Bila seseorang menyukai suatu *tweets* dia akan menandainya dengan fitur *favorite* agar *tweets*nya tidak hilang. Yang di *favorite*-kan biasanya *tweets* berupa *quotes* yang sesuai dengan keadaan penggunannya pada saat itu. |
| *7* | Pesan Langsung (*Direct Massage*) | Pesan langsung digunakan untuk mengirim pesan ke pengguna lain. Pesan ini hanya bisa dilihat oleh kedua Pengguna yang sedang berinteraksi |
| *8* | *Hashtag* | *Hastag* "#" digunakan untuk menandai topik tertentu, biasanya topik yang sedang viral di kala itu sedang terjadi. |
| *9* | *List* | *List* berfungsi untuk mengelompokkan pengguna lain yang diikuti oleh pengguna sehingga memudahkan melihat keseluruhan nama para penggunanya. |

### 2.2 Trending topic

Kejadian yang saat ini terjadi seringkali disebarkan melalui media sosial .Twitter memiliki satu fitur yang disebut Trending Topik, yaitu suatu daftar real-time mengenai topik yang sedang populer/ dibicarakan banyak orang[9]. Trending topik itu sendiri merupakan sebuah masalah yang dibicarakan banyak orang atau kejadian yang sedang terjadi. Saat topik baru menjadi populer di twitter, maka topik tersebut termasuk kedalam *list* trending topik yang berupa sebuah frasa pendek atau hastags[8]. Isi dari trending topik berupa informasi dari trend yang terjadi di *twitter* seringkali sulit dimengerti, maka penting untuk melakukan klasifikasi topik-topik tersebut agar mudah dipahami. Seperti contohnya tren bernama #booneLogan. Bagi yang tidak mengikuti *american major league baseball* tidak akan mengetahui tentang boone logan yang

merupakan *pitcher* dari new York yankee, maka dari itu pengolahan trending topik menjadi informasi yang dapat dimengerti oleh orang lain**.**

### *2.3* Pembobotan

Pembobotan merupakan teknik pengambilan keputusan pada suatu proses yang melibatkan berbagai faktor secara bersama-sama dengan cara memberi bobot pada masing-masing faktor tersebut.
Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan istilah dalam dokumen karena dipercaya bahwa frekuensi kemunculan istilah merupakan petunjuk sejauh mana istilah tersebut mewakili isi dokumen[14] *Term Frequency-Inverse Document Frequency* merupakan suatu cara untuk pembobotan sebuah kata dalam sebuah dokumen atau korpus.TF-IDF sendiri merupakan produk dari dua statistik, yaitu TF (*Term Frequency*) dan IDF (*Inverse Document Frequency*).[15]. Pada proses ini data yang telah dikumpulkan sebelumnya akan diproses untuk dilakukan pembobotan yang bertujuan untuk mendapatkan *rating* pada kata-kata yang didapatkan. Metode ini akan digunakan untuk mengetahui seberapa berpengaruh hasil yang akan dikeluarkan. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus sebagai berikut[11]:

$$tfidf_t = f_{t,d} \times log \frac{N}{df_t} \quad (1)$$

Dimana:

$tfidft$ = bobot dari kata ke-t

$ft,$ = kemuculan kata-t dalam dokumen ke-d

$N$ = total dokumen

$dft$ = banyaknya dokumen yang mengandung kata-t

Dengan $tfidf_t$ merupakan bobot kata t, $f_{t,d}$ merupakan frekuensi kemunculan kata t pada dokumen d dan n adalah total dokumen dan $df_t$ merupakan banyaknya dokumen yang mendukung kata ke t.

### 2.4 Algoritma C4.5 (tuliskan

Algoritma C4.5 adalah metode klasifikasi pengembangan dari algoritma ID3 ( *iterative Dychotomizer version 3).* Algoritma C4.5 bekerja dengan membentuk suatu aturan dan membagi data menjadi beberapa kelompok yang akan digunakan sebagai aturan pembuatan pohon keputusan. Algoritma ini dapat digunakan untuk menangani data kontinu dan *missing value (lengkapi urutan proses algoritmanya)*

Terdapat proses yang harus dilakukan sebelum menentukan aturan yang akan digunakan untuk menentukan pohon keputusan, yaitu menentukan atribut yang akan menjadi *root* berdasarkan nilai *gain ratio* terbesar. Kemudian menentukan atribut yang akan menjadi *internal node* untuk setiap *branch* dari *parent node*, dan membuat *decision node* Ketika seleksi atribut tidak dapat digunakan. Persamaan yang digunakan adalah sebagai berikut[13]

1. Entropy

$$\boldsymbol{Entropy(s) = \Sigma_{i=0}^{n} - pi * log_2(pi)} \quad (2)$$

Rumus (2) merupakan rumus yang digunakan dalam perhitungan entropy yang digunakan untuk menentukan seberapa informatif atribut tersebut. Berikut keterangannya :

s : Himpunan kasus

n : Jumlah partisi

pi : Jumlah kasus pada partisi ke-i

2. Information Gain

Information Gain adalah informasi yang didapatkan dari perubahan entropy pada suatu kumpulan data, baik melalui observasi atau bisa juga disimpulkan dengan cara melakukan partisipasi terhadap suatu set data[12]

$$GAIN\ (S,A) = Entropy(S) - \Sigma_{i=1}^{n} - \frac{Si}{S} * Entropy\ (S) \qquad (3)$$

Rumus (3) merupakan rumus yang digunakan dalam perhitungan information gain setelah melakukan perhitungan entropy. Berikut keterangannya :

s : Himpunan kasus

n : Jumlah partisi atribut A

Si : Jumlah kasus pada partisi ke-i

S : Jumlah kasus dalam s

3. Gain Ratio

*Gain Ratio* merupakan modifikasi dari *information gain* untuk mengurangi bias atribut yang memiliki banyak cabang. Dimana (Clearkan masalah data yang kontinu. Beri contoh)

$$Gain\ Ratio = \frac{Gain(S,A)}{Split\ Information\ (S,A)} \qquad (4)$$

Dengan *split information(S,A)* sebagai berikut

$$Split\ Information(S,A) = -\sum_{i=1}^{n} \frac{si}{s} log_2\ \frac{si}{s} \qquad (5)$$

Contoh data latih berdasarkan hasil pembobotan:

**Tabel 1 Contoh Data Latih**

| ID tweet | TFIDF_Unigram | TFIDF_Bigram | TFIDF_Trigram | label |
|---|---|---|---|---|
| 5992 | 2.375194305 | 3.627957274 | 7.277576044 | Teknologi |
| 6002 | 0.423898007 | 0.744295246 | 0.744956317 | Sosial |
| 11030 | 5.465346263 | 0 | 0 | Senbud |
| 14695 | 2.717413731 | 0 | 0 | Ekonomi |
| 5996 | 1.186209348 | 6.822278562 | 0 | Teknologi |
| 11032 | 2.634020099 | 0 | 0 | Senbud |
| 11641 | 2.763395324 | 0 | 0 | Pendidikan |
| 14698 | 1.811609154 | 0 | 0 | ekonomi |
| 1589 | 1.059745016 | 2.223040826 | 4.446511852 | Hiburan |

Pada tabel 1, data latih disusun oleh beberapa atribut yaitu *jumlah following, jumlah followers, jumlah like, tweet,* dan juga *predefined class* yaitu berupa label. Atribut diatas bertipe kontinu seperti *jumlah followes, jumlah following, jumlah like,* nilai *gain ratio* harus dihitung untuk seluruh threshold yang merupakan *mean* dari *discticnt value* atau mencari *median* yang tersusun secara urut dari yang terkecil hingga terbesar

Untuk mengetahui *Gain Ratio* setiap atribut, nilai *entropy* awal sebelum himpunan data latih pada tabel 1 dipartisi perlu dihitung sebagai berikut:
Jumlah label Teknologi: 2
Jumlah label Sosial: 1
Jumlah label Senbud: 2
Jumlah label Ekonomi: 2
Jumlah label Pendidikan : 1
Jumlah label Hiburan: 1

$$\boldsymbol{Entropy(S)} = \left(-\frac{2}{9} * \boldsymbol{log_2}\left(\frac{2}{9}\right)\right) + \left(-\frac{1}{9} * \boldsymbol{log_2}\left(\frac{1}{9}\right)\right) + \left(-\frac{2}{9} * \boldsymbol{log_2}\left(\frac{2}{9}\right)\right) + \left(-\frac{2}{9} * \boldsymbol{log_2}\left(\frac{2}{9}\right)\right) + \left(-\frac{1}{9} * \boldsymbol{log_2}\left(\frac{1}{9}\right)\right) + \left(-\frac{1}{9} * \boldsymbol{log_2}\left(\frac{1}{9}\right)\right) = 2.503258$$

Setelah mendapatkan nilai *entropi*,data latih akan dipartisi dengan atribut manapun yang bernilai kontinu, dengan nilai tengah adalah *entropy* = 2.503258 maka data dapat dibagi berdasarkan *threshold* nilai tengah atribut 2.503258 untuk menghitung *gain*

**tabel 2 distribusi kelas data kontinu**

| *Threshold* 2.503258 | Teknologi | Hiburan | Senbud | Ekonomi | Sosial | pendidikan | total |
|---|---|---|---|---|---|---|---|
| <2.503258 | 3 | 2 | 4 | 5 | 3 | 2 | 19 |
| >2.503258 | 3 | 1 | 2 | 1 | 0 | 1 | 8 |

$$\boldsymbol{Gain(threshold)} = \frac{\mathbf{19}}{\mathbf{27}} \times \left( -\frac{\mathbf{3}}{\mathbf{19}} * \boldsymbol{log_2}\left(\frac{\mathbf{3}}{\mathbf{19}}\right) - \frac{\mathbf{2}}{\mathbf{19}} * \boldsymbol{log_2}\left(\frac{\mathbf{2}}{\mathbf{19}}\right) - \frac{\mathbf{4}}{\mathbf{19}} * \boldsymbol{log_2}\left(\frac{\mathbf{4}}{\mathbf{19}}\right) - \frac{\mathbf{5}}{\mathbf{19}} * \boldsymbol{log_2}\left(\frac{\mathbf{5}}{\mathbf{19}}\right) - \frac{\mathbf{3}}{\mathbf{19}} * \boldsymbol{log_2}\left(\frac{\mathbf{3}}{\mathbf{19}}\right) - \frac{\mathbf{2}}{\mathbf{19}} * \boldsymbol{log_2}\left(\frac{\mathbf{2}}{\mathbf{19}}\right) \right) + \frac{\mathbf{8}}{\mathbf{27}} \times \left( -\frac{\mathbf{3}}{\mathbf{8}} * \boldsymbol{log_2}\left(\frac{\mathbf{3}}{\mathbf{8}}\right) - \frac{\mathbf{1}}{\mathbf{8}} * \boldsymbol{log_2}\left(\frac{\mathbf{1}}{\mathbf{8}}\right) - \frac{\mathbf{2}}{\mathbf{8}} * \boldsymbol{log_2}\left(\frac{\mathbf{2}}{\mathbf{8}}\right) - \frac{\mathbf{1}}{\mathbf{8}} * \boldsymbol{log_2}\left(\frac{\mathbf{1}}{\mathbf{8}}\right) - \frac{\mathbf{1}}{\mathbf{8}} * \boldsymbol{log_2}\left(\frac{\mathbf{1}}{\mathbf{8}}\right) - \mathbf{0} \right) = \mathbf{2.401345}$$

Setelah menghitung nilai *entropy* sebelum dan sesudah proses partisi, maka nilai *gain*, *split info* dan *gain ratio* dapat dihitung

$$\boldsymbol{GAIN(X) = 2.021053 - 1.004547 = 0.101913}$$

$$Split\ Info(X) = -\frac{19}{27} \times \log_2\left(\frac{19}{27}\right) - \frac{8}{27} \times \log_2\left(\frac{8}{27}\right) = 0.876716$$

$$Gain\ Ratio(X) = \frac{\mathbf{0.101913}}{0.876716} = 0.116244$$

Penghitungan di atas hanya menghitung *Gain Ratio* untuk satu *threshold* atribut *jml_follower* saja dan untuk mengetahui *Gain Ratio* untuk atribut ini perlu dilakukan penghitungan untuk seluruh *threshold* yang dimilikinya kemudian mencari nilai terbesar yang ada

## 2.6 Analisis performansi

Dalam tugas akhir ini performansi dari setiap metode *term weighting* yang telah digunakan akan diukur dengan menggunakan nilai akurasi, *precision*, dan *recall*. Berikut adalah tabel *confusion matrix*:

**Tabel 3 *Confusion Matrix***

| **Kategori** | **Kelas Prediksi** | | |
|---|---|---|---|
| Kelas Sebenarnya | | Kelas = *Yes* | Kelas = *No* |
| | Kelas = *Yes* | *Tp* | *Fn* |
| | Kelas = *No* | *Fp* | *Tn* |

Dimana:
TP (Benar Positif) : Kelas yang diprediksi *yes*, dan faktanya adalah *yes.*
TN (Benar Negatif) : Kelas yang diprediksi *no*, dan faktanya adalah *no.*
FP (Salah Positif) : Kelas yang diprediksi *yes*, dan faktanya adalah *no.*
FN (Salah Negatif) : Kelas yang diprediksi *no*, dan faktanya adalah *yes.*

1. Akurasi
   Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aslinya. Akurasi digunakan untuk mengevaluasi banyaknya label prediksi yang sesuai dengan label aktual. Semakin besar nilai akurasinya, maka performansi klasifikasi semakin baik. Berikut persamaan dari akurasi [14]:

$$\text{Akurasi} = \frac{TP + TN}{(TP + FP + TN + FN)} \quad (10)$$

2. *Precission*

*Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Bila di data *mining precision* adalah jumlah dokumen yang dengan benar diklasifikasikan dalam sebuah kelas dibagi jumlah total dokumen dalam kelas tersebut. Dengan persamaan [6]:

$$Precision(P) = \frac{TP}{(TP + FP)} \quad (11)$$

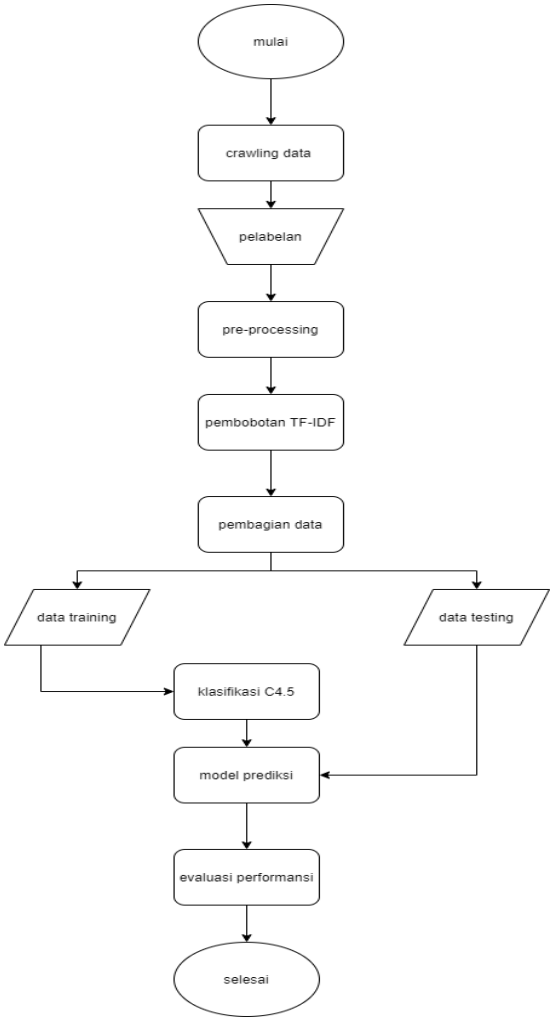
3. *Recall*

*Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Dalam data *mining*, *recall* dapat didefinisikan sebagai jumlah dokumen yang dengan benar diklasifikasikan dalam sebuah kelas dibagi jumlah total dokumen yang diklasifikasikan dalam kelas tersebut. Dengan persamaan [6]:

$$Recall(R) = \frac{TP}{(TP + FN)} \quad (12)$$

## 3. Sistem yang dibangun

Rangkaian proses dalam sistem ini akan dibangun seperti **Gambar 1**.




**Gambar 1. Rancangan sistem Alur analisis trending topik**

1. *Crawling Data*
   *Crawling data* adalah proses mengambil data dari sebuah *website twitter* dengan memanfaatkan *API twitter*. *crawling data* di twitter dapat menggunaan system pencarian *by user*, dan *by keyword.* Jumlah data yang didapatkan pada tahap ini adalah sebanyak147776 *tweet.*
2. Pelabelan
   Pada tahap ini data yang sudah dikumpulkan akan dilabelkan menjadi : Ekonomi, Umum, Kesehatan, otomotif, Hukam, Politik, hiburan, olahraga, kesenian, seni budaya, sosial
3. *Preprocessing*
   Pada *preprocessing* ini bertujuan untuk mengubah data yang tidak terstruktur menjadi data yang dapat diolah dengan mudah sesuai dengan kebutuhannya. Proses *preprocessing* yang pertama adalah penghapusan *retweet* dari tweet, karena *retweet* tidak dapat menjelaskan kepribadian si pengguna, kemudian diikuti dengan penghapusan URL. *Case folding* yaitu mengubah huruf besar menjadi huruf kecil. *Tokenizing* merupakan proses penghapusan karakter seperti titik (.) dan koma (,) lalu, *tweet* diuraikan menjadi satuan kata. *Filtering* yaitu pemilihan kata-kata penting setelah proses *tokenizing*. Dan yang terakhir adalah *stemming* yaitu proses mengubah kata yang berimbuhan menjadi kata dasar.
4. Pembobotan
   Pada proses ini masing-masing data *tweet* yang sudah selesai di-*preprocessing* akan diberi nilai atau bobot dengan menggunakan pembobotan TF, TF-IDF. Metode ini dilakukan untuk mendapatkan nilai terhadap kata yang berpengaruh terhadap dokumen
5. Klasifikasi Algoritma C4.5
   Data yang telah di-*preprocessing* akan masuk ke tahap klasifikasi Dengan menggunakan algoritma C4.5 Dimana data latih akan diinputkan dan dihitung berdasarkan proses *decision tree* C4.5 yaitu perhitungan nilai *entropy, information gain*, *split info*, dan nilai *Gain Ratio*. Output dari proses ini adalah model prediksi yang nanti akan diujikan performansinya.
6. Model Prediksi
   Model prediksi adalah sistem pembelajaran yang sudah dibuat dari klasifikasi *decision tree* C4.5. Data uji yang sudah dibuatkan skenarionya akan diuji terhadap model yang telah dibuat. *Output*-nya adalah nilai performansi dari mode yang dibuat.
7. Evaluasi performansi
   Pada proses ini akan dilakukan perhitungan akurasi, *precision*, dan *recall* untuk mengukur performansi dari sistem yang telah dibuat.

## 4. Hasil Analisa dan Uji

Pada bagian ini akan dijelaskan bagaimana hasil uji dari sistem yang telah dibangun sesuai dengan *flowchart* yang telah dibuat sebelumnya, serta akurasi yang didapat.

### 4.1 Data Set dan Pelabelan

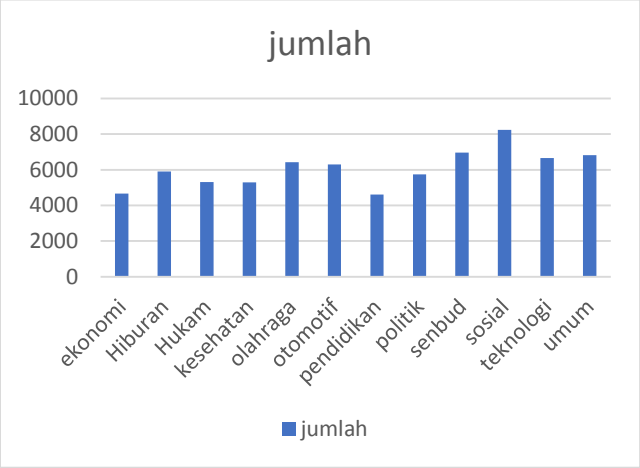
Berdasarkan hasil *crawling data twitter* , sebanyak 147776 *twee*t yag didapatkan akan digunakan sebanyak 77792 *tweet* yang akan dilabelkan kedalam 12 topik berdasarkan kata dan tagar (#) pembangunnya. Pelabelan dilakukan dengan menggunakan *MySql database* dan *datasaur*

**Tabel 4 hasil *crawling data twitter***

| ID | id_tweet | id_user | Nama_akun | Jml like | Jml emoji | … | tweet | label |
|---|---|---|---|---|---|---|---|---|
| 1 | 14699 | 2843601854 | Seputar Ekonomi | 0 | 0 | … | Kelaziman pembayaran kupon obligasi | Ekonomi |
| 2 | 2441 | 1150000000 | Riana Dewie | 2573 | 0 | … | parade budaya lomba lukis | Hiburan |
| 3 | 39484 | 1160000000 | Putriarna | | | … | beritahu polisi segera | Hukam |
| 4 | 1437 | 2363027508 | PSSI | 29 | 0 | … | yanto basna berhasil menyelamatkan gawang | Olahraga |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | indonesia | |
| … | … | | | | | …. | … | … |
| 77792 | 43632 | 113000000 | IndraPohan7 | 1 | 1 | … | event idul adha free fire | Umum |

Hasil dari *crawling data* dan pelabelan, Penulis mendapatkan sebanyak 72590 tweet yang dapat di proses dengan optimal setelah melewate fase *preprocessing* yang terbagi menjadi 12 label topik yang kemudian akan diuji menggunakan algoritma C4.5




**Gambar 2 jumlah tweet terhadap tabel**

### 4.2 hasil uji

Pada bagian ini akan dijelaskan hasil uji dari sistem yang telah dibuat dan dirancang sesuai dengan *flowchart* yang telah dibuat sebelumnya, serta akurasi, dan kompetensi pengguna *twitter* yang didapat.

#### 4.2.1 Perhitungan dan Pembobotan TF-IDF

Pada proses ini dilakukan perhitungan pembobotan TF-IDF. Dimana kata ke 1,2,3..N akan dihitung kemunculannya (TF N) yang kemudian nilainya akan dikalikan dengan nilai IDF. Nilai dari setiap kata nantinya akan diproses Kembali kedalam algoritma C4.5

**Tabel 5 contoh hasil pembobotan TF-IDF**

| tweet | Indonesia | Diteruskan | Republik | Ancaman | Pusaka | Karya | Dimana | Hanya | … |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5.6278 | 12.982 | 0 | 0 | 0 | 0.2341 | 0 | 0 | … |
| 2 | 0 | 0 | 0 | 2.22397 | 0 | 0 | 2.5461 | 0 | … |
| 3 | 7.912 | 0 | 2.0918 | 0 | 0 | 0 | 0 | 0 | … |
| 4 | 0 | 0 | 0 | 0 | 12.204 | 3.429 | 0 | 4.123 | … |
| .. | .. | | | | | | | | |
| 72950 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … |

**Tabel 6 contoh hasil pembobotan TF-IDF Ngram**

| Dokumen | TFIDF_UNIGRAM | TFIDF_BIGRAM | TFIDF_TRIGRAM |
|---|---|---|---|
| 1 | 5.60588323 | 0.000 | 0.000 |
| 2 | 2.67160123 | 0.000 | 0.000 |
| 3 | 1.71379385 | 0.000 | 0.000 |
| 4 | 0.000 | 2.2344452 | 0.000 |
| 5 | 0.000 | 0.000 | 0.000 |
| …. | … | … | … |
| 77792 | 0.000 | 0.000 | 0.000 |

Pada fase pembobotan ini, kata-kata yang memiliki nilai TF-IDF yang tinggi dapat diurutkan terhadap topik dari sebuah dokumen, pengektraksian fitur ini dilakukan guna melihat kata yang berpengaruh untuk setiap tagar(#) atau label

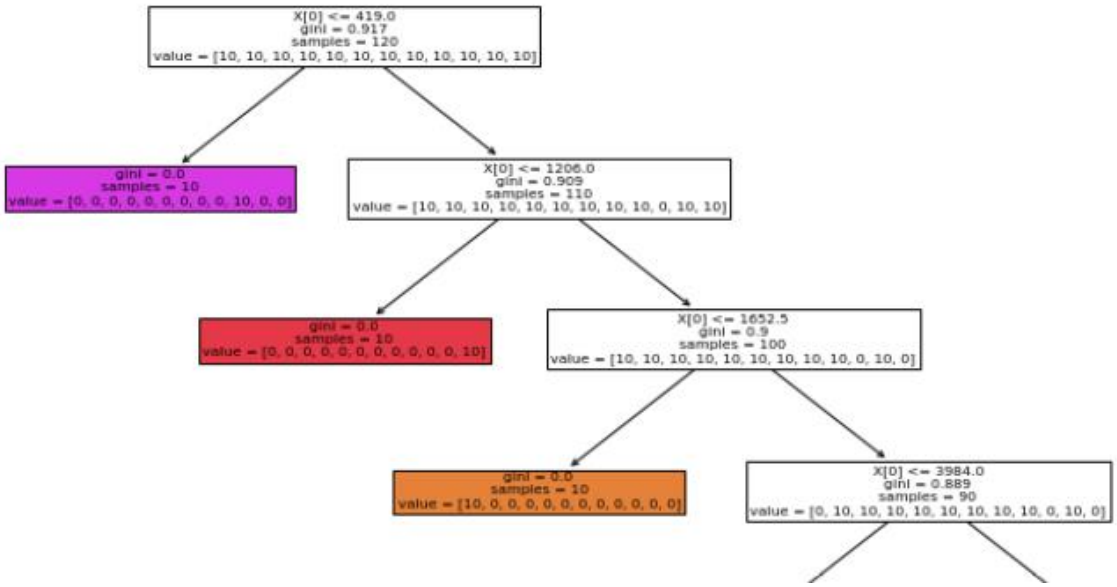
**Tabel 7 kata yang berpengaruh untuk setiap topik**

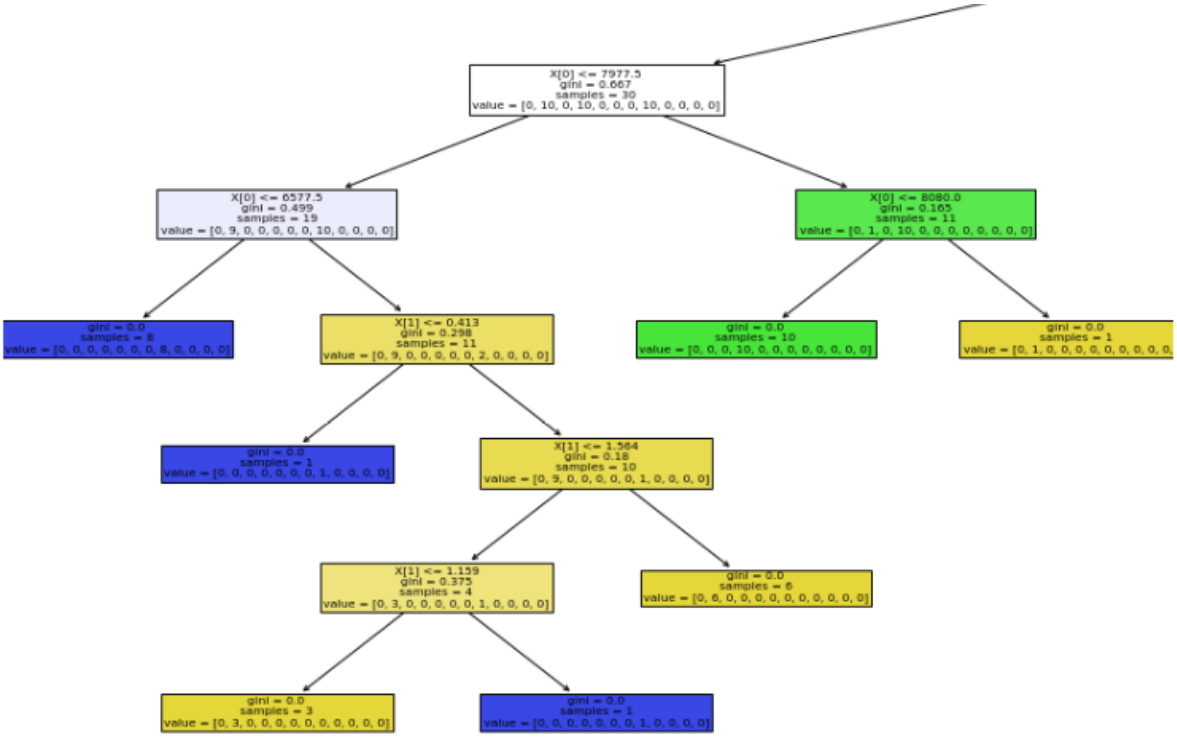
| Kata yang berpengaruh | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ekonomi | hiburan | hukam | olahraga | otomotif | pendidikan | politik | sosial | teknologi | umum | kesehatan | senbud |
| Gejolak | Trailer | Narkoba | Babak | motogp | Universitas | Ideologi | Erupsi | Apple | Nikah | Buah | Seni |
| Panggil | Movie | Ham | Tiket | Suzuki | Seleksi | Presiden | Perahu | Pintar | Kapan | Bahaya | Senbud |
| Jaman | Tv | Anak | Pssi | Honda | Kuliah | Republik | Tsunami | Android | Bersatu | Konsumsi | Budaya |
| Bank | Nonton | Keras | Ikutan | Toyota | Sbmptn | Anies | Gempa | Spesifikasi | Jabodetabek | Manfaat | Expo |
| saldo | Film | komnas | sub | mobil | snmptn | jokowi | longsor | google | nyala | tubuh | tari |

Pada tabel 7, Setelah melakukan pembobotan, hasil pembobotan yang didapatkan dapat digunakan untuk melihat kata-kata pembangun yang dominan pada setiap label dengan menggunakan ekstraksi fitur.

### *4.2.2 Decision tree C4.5* **(gunakan graphviz untuk visualisasi tree. Gunakan contoh dari dokumen)**

Selanjutnya pengerjaan metode decision tree C4.5, pada pengerjaan metode untuk menentukan akar dilihat dari *Gain Ratio* tertinggi hingga menghitung semua atribut yang bisa menghasilkan akar pohon, batang, dan daun. Representasi tree adalah sebagai berikut. Pada penjelasan pada bab 2 mengenai tentang perhitungan gain dan entropy, perhitungan tersebut menjadi parameter yang menciptakan cabang dan daun dari pohon keputusan C4.5 dan berikut adalah hasil pembetukan pohon keputusan yang terbentuk:






**Gambar 3 hasil pembentukan *decision tree***

### 4.2.3 Performansi

Selanjutnya menghitung performansi dari model prediksi yang telah dibuat, pada perhitungan performansi dilakukan percobaan sebanyak 5 kali dengan memilih data latih dan data uji secara acak kemudian dihitung nilai rata-rata performansinya di setiap skenario. Berikut adalah hasil akurasi, *recall* dan *precision* dan *f-1 score*

**Tabel 6 1analisis performansi**

| Dataset | Precision | Recall | Accuracy | F-1 score |
|---|---|---|---|---|
| 60:40 | **55%** | **54%** | **56%** | **55%** |
| 70:30 | **54%** | **55%** | **55%** | **54%** |
| 80:20 | **55%** | **56%** | **56%** | **55%** |
| 90:10 | **55%** | **55%** | **56%** | **55%** |
| Rata-rata | **54.75%** | **55%** | **55.75%** | **54.74%** |

Pada hasil percobaan diatas, percobaan dilakukan dengan menyatukan seluruh nilai TFIDF seluruh Ngram pada satu dokumen guna memperkecil penggunaan memori dan algoritma dilakukan berulang sebanyak 5 kali, didapatkan hasil tertinggi pada dataset 60:40 dengan akurasi terbaiknya yaitu 96%

### 4.3 Analisis hasil uji

Dari hasil percobaan menggunakan penggabungan nilai TF-IDF untuk setiap *term N-gram* serta penggunaan rasio dataset yang berbeda, analisis trending topik memiliki hasil akurasi terbesar sebesar 96% diperbandingan data latih dan uji sebesar 60:40 dengan pengulangan algoritma *decision tree c4.5* sebanyak 5 kali untuk mendapatkan nilai tree sebaik mungkin untuk data sebanyak 10% dari total data yang dibobotkan.

Pemilihan kata terbaik dipilih dengan mengurutkan nilai TFIDF total pada keseluruhan dokumen menggunakan fitur ektraksi chi2

## 5. Kesimpulan dan saran

Pada penelitian ini, Analisis Trending Topik pada media social *twitter,* penggunakan algoritma C4.5 dan pembobotan TF-IDF berjalan dengan cukup baik. Dari data yang diambil dari *API Twitter* selama bulan maret sampai dengan oktober 2019 masih didapatkan data yang cukup banyak dan cenderung merujuk pada satu topik. Jumlah data yang sangat banyak dapat mempengaruhi nilai pengujian dan akurasi model. Pengujian yang dilakukan dengan melakukan percobaan dengan beragam jumlah data dimana hasil dengan data sebanyak 72590 data mendapatkan hasil terbesar pada pengujian 60:40, 80:20 dan 90:10 dengan hasil sebesar 56%. disini penulis simpulkan bahwa banyak data, jumlah label per topik dan jumlah label keseluruhan pada dokumen, dapat mempengaruhi hasil dari pengujian yang dilakukan. Pada proses pembobotan TFIDF, dilakukan proses ekstraksi fitur dan mendapatkan kata yang dominan pada topik tertentu, dan disini penulis simpulkan bahwa pembentuk Ngram pada proses pembobotan itu sendiri tidak berasal dari kata yang dominan pada topik tertentu. Melainkan kata yang dominan pada keseluruhan dokumen

Pengujian terhadap data membuktikan algoritma C4.5 dapat digunakan untuk klasifikasi terhadap data trending topik twitter. sumber daya perangkat bisa menjadi alasan mengapa penggunaan data hanya digunakan sebesar 10% guna pengoptimalan hasil akhir, seperti proses *preprocessing* yang tidak membuang kata berbahasa non-indonesia, normalisasi kata, *retweet* yang masuk kedalam *tweet,* serta pembuangan huruf pada tautan yang tidak sempurna, membuat hasil pada pembobotan menjadi tidak sempurna atau dikatakan sebagai nilai yang tidak ada (0)

Saran untuk penelitian selanjutnya adalah untuk mempersiapkan fase *preprocessing* yang sempurna agar dokumen menjadi bersih tanpa adanya data yang dapat merubah nilai dari pembobotan. Penggunaan algoritma yang efisien sehingga proses yang tidak terlalu menggunakan banyak sumber daya, waktu dan memori pada perangkat dan dapat menghasilkan hasil yang akurat. Algoritma juga harus dirancang sehingga dapat menghasilkan hasil yang diinginkan

## DAFTAR PUSTAKA


[1] A. Farzindar and K. Wael, "a Survey of Techniques for Event Detection in Twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, 2015.

[2] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature Expansion using Word Embedding for Tweet Topic Classification," 2012.

[3] R. Lu and Q. Yang, "Trend Analysis of News Topics on Twitter," *Int. J. Mach. Learn. Comput.*, vol. 2, no. 3, pp. 327–332, 2013.

[4] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 251–258, 2011.

[5] J. Han, M. Kamber, and J. Pei, *Data Transformation by Normalization*. 2011.

[6] L. Egghe, "BRIEF COMMUNICATION A New Short Proof of Naranan ' s Theorem , Explaining Lotka ' s Law and Zipf ' s Law," *J. Am. Soc. Inf. Sci.*, vol. 1, no. 6, pp. 2581–2583, 2010.

[7] K. Adhatrao, A. Gaykar, V. Honrao, R. Jha, and A. Dhawan, "Predicting Students' Performance Using ID3 and C4.5 Classification Algorithms," *Int. J. Data Min. Knowl. Manag. Process*, vol. 3, no. 5, pp. 39–52, 2013.

[8] R. Aditya, "Pengaruh Media Sosial Instagram Terhadap Minat Fotografi Pada Komunitas Fotografi Pekanbaru," *Jom FISIP Oktober H.R. Soebrantas Km*, vol. 2, no. 12, pp. 1–14, 2015.

[9] M. H. Syahnur, M. A. Bijaksana, and M. S. Mubarok, "Kategorisasi Topik Tweet di Kota Jakarta , Bandung , dan Makassar dengan Metode Multinomial Naïve Bayes Classifier Tweet Topic Categorization in Jakarta , Bandung , and Makassar with Multinomial Naïve Bayes Classifier," *e-Proceeding Eng.*, vol. 3, no. 2, pp. 3612–3620, 2016.

[10] A. Arora, P. K. Malhotra, S. Marwah, A. Bhardwaj, and S. Dahiya, "Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets," 2012.

[11] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, no. September 2018, pp. 170–174,

2016.
[12] M. F. Arifin and D. Fitrianah, "Penerapan Algoritma Klasifikasi C4.5 Dalam Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada," *InComTech*, vol. 8, no. 2, pp. 87–102, 2018.
[13] K. B. J, "P enerapan A lgoritma C4 . 5 pada A nalisis ( Studi K asus : PT K ayu L apis A sli M urni )," no. October, 2016.
[14] M. Fitri, "Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi Tf-Idf Untuk Pencarian Dokumen Berbahasa Indonesia," *Peranc. Sist. Temu Balik Inf. Dengan Metod. Pembobotan Komb. Tf-Idf Untuk Pencarian Dok. Berbahasa Indones.*, 2013.
[15] "Tf-idf - Wikipedia." .
[16] Evi P. Marpaung, "IMPLEMENTASI CONTENT BASED VIDEO RETRIEVAL MENGGUNAKAN SPEEDED-UP ROBUST FEATURES (SURF)," pp. 5–32, 1998.
[17] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.