Analisis Perbandingan Pembobotan TF-IDF dan TF-RF pada *Trending Topic* di Twitter dengan Menggunakan Klasifikasi *K-Nearest Neighbor*

Agung N Assidyk[1], Erwin Budi Setiawan, S.Si., M.T.[2], Isman Kurniawan S.Pd., M.Si., M.Sc., Ph.D[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]agungas@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id,
[3]ismankrn@telkomuniversity.ac.id,

Abstrak

Media sosial yang sedang berkembang saat ini adalah twitter. Twitter merupakan media sosial yang di dalamnya berisikan informasi seperti biografi seseorang, informasi, *tweet* atau cuitan dari penggunanya. Informasi yang didapatkan dari twitter dapat dimanfaatkan untuk memprediksi suatu topik yang sedang tren atau trending. Pada penelitian ini membahas perbandingan metode pembobotan yang digunakan di suatu topik yang sedang *trending topic* yaitu TF-RF dan TF-IDF untuk memberikan suatu nilai/bobot pada term yang terdapat pada suatu dokumen. dan menggunakan metode pengklasifikasian dari data mining dimana metode yang digunakan adalah metode pengklasifikasian *K-Nearest Neighbor,* Hasil penelitian dilakukan berdasarkan berita dan percakapan diambil dari media twitter. Akurasi *K-Nearest Neighbor* nilai terbaik mengunakan K=1 dengan pembagian data training dan data testing (90:10) pembobotan TF-IDF adalah 63,12% dengan *precision* 0,633 dan *recall* 0,633 sedangkan TF-RF yaitu 62,48 % dengan *precision* 0,623 dan *recall* 0,623.

Kata kunci : Trending, TF-IDF, TF-RF, *K-Nearest Neighbor*

Abstract

The social media that is currently developing is Twitter. Twitter is a social media that contains information such as a person biography, information, tweets or tweets from users. Information obtained from Twitter can be used to predict a *trending topic*. This research discusses comparison of the weighting methods used in a trending topic, that is TF-RF and TF-IDF to give a weight to the term contained in a document. and using the classification method of data mining where the method used is the *K-Nearest Neighbor* classification method. The results of the study are based on news and conversations taken from Twitter . Accuracy of *K-Nearest Neighbor* the best value using K = 1 with the distribution of training data and testing data (90:10) weighting TF-IDF is 63,10% with *precision* 0.633 and *recall* 0.633 while TF-RF is 62,48% with *precision* 0.623 and *recall* 0.623.

Keywords: *Trending*, TF-IDF, TF-RF, *K-Nearest Neighbor*

1. Pendahuluan

Pada saat ini media sosial sangat memiliki peran penting di masyarakat untuk berkomunikasi, informasi dan bisa dimanfaatkan untuk sarana pembelajaran. Media sosial adalah suatu tempat dimana seseorang bisa mengespresikan dirinya sebebas mungkin kepada dunia, mengungkapkan detail dan wawasan pribadi kedalam kehidupan mereka [1]. Twitter adalah layanan *microblogging* yang banyak digunakan untuk menberi informasi dan mendapatkan informasi dan layanan ini menyediakan fitur mengirimkan pesan ke dalam aplikasi yang disebut *tweet*. Karena adanya batasan pada Twitter, maka setiap *tweet* yang ditulis oleh pengguna mempunyai banyak variasi singkatan kata, dan tidak menggunakan tata bahasa yang benar. Banyaknya variasi membuat *tweet* pada Twitter sulit dipahami[11]. *Tweet* yang dikirimkan oleh pengguna dapat mengandung sebuah topik pembahasan, topik yang menarik dan dibicarakan oleh banyak orang dalam jangka waktu yang singkat dapat disebut dengan *trending topic* [2]. Analisis terhadap media sosial adalah untuk mengetahui suatu informasi dan berita mengenai sebuah informasi, preferensi dan opini masyarakat. Beragam topik pembicaraan yang kemudian diklasifikasikan berdasarkan kelompok yang lebih umum dimana beberapa topik pembicaraan dapat digolongkan sebagai topik pembicaraan yang sama atau kategori tertentu. Penggolongan trend ke dalam kategori tertentu tersebut [3]. *Trending topic* menjadi hal yang penting, karena dalam keadaan darurat seperti kecelakaan, bencana alam, dan terorisme, *trending topic* memberikan laporan yang lebih cepat dibandingkan media konvensional [11]. Oleh karena itu *trending topic* dapat diteliti, dengan data yang banyak. Untuk mempermudah mengetahui informasi dan berita yang sedang *trending*, pengguna dapat membaca terhadap hashtag dan membaca setiap tweet yang muncul untuk mengetahui informasi atau berita yang akurat pada *trending topic*.

Penelitian terkait yang telah dilakukan oleh Syarif, Anwar dan Dewiani tentang prediksi *trending topic* dengan algoritma *K-Nearest Neighbor* yang dioptimasi, penelitian tersebut menggunakan sebanyak 2007 data dan menghasilkan nilai akurasi 81,13% [4]. Tujuan dari penelitian ini adalah untuk mengetahui perbandingan penggunaan metode TF-IDF dan TF-RF dengan mengengunakan klasifikasi *K-Nearest Neighbor* dalam menentukan *trending topic*. Penelitian ini data yang digunakan adalah sebanyak 77793 data *tweet* yang didapat dengan cara crawling dan pelabelan dilakukan secara manual, data yang telah memiliki label akan di *preprocessing* untuk mengubah data menjadi lebih baik lalu data dibobotkan menggunakan metode TF-IDF dan TF-RF, data yang telah diberi bobot akan dibagi menjadi data *train* dan data *test* digunakan untuk proses klasifikasi, *tweet* akan diklasifikasikan. Batasan masalah dalam tugas akhir ini adalah hanya menggunakan 12 kategori dan tidak membandingkan dengan klasifikasi lain. Untuk setiap kategori dalam menentukan suatu data *tweet* yang dikategori, ditentukan oleh penulis, beserta anggota kelompok yang lain. Data yang digunakan pada penelitian ini berasal dari tweet berbahasa Indonesia dari media social Twitter. Tweet tersebut diambil dari bulan Juli 2019 sampai bulan Agustus 2019, dengan jumlah tweet yang berbeda setiap harinya.

2. Studi Terkait

2.1 *Trending Topic*

Trending Topic adalah mengambarkan suatu kejadian yang sedang dibicarakan atau dibahas pada Twitter. Hal tersebut membuat para penguna Twitter (user) untuk membahas kejadian tersebut di media Twitter dan bilamana kejadian tersebut semakin menarik untuk dibahas oleh penguna Twitter(user) di media twitter akan semakin trending oleh karena itu semakin banyak user yang membahasnya, maka kejadian itu semakin terkenal kejadian tersebut [5]. Untuk membuat sebuah *Trending Topic* pengguna dapat menambahkan berupa *hashtag* sebagai kata kunci seperti (#Timnas). *hashtag* tersebut secara otomatis bertindak sebagai tag untuk memberikan sedikit keterangan yang sedang dibicarakan. dimana beberapa topik pembicaraan dapat digolongkan sebagai topik pembicaraan yang sama atau kategori tertentu. Penggolongan trending ke dalam kategori tertentu tersebut [3].

2.2 *Preprocessing*

Preprocessing merupakan suatu tahap untuk memperbaiki data agar lebih terstruktur diproses pada setiap dokumen yang digunakan. *Preprocessing* bertujuan untuk menyiapkan data sebelum menentukan yang sedang *Trending Topic* [4], berikur merupakan tahapan dari *Preprocessing* [4] :

a. *Case Folding* adalah merubah semua huruf dalam dokumen menjadi huruf kecil.
b. *Tokenizing* adalah memecah kalimat menjadi sebuah kata.
c. *Filtering* adalah menghilangkan suatu kata pada sebuah dokumen dengan mengunakan cara *stopword*. *Stopword* adalah kata yang diperlukan dalam dokumen agar data yang digunakan akan lebih baik.
d. *Stemming* untuk memproses kata menjadi kata dasar, dengan cara menghapus bagian imbuhan.
e. Normalisasi teks untuk kata singkatan menjadi kata baku menggunakan kamus

2.3 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term Frequency-Inverse Document Frequency merupakan salah satu metode pembobotan yang menggabungkan antara *term frequency* (TF) dan *inverse document frequency* (IDF) atau pembobotan pada setiap kata dalam setiap dokumen teks [6]. *Term frequensy*(TF) adalah kemunculan sebuah kata dalam suatu dokumen sedangkan *inverse document frequency* (IDF) adalah jumlah seluruh dokumen yang mengandung kata tetentu, Maka metode TF-IDF ini dapat menghitung total bobot dari kata dalam sebuah dokumen dengan persamaan [6]:

$$tf_{td}idf_t = tf_{td} * log\left(\frac{N}{df_t}\right) \quad (2.1)$$

Dimana
$tf * idf$: Bobot total dari kata t
tf_{td} : Jumlah kemunculan kata t dalam suatu dokumen
N :Total dokumen
df_t :Jumlah dari seluruh dokumen yang mengangdung kata t

2.4 *Term Frequency -Relevance Frequency* (TF-RF)

Relevance Frequency atau disebut juga RF, adalah sebuah metode yang tergolong baru, yang muncul sebagai upaya perbaikan terhadap metode-metode yang sudah ada. metode ini mempertimbangkan relevansi dokumen dilihat dari frekuensi kemunculan term di kategori yang berkaitan[7]. Pada RF, bobot dari suatu term dihitung dengan menggunakan persamaan [7]:

$$tf_{td}rf = tf_{td} * log(2 + \frac{b}{max1, c}) \quad (2.2)$$

Dimana:
$tf * rf$: Pembobotan dokumen ke dalam model ruang vektor
tf_{td} : Jumlah kemuncualan kata t dalam dokumen
b : Jumlah dokumen yang mengandung kata t
c : Jumlah dokumen yang tidak mengadung kata t

2.5 *K-Nearest Neighbor* (KNN)

Klasifikasi merupakan sebuah metode untuk mengelompokan data. *K-Nearest Neighbor* adalah algoritma yang befungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train dataset*), diambil dari k tetangga terdekatnya (*K-Nearest Neighbor*), dengan k merupakan banyaknya tetangga [8]. Untuk mengetahui jarak antara dua titik yaitu *data train* (x_1) dan *data test* (x_2) mengunakan rumus *euclidean distance*. Berikut merupakan perhitungan [12] :

$$d = \sqrt{\sum_{i=1}^{N} (x_{2\,i} - x_{1\,i})^2} \quad (2.3)$$

Dimana :
d : Jarak antara titik
x_1 : Data *training*
x_2 : Data *testing*
i : Variable

2.6 *Confusion Matrix*

Pengukuran performa dikerjakan untuk memperoleh tingkat kesesuaian dari hasil klasifikasi dari sistem yang di bangun. Beberapa cara yang sering di gunakan pada pengukuran performa dengan menghitung akurasi total, *Recall, Precission dan Accuracy* dengan mengunakan persamaan sebagai berikut :

Tabel 1 nilai prediksi

	Nilai Sebenarnya		
	Kelas	TRUE	FALSE
Nilai Prediksi	TRUE	TP (True Positive) Corect result	FP (False Positive) Unexpected result
	FALSE	FN (False Negative) Missing result	TN (True Negative) Corect absence of result

a. *Precision* adalah jumlah user yang dengan benar diklasifikasikan dalam sebuah kelas dibagi dengan jumlah total user yang diklasifikasikan dalam kelas tersebut [9]. Precision juga sering disebut sebagai tingkat ketepatan antara informasi yang diminta oleh pengguna. Berikut rumus dari *Precision* :

$$Precision = \frac{TP}{TP + FP} \quad (2.4)$$

b. *Recall* adalah jumlah user yang dengan benar diklasifikasikan dalam sebuah kelas dibagi dengan jumlah total user dalam kelas tersebut [9].*Recall* adalah tingkat keberhasilan system dalam menemukan kembali sebuah informasi berikut rumus *Recall :*

$$Recall = \frac{TP}{TP + FN} \quad (2.5)$$

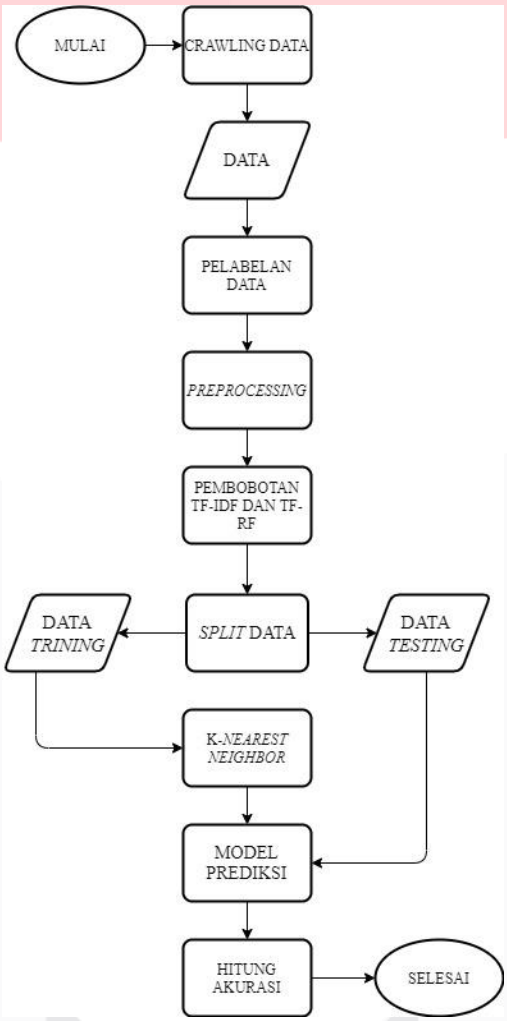
c. *Accuracy* di definisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual. Berikut rumus *Accuracy*:

$$Accuracy = \frac{TP}{TP + TN + FP + FN} \quad (2.6)$$

3. Sistem Prediksi *Trending Topic*

Rancangan sistem prediksi *Trending topic* dengan mengunakan klasifikasi *K- Nearest Neighbor* pada gambar 1

Gambar 1. Alur sistem



Pada penelitian ini, system yang dibuat adalah system yang mengimplementasikan metode pembobotan TF-IDF dan TF-RF dengan klasifikasi KNN untuk menentukan *Trending topic* pada media twitter. Data yang digunakan berupa tweet pada akun twitter dan melakukan crawling data menggunakan API twitter. Hasil yang didapat bisa mengklasifikasikan *Trending topic* pada twitter.

3.1 Crawling Data

Data yang digunakan berupa tweet dengan cara crawling merupakan tahap dalam penelitian yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu database proses crawling dibatasi oleh API Twitter sebanyak 100 tweet percrawling. Crawling data di Twitter dapat menggunakan dua sistem pencarian, by user dan by keyword [10]. Pengumpulan data dari penelitian ini yaitu data yang diunduh dari server twitter berupa user dan tweet beserta atribut-atributnya dan dibagi kedalam 12 katagori yaitu ekonomi, hiburan, hukam, kesehatan, olahraga, otomotif, Pendidikan, politik, seni budaya, sosial, teknologi, umum. Dasar pengunaan kategori tersebut dikarenakan pada media berita seringnya digunakan kategori tersebut dimedia berita untuk membedakan berita yang memiliki berbagai kategori Tabel 3. Untuk setiap kategori dalam menentukan suatu data tweet yang yang telah dikategorikan dengan jumlah data yang berhasil didapatkan pada tahap ini adalah sebanyak 77.793 tweet dan rentang waktu

dilakukan proses crawling adalah dari tanggal 28 Juli 2019 sampai dengan 25 Agustus 2019. Rincian tabel data yang digunakan pada Tabel 2.

Tabel 2. Jumlah data

No	Kategori	Keyword	Jumlah Data
1	Ekonomi	korupsi, pasar saham, investasi, ekonomi, bank, modal, warta ekonomi, menkoperekonomian, rupiah, inflasi	6094
2	Hiburan	NetMediaTama, Gundala, ZaraJKT48, bumilangit, NET TV, jagatsinema, wishnutama, koboykampus	7838
3	Hukam	FPI, #KPKJemputPaksaMendag, #PapuaBergejolakJaeKemana, semoga papua, #JaeKemana	6133
4	Kesehatan	Infokesehatan, info dokter, BPJS, obat kanker, makanan sehat	6553
5	Olahraga	#PersibDay, Persija, Timnas-u18, #TimnasDay, #IndonesiaOpen2019	6173
6	Otomotif	#GIIAS2019, MotoGPIndonesia, IIMS2019, GarasiDrift, mobil listrik	6356
7	Pendidikan	UTBK, Ruang Guru. #SBMPTN, #SBMPTN2019, #SNMPTN2019, masuk universitas	6042
8	Politik	Jokowi, #demokrasiTercorengNasiGoreng, #AniesdiBully, #ibukotabaru, #kalimantanibukota, #pindahibukota	6396
9	Senbud	HeritageBandung, Danau Toba, #FestivalDieng, DiengCultureFestival, GebyarNusantara, #SailNias2019, Kesenian Indonesia	6156
10	Sosial	BMKG, Erupsi, Tangkuban Perahu, Potensi Tsunami	7129
11	Teknologi	Facebook, 5G, GoogleIndonesia, Golang, Huawei, goyangshoope, Android	6272
12	Umum	#matilampu, #maribertemen, #malamminggu, terimakasih pln, giveaway, Jakarta, #rebahan	6648
Jumlah Data			77.793

Tabel 3. Media Berita

No	Media berita	Kategori										
		News	ekonomi	teknologi	entertaiment	olahraga	otomotif	travel	kesehatan	edukasi	gaya hidup	Politik
1	tribun	1		1	1	1	1		1	1	1	
2	detik	1		1	1	1	1	1	1		1	
3	liputan 6	1		1	1	1	1		1		1	
4	kompas	1	1	1	1	1	1	1		1	1	
5	sindonews			1		1				1	1	1
6	kumparan	1		1	1	1	1	1			1	
7	merdeka		1		1	1	1		1		1	1
8	CNN indonesia		1	1	1	1	1	1	1		1	1
9	Okezone	1	1	1		1	1		1		1	
10	Viva	1			1	1	1				1	
11	Tempo	1		1		1					1	
	Jumlah	8	4	9	8	11	9	4	6	3	11	3

3.2 Pelabelan

Pada proses pelabelan akan diberikan label atau pengelompokan berdasarkan katagori. Data pelabelan menggunakan data tweet yang telah diunduh lalu diberikan label sesuai dengan kategori agar mengetahui kategori.Contoh pelabelan tweet dengan kategori:

Tabel 4. Pelabelan

No	Tweet	Label
1	Pemain muda Timnas Indonesia tak minder latihan bareng senior! #Timnas #TimnasIndonesia #PSSI #KitaGaruda #Sepakbolaindo	Olahraga
2	Fabio Quartararo Bikin Posisi Rosisi Semakin Terancam #BeritaMotoGP #Yamaha	Otomotif
3	Ada fitur apa nih yg spesial dari #laptop @asusid ya #asusindonesia	Teknologi

3.3 *Preprocessing*

Data yang telah didapatkan akan diproses untuk mengubah datanya menjadi lebih baik, berikut merupakan tahapan yang dilakukan *preprocessing* :

a. *Cleanning*, pada proses ini menghilangakan karakter yang tidak digunakan seperti tanda baca , mention, hastag, lambang emotikon dan url.
b. *Case folding*, megubah huruf besar menjadi huruf kecil.
c. *Tokenizing*, membagi kalimat menjadi sebuah kata.
d. *Filtering*, menghilangkan suatu kata pada sebuah dokumen dengan mengunakan cara stopword.
e. *Stemming*, memproses kata menjadi kata dasar, dengan cara menghapus bagian imbuhan.
f. Normalisasi teks untuk kata singkatan menjadi kata baku menggunakan kamus

Tabel 5. Preprocessing

Id_tweet	sebelum di preprocesing	Ssesudah preprpcesing
20218207132	Ilmu Digital adalah Pengetahuan yang sangat dibutuhkan pada dunia bisnis	ilmu digital pengetahuan butuh dunia bisnis

3.4 Pembobotan

Pada proses ini data yang dihasilkan oleh *Preprocessing* akan mendapatkan bobot atau nilai dengan mengunakan metode pembobotan TF-IDF dan TF-RF. Dalam pembobotan yang dikerjakan bertujuan untuk mengukur dampak pengaruh kata dari suatu data atau dokumen dan Menentukan metode pembobotan kata yang paling baik diantara TF•IDF dan TF•RF. Sebelum masuk ke dalam klasifikasi data yang digunakan akan diproses sesuai rumus yang telah di jelaskan dibagian studi terkait perhitungan pembobotan dilakukan perhitungan per *id tweet* dibwah ini merupakan contoh perhitungan pembobotan:

Tabel 6. Contoh TF-IDF

TF-IDF	Kata				
	Sehat	Obat	Bandung	Makan	Banyak
TF 1	9	5	6	4	2
TF 2	1	0	0	0	0
IDF	0.8956	0.7985	0.5067	0.5589	0.6754
IDF 1	8.0604	3.9925	3.0402	2.2356	1.3508
IDF 2	0.8956	0	0	0	0

Tabel 7. Contoh TF-RF

TF-RF	Kata				
	Jadi	Minum	Bantu	Tingkat	Istirahat
TF 1	9	5	6	4	2
TF 2	1	0	0	0	0
RF	0.8956	0.7985	0.5067	0.5589	0.6754
RF 1	8.0604	3.9925	3.0402	2.2356	1.3508
RF 2	0.8956	0	0	0	0

3.5 Klasifikasi *K-Nearest Neighbor*

Pada tahapan ini data yang digunakan telah memiliki bobot yang telah diproses pada pembobotan sebelumnya setelah itu data diproses klasifiksi *K-Nearest Neighbor* dengan melihat pengaruh perbandingan metode pembobotan TF-IDF dan TF-RF terhadap scenario yang memiliki nilai tertinggi. Berikut merupakan langkah dalam menghitung klasifikasi *K-Nearest Neighbor* :

- Menentukan parameter k
- Menghitung jarak antara data yang dievaluasi dengan data testing.
- Memilih jarak yang terbentuk
- Menentukan jarak terdekat hingga k yang ditentukan
- Memasangkan kelas yang sesuai
- Mencari jumlah kelas tetangga terdekat
- Pembagian data training dan data testing dengan komposisi (50-50), (60-40), (70-30), (80-20), (90-10)
- Setelah dilakukan pengujian , diperoleh nilai k ditentukan

3.6 Kinerja Sistem

Pada tahapan ini untuk mengevaluasi kinerja sistem setelah mendapatkan hasil maka akan diproses mengunakan confusion matrix dpat mengukur hasil *accuracy*, *precision* dan *recall* dari sistem yang dikerjakan.

4. Hasil dan Analisis

4.1 Skenario pengujian

Pada penelitian ini agar dapat melihat hasil kinerja klasifikasi yang baik. Skenario pengujian dilakukan untuk mengetahui rasio pembagian data *train* dan data *test* yang terbaik dengan menggunakan 5 rasio, terdiri dari 50:50, 60:40, 70:30, 80:20, 90:10 dalam bentuk data *train* : data *test*. Setelah didapatkan rasio pembagian data yang terbaik. Tahapan awal data akan dibobotkan mengunakan metode TF-IDF dan TF-RF setelah itu data yang telah di bobotkan telah memiliki nilai/bobot. Selanjutnya proses klasifikasi terhadap data *test* dan data *train* menggunakan metode *K-Nearest Neighbor* (KNN).

4.2 Klasifikasi *K-Nearest Neighbor*(KNN)

Pada proses kalsifikasi data train dan data bobot yang telah diproses sebelumnya akan diproses prediksi klasifikasi menggunakan metode klasifikasi *K-Nearest Neighbor* .Setelah itu proses selanjutnya adalah mengklasifikasikan data test ke dalam 12 kategori yaitu ekonimi, hiburan, hukum, kesehatan, olahraga, otomotif, Pendidikan, politik, seni budaya, sosial, teknologi, umum .Hasil yang didapatkan dari tahap ini akan digunakan untuk menghitung nilai evaluasi kinerja sistem. Berikut merupakan hasil kinerja sistem yang telah dikerjakan:

Tabel 7.Hasil TF-IDF

TF-IDF

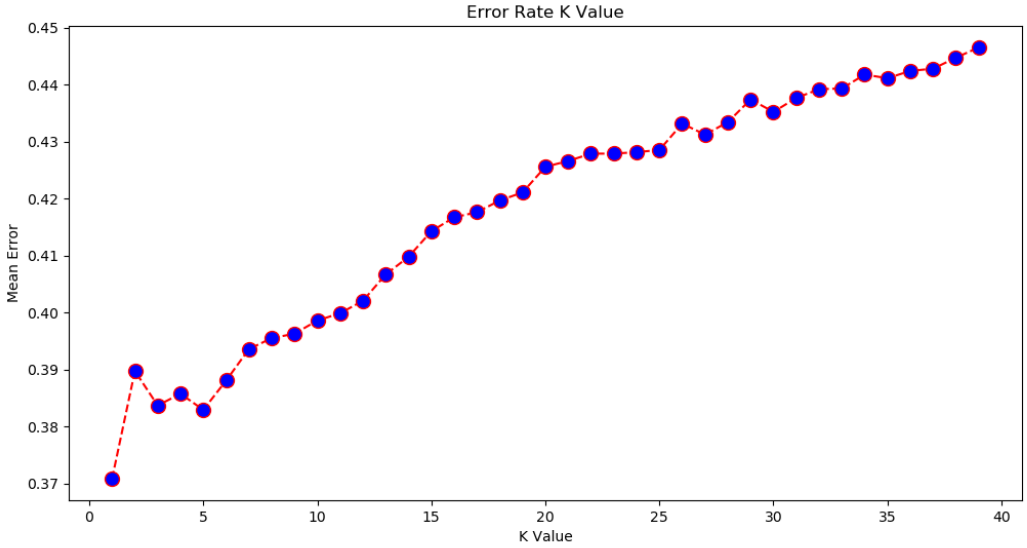
No	Komposisi	Precision				Recall				Accuracy			
1	(50-50)	0,59	0,59	0,58	0,573	0,59	0,576	0,573	0,57	59,45%	57,63%	57,35%	56,92%
2	(60-40)	0,606	0,60	0,59	0,583	0,606	0,59	0,586	0,58	60,42%	58,69%	58,49%	57,89%
3	(70-30)	0,613	0,61	0,60	0,59	0,613	0,60	0,596	0,586	61,46%	59,67%	59,59%	58,66%
4	(80-20)	0,623	0,62	0,61	0,60	0,62	0,606	0,60	0,593	62,16%	60,57%	60,35%	59,54%
5	(90-10)	0,633	0,633	0,62	0,606	0,633	0,613	0,616	0,603	63,12%	61,63%	61,47%	60,25%
Jumlah K		K=1	K=3	K=5	K=7	K=1	K=3	K=5	K=7	K=1	K=3	K=5	K=7

Tabel 8.Hasil TF-RF

TF-RF

No	Komposisi	Precision				Recall				Accuracy			
1	(50-50)	0,596	0,593	0,58	0,57	0,596	0,58	0,573	0,57	59,52%	57,84%	57,53%	56,74%
2	(60-40)	0,61	0,603	0,59	0,583	0,61	0,59	0,586	0,576	60,61%	58,80%	58,45%	57,75%
3	(70-30)	0,613	0,613	0,60	0,59	0,613	0,596	0,59	0,59	61,45%	59,68%	59,28%	58,81%
4	(80-20)	0,62	0,626	0,61	0,60	0,62	0,61	0,606	0,60	61,64%	60,97%	60,47%	59,61%
5	(90-10)	0,623	0,63	0,613	0,61	0,623	0,616	0,606	0,606	62,48%	61,65%	60,81%	60,58%
Jumlah K		K=1	K=3	K=5	K=7	K=1	K=3	K=5	K=7	K=1	K=3	K=5	K=7

Gambar 1.Tingkat error setiap k



4.3 Analisis Hasil Pengujian

Berdasarkan hasil pengujian yang dilakukan, maka penulis mengambil analisis mengenai hasil yang diperoleh, proses dari *preprocessing* berpengaruh pada hasil ouput data saat melakukan pengujian pada sistem yang dibuat, data yang telah di crawling dan telah di labelling sangat berpengaruh pada hasil akurasi. Analisis hasil pengujian dari skenario yang telah dijabarkan sebelumnya, telah diperoleh hasil pengujian. Skenario adalah perbandingan antara TF-IDF dan TF-RF dengan pembagian data (50:50), (60:40), (70:30), (80:20), (90:10) dan mengunakan nilai K = 1 ,3 ,5 ,7 , setiap komposisi dilakukan pengujian sebanyak 3 kali dan komposisi dengan rata-rata hasil yang terbaik akan dipilih sebagai nilai terbaik dari seluruh perbanding pembobotan antara TF-IDF dan TF-RF yang diklasifikasi oleh *K-Nearest Neighbor* dan hasil yang didapat pada 5 komposisi dan pada setiap k yang telah di uji. Diantara semua k yang diuji hasil yang paling optimal k=1 namun hasil pengujian dengan k lain tidak memiliki perbedaan hasil performansi terlalu signifikan 1%-3% maupun dari pembobotan TF-IDF dan TF-RF bahwasannya semakin besar nilai k maka akan mempengaruhi hasil performansi akan menurun. Rata rata terbaik adalah pembobotan TF-IDF dengan pembagian data (90:10) hasil akurasi yang didapat 63,12% dengan *precision* 0,633 dan *recall* 0,633 sedangkan hasil terbaik yang dimiliki oleh TF-RF dengan pembagian data (90:10) hasil akurasi yang didapat 62,48% dengan *precision* 0,623 dan *recall* 0,623. Berdasarkan hasil pengujian, bahwa nilai *precision, recall, Accuracy* untuk TF-IDF lebih baik dibandingkan metode TF-RF pada pengujian yang dilakukan karena Setiap metode pembobotan memiliki kelebihan dalam menggunakan metode untuk mendapatkan performansi yang lebih baik dan perbandingan antara metode IDF hanya akan menilai term berdasarkan kemunculan sedangkan metode RF mempertimbangkan hubungan dokumen dilihat dari frekuensi kemunculan term di kategori yang berkaitan dan memperhitungkan *term*(kata) pada suatu dokumen muncul serta menormalisasikan kemuculan *term*(kata) tersebut ke keseluruhan dokumen. Setelah semua skenario pengujian dilakukan, maka langkah selanjunya adalah untuk melakukan prediksi *trending topic* Tabel 9 kata *trend*.

Table 9 Kata Trend

Kata Trends Pada Kategori

No	Label	Kata Trends
1	Hiburan	nonton, gundala, film, episode, tiket, sinema, tonton, rilis, lagu, bumilangit
2	Hukam	kasus , hukum , rusuh , perintah , lindung , adil , jaksa , hakim , dukung
3	Kesehatan	sehat, dokter, obat, sakit, tubuh, badan, tingkat, manfaat, cegah, jantung
4	Olahraga	main, piala, timnas, final, menang, liga, juara, laga, stadion, tim
5	Otomotif	mobil, honda, motor, balap, otomotif, modifikasi, mesin, kendara, yamaha
6	Pendidikan	sekolah, mahasiswa, guru, didik, lulus, osis, beasiswa, pramuka, kuliah,
7	Politik	presiden, gubernur, pimpin, pidato, dukung, mpr, partai, politik, negara
8	Senbud	budaya, festival, seni, wisata, destinasi, nusantara, tari, wisatawan
9	Sosial	gempa,tsunami,potensi, gunung, selamat, erupsi, doa, kedalaman, tangkuban
10	Teknologi	harga, android, ponsel, google, aplikasi, data, rilis, teknologi, internet, spesifikasi
11	Umum	lampu, listrik, buat, nyala, padam, pln, ganggu, dirut, berkat, akibat
12	Ekonomi	ekonomi, uang, usaha, bank, investasi, saham, dagang, modal, saldo, industri

5. Kesimpulan

Penelitian ini mempuayai tujuan untuk membandingkan kinerja pembobotan TF-IDF dan TF-RF terhadap proses kinerja klasifikasi *K-Nearest Neighbor* untuk menentukan *trending topic*. Dari hasil pengujian dengan jumlah data 77793 data tweet didapat dari media twetter yang dilabelkan secara manual dibagi kedalam 12 katagori yaitu ekonimi, hiburan, hukum, kesehatan, olahraga, otomotif, Pendidikan, politik, seni budaya, sosial, teknologi, umum. Berdasarkan hasil pengujian implementasi pembobotan TF-IDF dan TF-RF terhadap klasifikasi *K-Nearest neighbor* mendapatkan hasil akurasi tertinggi mengunakan k = 1 dengan skenario (90-10) dan hasil akurasi didapat adalah 63,12% dengan *precision* 0,633 dan *recall* 0,633. Dalam hal ini kinerja perbandingan antara TF-IDF dan TF-RF dengan mengunakan klasifikasi *K-Nearest Neighbor* bahwasannya TF-IDF lebih baik dalam *confution martix* tersebut.

Daftar Pustaka

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter , a Social Network or a News Media?," Int. World Wide Web Conf. Comm., pp. 1–10, 2010.

[2] S. Saquib and R. Ali, "Understanding dynamics of trending topics in Twitter," Proceeding - IEEE Int.Conf. Comput. Commun. Autom. ICCCA 2017, vol. 2017-Janua, pp. 98–103, 2017.

[3] Agustina, P. A., Matulatan, T., Tech, M., & Si, M. B. S. (2012). Klasifikasi Trending Topic Twitter dengan Penerapan Metode Naive Bayes,

[4] S. Syarif, Anwar, and Dewiani, "Trending topic prediction by optimizing K-nearest neighbor algorithm,"Proc. 2017 4th Int. Conf. Comput. Appl. Inf. Process. Technol. CAIPT 2017, vol. 2018-Janua, pp. 1–4,2018.

[5] Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. Icwsm, 1–17.

[6] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," Procedia Eng., vol. 69, pp. 1356–1364, 2014.

[7] Lan,Man. 2006. A New Term Weighting Method for Text Categorization. National University of Singapore. Singapore.

[8] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," Procedia Eng., vol. 69, pp. 1356–1364, 2014.

[9] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition, San Fransisco: Morgan Kauffman Publishers, 2012

[10] J. E. Sembodo, E. B. Setiawan, and A. Baizal, "Data Crawling Otomatis pada Twitter," no. September, pp. 10–16, 2016.

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

[12] Kursini & Luthfi E.T. *Algoritma data Mining*, ANDI,Yogyakarta, 2009