Analisis Trending Topik Pada Twitter menggunakan Metode Naive Bayes dengan Pembobotan TF-IDF

Saut Sihol Ritonga[1], Erwin Budi Setiawan[2], Isman Kurniawan[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]ritongasihol@gmail.com, [2]erwinbudisetiawan@telkomuniversity.ac.id,
[3]ismankrn@telkomuniversity.ac.id

Abstrak

Twitter adalah salah satu media sosial yang banyak sekali para penggunanya menceritakan berbagai macam banyak kejadian oleh karena itu perlu megklasifikasi topik menjadi dengan akurasi tinggi untuk lebih baik pengambilan informasi. Oleh karena itu penulis melakukan penelitian untuk mengatasi masalah ini dengan membagi beberapa tren topik *twitter*. Pembobotan yang digunakan adalah TF-IDF dengan menggunakan Naïve Bayes. Akurasi terbaik pada pembobotan TF-IDF menggunakan klasifikasi Naïve Bayes didapat pada skenario data training, data tesing 80:20 adalah 57.08% dan memiliki nilai *f-measure* 0.52. Trending pertama yang terdeteksi dari pengambilan data dari bulan 25 Juli sampai 28 Agustus adalah politik dengan persentase 26.88% lalu kedua senbud persentase 8.65% dan yang ketiga hukam 8.27%.

Kata Kunci: *Twitter*, Naive Bayes, TF-IDF, Topik.

Abstract

Twitter is one of the many social media users who tell a wide variety of events so it is necessary to classify topics into high accuracy for better information retrieval. Therefore, the authors conducted research to overcome this problem by dividing a number of Twitter topic trends. The weighting used is TF-IDF by using Naïve Bayes. The best accuracy on TF-IDF weighting using the Naïve Bayes classification is obtained in the training data scenario, the 80:20 testing data is 57.08% and has an f-measure value of 0.52. The first trend detected from data collection from July 25 to August 28 is politic with a percentage of 26.88%, then second senbud with a percentage of 8.65% and a third with 8.27% hukam.

Keyword*: Twitter*, Naive Bayes, TF-IDF, Topic.

1. Pendahuluan

Perkembangan teknologi informasi sampe saat ini semakin pesat dan memberi banyak manfaat di semua aspek sosial. Perkembangan teknologi informasi sangat membantu manusia dalam menyelesaikan suatu pekerjaan yang harus diiringi dengan tenaga Sumbe Daya Manusia (SDM). Perkembangan teknologi informasi dipermudah dengan adanya sosial media aplikasi seperti *Facebook, Twitter, Path, Snapchat, Instagram, Telegram* dan sebagainya. Dan salah satu teknologi informasi yang sampe saat ini paling sering di gunakan adalah *twitter.*

Twitter adalah situs mikroblog yang sangat populer[1], tempat pengguna mencari informasi sosial dan waktu yang tepat seperti berita terkini, posting tentang selebritas, dan trending topic. Pengguna memposting pesan teks pendek yang disebut tweet, yaitu dibatasi oleh 140 karakter dan dapat dilihat oleh pengikut pengguna. Siapa pun yang memilih untuk memiliki yang lain tweet yang diposting di timeline satu disebut pengikut. Tweet telah digunakan sebagai media untuk informasi real-time atau banyak dibicarakan banyak orang diseminasi ini telah digunakan dalam berbagai kampanye, pemilihan umum, dan sebagai media berita[2]. Bahkan Indonesia menduduki peringkat 5 di Dunia pengguna *Twitter*. persentase yang sangat tinggi dari trending topik adalah tagar dimana tagar tesebut mengatas namakan orang lain atau kata kata lainnya makanya perlu sekali mengklasifikasi topik ini kedalam kategori umum sehingga mudah di pahami dalam pengambilan informasi. Untuk mendapatkan data berita dan tweet dari *Twitter* kita lakukan crawling. Crawling data merupakan tahap dalam penelitian yang bertujuan untuk mengumpulkan data atau mengunduh data dari suatu database. Pengumpulan data dari penelitian ini yaitu data yang di unduh dari server *Twitter* berupa user dan tweet beserta atribut-atributnya[9].

Trending Topic adalah suatu kejadian yang paling terkenal dan terjadi di dunia nyata dan banyak dibahas di media sosial terutama pada *Twitter*. Kejadian-kejadian yang terkenal ini membuat semua orang terutama para pengguna *Twitter* tertarik untuk membahasnya di media sosial. Semakin banyak pengguna media sosial yang membahas kejadian tersebut, maka kejadian itu semakin terkenal[7].

Jika ingin mengetahui informasi dan berita yang sedang Trending Topic, pengguna bisa klik Hashtag tersebut, maka muncul hasil tweet dari Hashtag tersebut. Tetapi untuk melakukan hal itu sangat mempersulit pengguna melihat berita trending topik, karena pengguna harus baca satu persatu tweet untuk lebih dulu mengetahui informasi atau berita yang sedang trending. Maka dari itu untuk melakukan analisa Trending Topic pada *Twitter* diperlukan mengubah tweet menjadi data yang mempunyai makna, dan diperlukan metode penelitian yang dapat melakukan analisa dengan cara mengklasifikasikan teks dari *tweet* pada *Twitter*. Maka dalam penelitian ini yang digunakan untuk pengkategorian masing-masing fitur di tentukan metode pembobotan TF-IDF. Kemudian melakukan penggolongan trend menggunakan metode Naïve Bayes[3]. Metode ini klasifikasi kata dari topik pembicaraan yang sama dengan membandingkan setiap fitur yang dimiliki oleh setiap kategori[10].

Batasan dalam mengerjakan topik tugas akhir ini adalah hanya mencakup tweet berita yang terdapat pada akun-akun di Indonesia untuk digunakan sebagai data. Data yang digunakan adalah 77.793 tweet diambil perbulan 25 Juli sampai 28 Agustus. Data yang di gunakan sebagai data training dan testing. Tujuan yang ingin dicapai mendapatkan *trending topik* sesuai dengan kejadian yang terjadi di dunia nyata.

2. Studi Terkait

2.1 Media Sosial

Pengertian media sosial menurut beberapa ahli seperti McGraw Hill Dictionary adalah sarana yang digunakan orang-orang untuk berinteraksi satu sama lain dengan cara menciptakan, berbagai, serta bertukar informasi dan gagasan dalam sebuah jaringan dan komunitas virtual[12].

Media sosial seperti *facebook, Twitter, instagram* dan sebagainya. Situs ini memungkinkan seseorang untuk membantu halaman web pribadi dan terhubung dengan teman-temannya untuk berbagi informasi. Selain itu *Facebook* merupakan situs teknologi informasi yang digunakan *user* untuk memberi informasi dan berbagi foto, video, atau lainnya tanpa batas.

Twitter pada saat ini adalah salah satu media sosial yang paling sering banyak digunakan. *Twitter* ini memiliki aplikasi pengembangan *yaitu Application Programming Interface* (API). *Twitter* API dapat mempermudah para pengembang untuk mendapatkan suatu informasi dan mengambil serta mengolah data melalui tweet pengguna *Twitter*. Dan *Twitter* memiliki fitur trending topik. Pengguna *Twitter* biasanya menggunakan "tagar" untuk berpartisipasi dalam topic yang sedang hangat. Untuk menggunakan tagar, anda memasukan berita topik setelah simbol hash (#) dalam tweet anda [4].

2.2 Pre-Processing

Pre-processing data merupakan langkah-langkah dalam mengolah data mentah yang berikutnya akan dimasukan ke dalam sistem klasifikasi. Proses ini bertujuan untuk menyiapkan data yang akan digunakan secara efisien ke dalam sistem klasifikasi dan membuat data menjadi data yang berkualitas (input data untuk data mining tools) [15].

a) *Case Folding*, proses dimana kata atau frasa masuk teks tweet akan dikonversi menjadi huruf kecil (a ke z). Itu membantu mengatasi masalah ketika kata-kata ditulis dengan kapitalisasi berbeda.
b) *Tokenization*, dilakukan untuk memotong input tweet ke oleh kata-kata yang membentuknya. Pada prinsipnya, ini memisahkan setiap kata dalam teks tweet Proses ini termasuk penghapusan angka, tanda baca, dan karakter selain dari huruf alfabet. Karakter-karakter ini dianggap sebagai pemisah kata (pembatas) sehingga mereka akan dihapus untuk mencegah Terjadinya kebisingan dalam proses lebih lanjut.
c) *Stop Removal*, adalah adalah menghilangkan kata-kata non-topikal yang tidak dianggap penting seperti: "dan", "ini", "itu", "Adalah", "atau", "yang", "via", dan lainnya. Preprocessing ini membantu mengurangi fitur yang tidak relevan dalam data.
d) *Stemming*, adalah proses menemukan akar a kata dengan menghilangkan awalan, infiks, sufiks, dan konfiks (a kombinasi awalan dan akhiran) pada kata turunan. Oleh berasal, variasi kata yang memiliki akar yang sama akan dianggap sebagai token (fitur) yang sama. Dalam Informasi Pengambilan, ini membantu meningkatkan kinerja pengambilan.

2.3 Term Weighting

Term weighting adalah sebuah metode pembobotan kata (term) untuk memberikan sebuah bobot atau nilai untuk kata (term) yang terkandung dalam sebuah dokumen. Bobot nilai ini menjadi ukuran besarnya jumlah dan tingkat kontribusi sebuah kata (term) untuk penentuan suatu kelas atau kategori dalam suatu dokumen. Terdapat beberapa metode pembobotan kata (term weighting) diantaranya adalah TF, TF-IDF, WIDF, TF-CHI, dan TF-RF. Dalam penelitian tugas akhir ini mencoba menguji pembobotan TF-IDF.

2.3.1 *Term Frequency Inverse Document Frequency*

Langkah wajib suatu proses peringkasan menggunakan pendekatan ekstraktif [16]. Seberapa sering kata muncul dalam dokumen menunjukan tingkat kepentingan kata tertentu pada suatu dokumen. TF adalah Kemunculan kata tertentu dalam suatu dokumen[16]. TF-IDF dapat ditulis dengan persamaan:

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

Dengan $tfidf_t$ bobot dari term *t,* $f_{t,d}$ frekuensi munculnya *term t* pada dokumen *d,* *N* jumlah kumpulan dokumen, df_t jumlah dokumen yang mengandung *term t.*

2.4 Naïve Bayes

Naive bayes adalah algoritma yang menggunakan teorema bayes untuk mengklasifikasi objek. Algoritma ini di anggap indepedensi yang kuat antara atribut,point,dan data. Klasifikasi ini banyak di gunakan untuk pembelajaran mesin karena mudah di terapkan [5]. Selain itu naive bayes dinyatakan sebagai algoritma yang memiliki sifat kesederhanaan,kekokohan,dan memiliki akurasi yang tinggi[6]. Persamaan NBC [11] :

$$P(A|B) = \frac{P(B|A)xP(A)}{(P(B))} \qquad (2)$$

Dengan A Data kelas belum diketahui, B Hipotesis data merupakan suatu kelas spesifik, P(A|B) Probalitas terjadinya nilai A jika B diketahui. Disebut probalitas posterior,karena peluang A bergantung pada nilai B, P(B|A) Probalitas terjadinya nilai B jika A diketahui. Disebut likelihood function, P(A) Probabilitas prior A yang mendahului terjadinya B. Disebut "prior", karena nilainya bisa diperoleh tanpa perlu mempertimbangkan informasi apapun mengenai B terlebih dahulu. P(A) juga berarti probabilitas ini diperoleh dari data sampel yang telah diketahui berkelas A, P(B) Probabilitas prior B, dan bertindak sebagai normalizing constant. Secara intuitif, teorema Bayes menggambarkan bahwa perubahan pada "A" dapat diamati apabila "B" terlebih dahulu diamati.

klasifikasi pada kategori yang memiliki fitur yang sangat besar. Berikut rumusan persamaan probabilitas NBC [11]:

$$f(W_{kj}|C_j) = \frac{f(W_{kj}|C_j)+1}{f(C_j)+|W|} \qquad (3)$$

Dengan $f(Wkj|Cj)$ adalah nilai kemunculan fitur pada Wkj kategori Cj, Wkj adalah nilai dari kemunculan fitur di satu kategori, Cj adalah kategori, $f(Cj)$ adalah jumlah keseluruhan fitur yang muncul pada kategori, $|W|$ adalah jumlah keseluruhan kata / fitur yang digunakan.

Sedangkan untuk menentukan klasifikasi pada data uji,sebagai berikut persamaan yang digunakan[11]:

$$\operatorname{argmax}_{c \in C} p \prod_k p(W_{kj} | C_j) \qquad (4)$$

Hasil pencarian probabilitas setiap kata dan kategori sudah didapat dijadikan acuan untuk mencari kategori dari tweet berikutnya dengan fitur yang sudah diketahui[13].

2.5 Akurasi

Dalam tahapan pengukuran performansi adalah tahap analisis dan evalusai pada sistem yang akan dirancang. Dalam penelitian tugas akhir ini, digunakan performansi yang diukur dengan menggunakan nilai akurasi, *precision*, dan *recall*. Untuk mempermudah menghitung performansi maka digunakan *confusion matrix*.

Tabel 1 *confusion matrix*

		Prediksi kelas	
Kelas Asli		class = yes	class = no
	class = yes	TP	FN
	class = no	FP	TN

.

Dengan True Positive (TP) kelas yang diprediksi yes, dan ternyata faktanya yes (hasil yang benar), (TN) True Negative (TN) kelas yang diprediksi no, dan ternyata faktanya no (tidak adanya hasil yang benar), False Positive (FP) kelas yang diprediksi yes, tetapi faktanya no (hasil yang tidak diharapkan), False Negative (FN) Kelas yang diprediksi no, tetapi faktanya yes (hasil yang meleset).

a. 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 persamaannya [11].

$$Akurasi = \frac{TP + FN}{(TP\ +\ FP\ +\ TN\ +\ FN)} \qquad (5)$$

b. Precision

Precision merupakan rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total prediksi yang diklasifikasikan ke dalam kelas tersebut. Berikut rumus dari *precision*:

$$Precision\ (\mathrm{P}) = \frac{TP}{(TP+FP)} \qquad (6)$$

c. Recall

Recall merupakan rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total fakta yang diklasifikasikan ke dalam kelas tersebut. Berikut rumus dari *recall*:

$$Recall\ (\mathrm{R}) = \frac{TP}{(TP+FN)} \qquad (7)$$

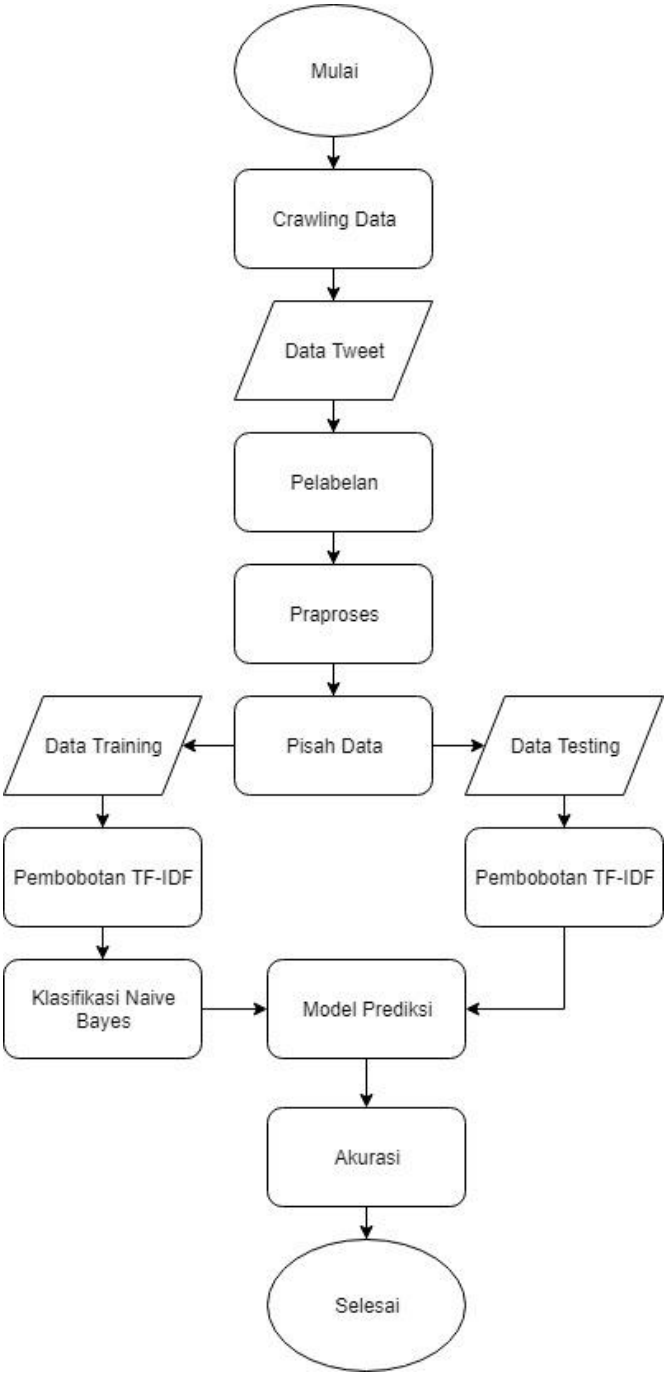
d. F-1 Score

Untuk menggabungkan rumus *precision* dan *recall* menjadi sebuah rumus tunggal yang disebut *F-Measure* atau *F-1 Score*, dapat dihitung dengan menggunakan rumus berikut:

$$F-1\ SCORE = 2\ \frac{\mathrm{Precision\ x\ Recall}}{\mathrm{Precision+Recall}} \qquad (8)$$

3 Sistem Klasifikasi Naïve Bayes

Dilakukan perancangan sistem yang akan dibangun untuk mengklasifikasi menggunakan metode Naïve Bayes dan pembobotan TF-IDF. Data yang digunakan adalah dari postingan tweets pengguna *Twitter* yang didapatkan melalui crawling data menggunakan API *Twitter*.



Gambar 1. Alur Sistem prediksi Trending Topik

1. *Crawling Data*

Crawling data di *Twitter* dapat menggunakan dua sistem pencarian, *by user* dan *by hastag*. Pencarian menggunakan *by keyword* yaitu pencarian menggunakan penggalan kata maupun hashtag dengan total tweet yang diunduh dalam sekali proses maksimum 200 tweet. Sedangkan pencarian dengan *by user* yaitu pencarian berdasarkan nama akun user *Twitter* dengan total tweet yang diunduh dalam sekali proses maksimum 3200 tweet. Ekstrasi fitur yang didapat dari *index Twitter* untuk data user berupa total tweet, total *follower*, total *following, total likes, website, source, bio profile*, id, akun, nama dan lokasi. Sedangkan ekstraksi fitur yang didapat dari index *Twitter* untuk data tweet berupa url, *mention*, *retweet, hashtag*, jumlah *likes* dan jumlah *retweet*[9]. Telah dijelaskan pada Tabel 3

2. Pelabelan

Lalu data tweet dilabelkan secara manual sesuai kategori berdasarkan bio akun atau artikel berita tersebut di website berita. Kategori yang ditentukan ada 12 yaitu: Senbud, Ekonomi, Hiburan, Kesehatan, Olahraga, Otomotif, Teknologi, Politik, Hukam, Pendidikan, Sosial, Umum. Pada Tabel 3 dijelaskan contoh pelabelan.

Tabel 2 Contoh Pelabelan Data

NO	Nama Akun	Tweet	Label
1	@InfoSehatku	"Sesuai dengan gaya hidup kita saat ini yang rentan dengan berbagai penyebab penyakit favorit saat ini"	Kesehatan
2	@bandungheritage	"Hari ini kami akan membahas detail acara Ngawawaas Bandung! ON AIR di Radio Raka 98.8 FM, pukul 17.00-18.00 WIB. Selamat menikmati 😊"	Senbud
3	@BareskrimPolri	"Poros Massa akan berada di sekitar Masjid Istiqlal, Tetap berhati-hati , oknum yg tdk bertanggung jwb akan membuat gaduh."	Hukam

3. *Pre-processing*

Pada tahap ini, *data training* dan *data testing* akan dilakukan *preprocessing data* untuk menghilangkan data yang tidak sempurna. Beberapa tahapan diantaranya ialah *case folding, tokenizing, filtering* dan *stemming*.

a. *Case folding*, yaitu mengubah seluruh huruf kapital menjadi huruf kecil.
b. *Stop Removal,* adalah adalah menghilangkan kata-kata non-topikal yang tidak dianggap penting.
c. *Stemming,* yaitu mengembalikan kata ke dalam bentuk dasar (kata dasar) dengan menghilangkan aditif yang ada.

4. Pembagian Data

Setelah melakukan pelabelan tahapan selanjutnya adalah pisah data. Data *tweet* akan dibagi menjadi data *training* dan data *test* dengan rasio berbeda-beda yaitu 50:50, 60:40, 70:30, 80:20, 90:10. Pembagian data ini dilakukan untuk mendapatkan hasil performansi yang baik.

5. Pembobotan TF-IDF

Pada proses ini data *tweet* yang telah dikumpulkan sebelumnya akan di proses untuk dilakukan pembobotan yang bertujuan untuk mendapatkan rating pada kata–kata yang didapatkan untuk dilakukan pengkasifikasian.

6. Klasifikasi NB

Pada pengujian NBC, data tweet yang digunakan adalah data testing yang telah dipisahkan pada proses seperti skenario diatas. Pengujian ini dilakukan untuk mengetahui hasil prediksi kategori oleh NBC terhadap kategori aktual tweet tersebut.

7. Akurasi

Proses ini merupakan tahapan terakhir yaitu menghitung akurasi, *precision, recall,* dan *f-measure* dari sistem yang sudah dibuat.

4. Evaluasi

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

4.1. Data Set & Pelabelan

Data yang digunakan adalah data tweet yang diunduh menggunakan aplikasi crawling yang dilakukan pada bulan 25 Juli sampai 28 Agustus. Data yang diunduh sebanyak 77.793 data. Data tweet yang telah didapatkan dilakukan pelabelan berdasarkan 12 kategori secara manual. Contoh data yang telah dilakukan pelabelan secara manual dapat dilihat pada Tabel 3.

Tabel 3 Jumlah data perlabel dan *keyword*

NO	Kategori	Kata Fitur	Jumlah Data
1	Ekonomi	bank, ekonomi, uang, rupiah, inflasi, kerja, hasil, korupsi	6094
2	Hiburan	netmediatama, music, nonton, tv, channel, premiere	7838
3	Hukam	polri, tinjau, tim, lapor, tinjau, yayasan, lindung, uu	6133
4	Pendidikan	mendikbud, itbofficial, unpad, ui, universitas, snmptn, sbmptn	6042
5	Politik	jokowi, prabowo, pilkada, dpr, politik, capres, cawapres, pilpres, kpu, ahok	6396
6	Olahraga	persib, arema, badminton, minions, piala, juara	6173
7	Kesehatan	infokesehatan, obat, sehat, jantung, olahraga, makan, dokter	6553
8	Senbud	tradisi, wisata, acara, festival, candi, tiket, foto	6159
9	Teknologi	microsoft, kamera, infokomputer, logitech, fifa	6272
10	Umum	remaja, teman, detik, percumanmaintwitter, jakarta, rebahan, potensi	6648
11	Sosial	papua, amal, hidup, sutupo, bmkg, erupsi	7129
12	Otomotif	gridoto, honda, mobil, iphone, xiomi, kendaraan, oppo	6356
		Total	77.793

4.2.1 Hasil Preprocessing

Pada tahap ini, semua tweet akan dilakukan *preprocessing* untuk mengubah data menjadi terstruktur. Proses *preprocessing* yang pertama adalah penghapusan *retweet* dari tweet, kemudian yang kedua diikuti dengan penghapusan URL. Ketiga *Case folding* mengubah huruf besar menjadi huruf kecil. Keempat *Tokenizing* merupakan proses penghapusan karakter seperti titik (.) dan koma (,) lalu, *tweet* diuraikan menjadi satuan kata. Kelima *Filtering* yaitu pemilihan kata-kata penting setelah proses *tokenizing*. Dan yang terakhir keenam adalah *stemming* yaitu proses mengubah kata yang berimbuhan menjadi kata dasar.

.

Tabel 4 Contoh Hasil Preprocessing

Sebelum Preprocessing	**"Hiii @InfoSehatku akan selalu meng update informasi seputar kesehatan"**
Setelah Preprocessing	infosehatku update informasi putar sehat

4.2.2 Korpus Kata Fitur

Korpus adalah kamus yang menjelaskan atau menunjukan bagaimana kata itu bekerja sama membentuk kalimat. Untuk membuat korpus kata fitur menggunakan pembobotan yaitu TF-IDF. Kata atau *term* yang sering muncul disetiap kategori diberi bobot. Contoh korpus kata fitur dapat dilihat pada Tabel 5.

Tabel 5 Korpus Kata Fitur

No	Korpus Kata Fitur	Label
1	bank, ekonomi, uang, rupiah, inflasi, kerja, hasil, korupsi	Ekonomi
2	netmediatama, music, nonton, tv, channel, premiere	Hiburan
3	polri, tinjau, tim, lapor, tinjau, yayasan, lindung, uu	Hukam
4	mendikbud, itbofficial, unpad, ui, universitas, snmptn, sbmptn	Pendidikan
5	jokowi, prabowo, pilkada, dpr, politik, capres, cawapres, pilpres, kpu, ahok	Politik
6	persib, arema, badminton, minions, piala, juara	Olahraga
7	infokesehatan, obat, sehat, jantung, olahraga, makan, dokter	Kesehatan
8	tradisi, wisata, acara, festival, candi, tiket, foto	Senbud
9	microsoft, kamera, infokomputer, logitech, fifa	Teknologi
10	remaja, teman, detik, percumanmaintwitter, jakarta, rebahan, potensi	Umum
11	remaja, teman, detik, percumanmaintwitter, jakarta, rebahan, potensi	Sosial
12	gridoto, honda, mobil, iphone, xiomi, kendaraan, oppo	Otomotif

4.2.3 Hasil Pegujian NBC

Setelah melalui tahap *preprocessing* dan pembobotan akan masuk ke tahap klasifikasi. Pengujian ini dilakukan untuk mengetahui hasil prediksi kategori oleh NBC terhadap kategori aktual tweet tersebut. Pada tahap pengklasifikasian ini data training yang dilakukan pengujian. Hasil persentase yang didapat dijelaskan pada Tabel 6.

Tabel 6 Hasil Pengujian

No	Skenario	Akurasi	Presisi	*Recall*	*F-measure*
1	50:50	56.97%	63%	51%	0.51
2	60:40	54.49%	64%	51%	0.51
3	70:30	54.82%	63%	51%	0.51
4	**80:20**	**57.08%**	**67%**	**52%**	**0.52**
5	90:10	55.15%	64%	51%	0.51

Dari percobaan **Tabel 6** mendapatkan hasil klasifikasi Naïve Bayes dengan pembobotan TF-IDF pada akurasi terbaik dengan scenario 90:10 adalah 55.08% dari 77.793 data tweet yang diproses dan mendapatkan hasil *f-measure* adalah 0.52.

4.3 Analis Hasil Pengujian

Analisis dari hasil pengujian TF-IDF dipengaruhi oleh pengambilan data secara acak. Dapat dilihat pada tabel 6 akurasi naik turun. Sehingga semakin banyak nilai perbandingan yang diberikan pada data training maka nilai akurasi klasifikasi menurun walau tidak signifikan. Pada pengujian ini didapat hasil terbaik adalah 57.08% pada scenario 80:20 dengan *presisi* 67% *recall 57%* dan *f-measure* 0.52 dari percobaan dataset dengan 5 buah skenario yang berbeda menggunakan jumlah data sebesar 77.793 data.

Kesalahan dalam pengujian deteksi trending topik ini adalah pada pemberian label secara manual yang tidak sesuai dengan tweetnya karna masih banyak data tweet yang dilabelkan tidak sesuai dengan tweet aktualnya atau bio pada akun. Dan tidaknya rapih data pada saat *preprocessing* sehingga masih ada kata-kata yang tidak jelas dan kata-kata tersebut didapatkan dengan jumlah besar pada perhitungan pembobotan TF-IDF. Dalam hasil akurasi tertinggi pada scenario 90:10, topik yang selalu dibicarakan dari pengambilan data bulan 25 Juli sampai 28 Agustus dapat dilihat pada Tabel 7.

Tabel 7 Trending Topik

No	Label	Akurasi
1	Senbud	8.65%
2	Ekonomi	6.87%
3	Hiburan	7.73%
4	Kesehatan	5.01%
5	Olahraga	7.17%
6	Otomotif	6.16%
7	Teknologi	6.71%
8	**Politik**	**26.88%**
9	Hukam	8.27%
10	Pendidikan	7.84%
11	Sosial	5.99%
12	Umum	7.07%

Dari **Tabel 7** telah diketahui persentase tertinggi pada kategori Politik 26.88% pada saat pengambilan data. Persentase dihitung dari jumlah tweet diprediksi tepat dibagi jumlah seluruh data.

5. Kesimpulan dan Saran

Penelitian ini bertujuan untuk mengatahui *trending topik* dan pengaruh TF-IDF terhadap proses klasifikasi Naïve Bayes. Pada penelitian analisis *trending topik***,** didapatkan hasil akurasi terbaik 57.08% menggunakan Naïve Bayes dan pembobotan TF-IDF pada scenario 80:20. Data tweet dan pelabelan dapat mempengaruhi peningkatan akurasi. Trending topik atau berita yang selalu dibicarakan oleh pengguna *Twitter* terdeteksi pada bulan 25 Juli sampai 28 Agustus adalah politik karna kategori tersebut memiliki persentase paling tinggi yaitu 26.88% diikuti dengan senbud 8.65% dan hukam 8.27% dari pembobotan TF-IDF dan pengklasifikasian Naïve Bayes dari 3 kali percobaan yang telah dilakukan pada setiap skenarionya.

Beberapa alasan lain yang mempengaruhi hasil akurasi yang didapatkan karna jumlah kategori yang digunakan terlalu banyak sehingga model prediksinya yang dihasilkan menjadi bias tidak akurat dan persisi sehingga sistem sulit untuk memprediksi kata yang bisa masuk disetiap labelnya karena banyak kata yang dimaksud disetiap labelnya.

Saran untuk pengembangan penilitian pendeteksi trending topik adalah mencoba untuk menggunakan lebih dari satu metode agar mendapatkan perbandingan yang lebih baik lagi dan pada saat melakukan pelabelan manual harus diliat secara teliti teks yang dimaksud. Pada saat tahapan *preprocessing* harus memperhatikan kalimat-kalimat yang dirasa tidak begitu penting, agar saat melakukan pembobotan mendapatkan nilai yang baik dari fitur term yang telah diproses. Dan disarankan juga menggunakan perangkat yang mempuni agar proses yang dilakukan tidak lama dan berjalan lancar karna data yang digunakan besar.

Daftar Pustaka

[1] 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.

[2] M. S. C. Sapul, T. H. Aung, and R. Jiamthapthaksin, "Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms," in Proceedings of the 2017 14th International Joint Conference on Computer Science and Software Engineering, JCSSE 2017, 2017.

[3] Samodra, J., Sumpeno, S., & Hariadi, M. (2009). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes. Seminar, 1–4

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

[5] J. Song, K. T. Kim, B. Lee, S. Kim, and H. Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis," KSII Trans. Internet Inf. Syst., vol. 11, no. 6, pp. 2996–3011, 20.

[6] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, dan D. Steinberg, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[7] Jiwanggi, Adriani - 2016 - Topic Summarization of Microblog Document in Bahasa Indonesia using the Phrase Reinforcement Algorithm-annotated.

[8] Xu, R. Grishman, A. Meyers and A. Ritter, "A Preliminary Study of Tweet Summarization using Information Extraction," in Proceedings of the Workshop on Language Analysis in Social Media pp. 20-29, Atlanta,Georgia, 2013.

[9] Eka Sembodo, J., Budi Setiawan, E., & Abdurahman Baizal, Z. (2016). Data Crawling Otomatis pada Twitter. Indosc 2016, (September 2016), 11–16.

[10] Agustina, P. A., Matulatan, T., Tech, M., & Si, M. B. S. (2012). Klasifikasi Trending Topic Twitter dengan Penerapan Metode Naive Bayes, 5.

[11] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," J. Inf. Sci., vol. 44, no. 1, pp. 48–59, 2018

[12] A. V. Member and G. Mohammadi, "Vinciarelli, A., and Mohammadi, G. (2014)," vol. 5, no. December, pp. 273–291, 2014.

[13] Yuan, Q., Cong, G., & Thalmann, N. M. (2012). Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification, 645–646.

[14] D.T. Larose, "Discovering Knowledge in Data" New Jersey: John Willey & Sons,2005.

[15] M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," 2017.

[16] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, 2004.