 ISSN: XXXX-XXXX

Analisis Penyakit Difteri Berbasis Twitter Menggunakan Algoritma Naïve Bayes

Ali Sholihin[1], Haviluddin[2], Novianti Puspitasari[3], Masna Wati[4], Islamiyah[5]

[1,2,3,4,5] *Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Mulawarman, Kalimantan Timur*
e-mail: ali.sholihin7@gmail.com, haviluddin@gmail.com, miechan.novianti@gmail.com, masnawati.ssi@gmail.com, islamiyah1601@yahoo.co.id

INFORMASI ARTIKEL

Histori Artikel
Diterima :
Direvisi :
Diterbitkan :

Kata Kunci:
Penyakit Difteri
Naïve Bayes
Klasifikasi
Opini

ABSTRAK

Antisipasi dan penanganan penyakit difteri dengan tepat sangat diperlukan oleh Pemerintah Indonesia. Oleh karena itu, informasi dari masyarakat terkait penyakit difteri sangat diperlukan oleh instansi yang berwenang. Hasil dari analisa informasi tersebut dapat menjadi salah satu rujukan dalam mengevaluasi antisipasi dan penanganan kepada masyarakat. Dalam penelitian ini, sebanyak 290 informasi terkait penyakit difteri dari masyarakat telah diambil dari data media sosial yaitu Twitter. Sedangkan, analisa data telah dilakukan menggunakan metode kecerdasan buatan berbasis semantic analysis yaitu Naïve Bayes (NB). Dalam percobaan ini, data yang dikenali telah diklasifikasikan ke dalam opini negatif dan positif. Berdasarkan hasil analisa data menunjukkan bahwa sebesar 94.5% bernilai negatif dan 5.5% bernilai positif. Hal ini menunjukkan bahwa masyarakat menganggap layanan Pemerintah terhadap penanganan penyakit difteri masih kurang percaya.

I. Pendahuluan

Saat ini, media sosial merupakan wadah yang popular dalam berkomunikasi dan bertukar informasi. Berbagai topik informasi seperti sosial, politik, ekonomi, kesehatan, pertahanan, keamanan dan lain-lain saling disebarkan (sharing) kepada sesama pengguna media social. Data per Januari 2019 menunjukkan pengguna media social sebanyak 7.7 milyar. Facebook merupakan pengguna terbanyak yaitu 2.3. Sedangkan, di Indonesia sebanyak 132 juta pengguna internet, 40 % didominasi oleh pengguna media sosial seperti Facebook, Linkedln, Pinterest, Instagram, dan Twitter. Dari beberapa media sosial yang ada, Twitter merupakan salah satu media sosial yang cukup banyak digunakan. Saat ini, pengguna Twitter di dunia sebanyak lebih dari 500 juta pengguna, sementara di Indonesia sebanyak 29 juta pengguna milyar (TetraPakIndex, 2017).

Berdasarkan banyaknya pengguna media sosial tersebut maka jumlah data yang tersimpan di media sosial juga semakin banyak. Sehingga para peneliti banyak melakukan penelitian dengan memanfaatkan data media sosial ini. Opinion mining (OM) atau sentiment analysis (SA) merupakan salah satu metode analisa data media sosial yang mengolah sebuah informasi yang terkandung dalam teks. OM/SA ini merupakan cabang ilmu dari text mining. Tujuan dari analisa suatu teks adalah untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang seperti kecenderungan berpandangan atau beropini negatif atau positif. Beberapa penelitian terkait OM/SA telah dilakukan oleh peneliti (Rozi, Pramono, & Dahlan, 2012) telah melakukan penelitian pada opini publik terhadap Perguruan Tinggi. Penelitian ini mengembangkan sub proses document subjectivity dengan teknik Part-of-Speech (POS) dan target detection menggunakan teknik Hidden Makov Model (HMM) kemudian dianalisa menggunakan metode Naïve Bayes (NB). Sebanyak 575 data teks telah digunakan. Hasil percobaan menunjukkan bahwa nilai precission dan recall untuk subproses document subjectivity adalah 0.99 dan 0.88, untuk subproses target detection adalah 0.92 dan 0.93, serta untuk subproses opinion orientation adalah 0.95 dan 0.94. Hal ini menunjukkan bahwa metode NB dapat digunakan dalam menganalisa opini publik Perguruan Tinggi. Pada penelitian yang dikerjakan (Adiyana & Hakim, 2015), melakukan analisis OM mengenai topik-topik terkait "KPK dan JOKOWI" pada pencarian twitter dengan menerapkan Algoritma clustering menghasilkan ukuran asosiasi kata dengan nilai korelasi tidak kurang dari 0.30, kata-kata yang berasosiasi dengan kata KPK adalah kata Polri dan Lapor. Sedangkan kata-kata yang

berasosiasi dengan Jokowi dimana nilai korelasi tidak kurang dari 0.30 adalah kata Widodo, Menghadiri, Izin, Pintu, Satu, Investor, Urus, Presiden, Nilai, Aktif, Bahaya, Manuver, Menang, Mulai, Relawan, dan Sejumlah. Dan pada penelitian yang dikerjakan (Nurhuda, Widya Sihwi, & Doewes, 2016) dengan judul Sentimen Analisis dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter, penelitian ini berhasil membangun sebuah model untuk melakukan klasifikasi tweet berdasarkan sentiment dan kategori dengan Naive Bayes Classifier. Hasil akurasi pengujian klasifikasi dengan fitur term frequency diperoleh sebesar 79,91% sedangkan fitur TF-IDF didapatkan akurasi sebesar 79,68%. Klasifikasi menggunakan tools RapidMiner dengan Naive Bayes dan fitur term frequency diperoleh sebesar 73,81% sedangkan dengan fitur TF-IDF diperoleh sebesar 71.11%. Klasifikasi dengan Support Vector Machine menghasilkan akurasi 83,14% untuk fitur term frequency dan 82,69% untuk fitur TF-IDF. Juga pada penelitian yang dilakukan (Nugroho, 2018) Analisis Sentiment Pada Media Sosial Twitter Menggunakan Naïve Bayes Classifier Dekat Ekstraksi Fitur N-Gram, dapat membangun sebuah model algoritma naiver bayes dengan tingkat akurasi mencapai 89,67%, dan pengaruh ekstraksi fitur n-gram yang diterapkan dapat meningkatkan nilai akurasi pada algoritma naïve bayes sekitar 2,33%, yaitu menjadi 92,00% dalam pengklasifikasian data tweet. Hal ini menunjukkan kemungkinan dilakukan analisis OM/SA pada data yang ada di twitter.

Paper ini bertujuan untuk menganalisa opini masyarakat kepada Pemerintah dalam memberikan pelayanan terhadap penyakit difteri. Sehingga, model antisipasi dan pelayanan dapat disebarluaskan kembali oleh Pemerintah terkait informasi kesehatan serta cara-cara penanganan penyakit difteri melalui Twitter. Paper ini terdiri dari, bagian 2 penjelasan terkait opinion mining/sentiment analisis. Bagian 3 hasil dan pembahasan percobaan. Kesimpulan dari penelitian disajikan pada bagian akhir paper.

II. Metodologi

A. Text Mining

Text Mining juga dikenal sebagai data mining text atau penemuan pengetahuan dari database tekstual. Sesuai dengan buku The Text Mining Handbook, text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Dalam memberikan solusi, text mining mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti Data Mining, Information Retrieval, Statistik dan Matematik, Machine Learning, Linguistic, Natural Language Processing (NLP) dan Visualization. Kegiatan riset untuk text mining antara lain ekstraksi dan penyimpanan teks, preprocessing akan konten teks, pengumpulan data statistik serta indexing dan analisis sentimen (Rozi et al., 2012).

Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian / pengelompokkan dan menganalisa unstructured data dalam jumlah besar, dalam hal ini data yang akan digunakan adalah data yang diambil dari Twitter. Dalam memberikan solusi, text mining mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti Data Mining, Information Retrieval, Statistik dan Matematik, Machine Learning, Linguistic, Natural Language Processing dan Visualization. Kegiatan riset untuk text mining antara lain ekstraksi dan penyimpanan teks, preprocessing akan konten teks, pengumpulan data statistik serta indexing dan analisis sentiment (Triawati, Bijaksana, Indrawati, & Saputro, 2009). Adapun, proses dalam NLP terhadap analisa data antara lain (Nurhuda et al., 2016): (a). Case Folding: Proses untuk mengubah seluruh data yang ada menjadi huruf kecil; (b). Tokenization: Proses pemecahan kalimat menjadi term. Term dipecah berdasarkan spasi dan tanda baca, pada tokenisasi juga menghilangkan tanda baca tersebut. Contoh: input: ini ayah budi. Maka menjadi: ini|ayah|budi; (c). Stop Word Removal: Tahap penghapusan kata yang kurang bermakna untuk penilaian. Contoh: dan, tapi, namun dan sebagainya; dan (d). Stemming*: P*roses transformasi kata menjadi bentuk awalnya (root), dengan menghilangkan imbuhan dari kata tersebut (Mihuandayani, 2018).

Dalam penelitian ini, proses analisa data penyakit difteri dilakukan dengan menormalisasikan text dengan preprocessing, sehingga text dapat dibaca dan diolah menggunakan algoritma.

B. Natural Language Processing (NLP)

Natural Language Processing (NLP) merupakan salah satu cabang kecerdasan buatan (Artificial Intelligent). Bagi komputer untuk memahami bahasa natural manusia memerlukan sebuah proses yang disebut NLP. Dalam NLP terdapat 3 aspek utama pada teori pemahaman NLP antara lain, Syntax, Semantics, dan Preagmatics (Poole & Mackworth, 2010). Terdapat 5 area utama dalam NPL, antara lain (Pustejovsky & Stubbs, 2012): (a). Question Answering system (QAS). Kemampuan komputer untuk menjawab pertanyaan yang diberikan oleh user. Daripada memasukkan keyword kedalam browser pencarian, dengan QAS, user bisa langsung bertanya dalam bahasa natural yang digunakannya, baik itu Inggris, Mandarin, ataupun Indonesia; (b). Summarization. Pembuatan ringkasan dari sekumpulan konten dokumen atau email. Dengan menmggunakan aplikasi ini, user bisa dibantu untuk mengkonversikan sebuah dokumen teks yang besar kedalam bentuk slide presentasi; (c). Machine Translation. Merupakan aplikasi yang dapat memahami bahasa manusia dan menerjemahkannya kedalam bahasa lain. Contohnya adalah google translate; (d). Speech

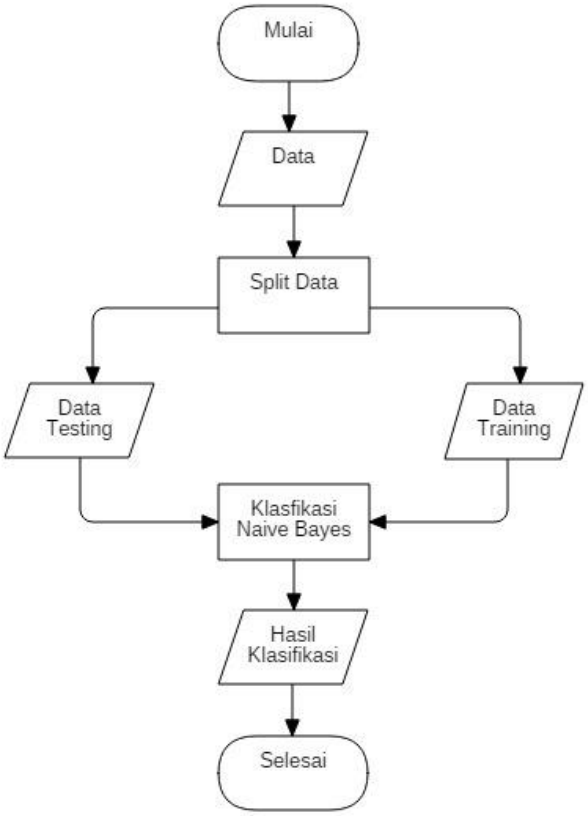
Recognition. Merupakan salah satu cabang NLP tersulit yang dimana aplikasi dapat mengenali apa yang diucapkan oleh user. Contoh aplikasi ini adalah Amazon Alexa dan Google home; dan (e). Document Classification. Merupakan penelitian yang paling banyak dilakukan saat ini. Yang dapat dilakukan aplikasi ini adalah mengelompokkan beberapa dokumen kedalam kelas-kelas tertentu.

Dalam penelitian ini, NLP dengan area Document Classifiaction telah digunakan untuk menganalisa data opini masyarakat terhadap penyakit difteri.

C. Algoritma Naïve Bayes (NB)

Naive Bayes (NB) merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan bahwa NB merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Saleh, 2015).

Algoritma NB didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Keuntungan penggunaan algoritma NB adalah hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Algoritma NB ini, sering digunakan dalam kebanyakan situasi dunia nyata yang kompleks (unsupervised) (Saleh, 2015). Persamaan dari Metode NB bisa dilihat pada Gambar 1.



Gambar. 1.Flowchart Naïve Bayes

Keterangan: Data: Tahap pertama membaca seluruh data yang ada; Split Data: Tahap kedua membagi data menjadi dua bagian; Data training: Data yang digunakan algoritma NB untuk membentuk sebuah model Classifier; dan Data Testing: Merupakan data yang akan diolah oleh Algoritma NB untuk di klasifikasikan.

Setelah dilakukan text preprocessing data akan dibagi menjadi dua bagian yaitu data testing dan data training, data testing akan menjadi referensi bagi algoritma NB untuk mengklasifikasikan data dan data testing untuk menguji hasil pembelajaran dari algoritma NB, setelah splitting data selanjutnya dilakukan penghitungan algoritma NB, kemudian didapatkan kelompok data hasil klasifikasi.

Adapun, persamaan dari Teorema Bayes dapat dilihat pada persamaan (1).

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

Dimana, X adalah data class yang sebelumnya diketahui; H adalah hipotesis data merupakan suatu class spesifik; P (H|X) adalah probabilitas hipotesis H berdasarkan kondisi X (Posteriori Probabilitas); P (H) adalah Probabilitas hipotesis H (prior probabilitas); P (X|H) adalah probabilitas X berdasarkan kondisi pada hipotesis H; P (X) adalah probabilitas X.

D. Pengukuran Akurasi Analisa

1) K-Fold Cross Validation

K-Fold Cross Validation adalah metode validasi yang membagi data ke dalam subset, kemudian dilakukan perulangan (iterasi) pengujian sebanyak K. Pada setiap pengulangan, digunakan satu subset sebagai data uji dan subset lainnya sebagai data pembelajaran. Keuntungan dari metode ini adalah setiap data, minimal akan menjadi data uji sebanyak satu kali dan akan menjadi data learning juga minimal sebanyak satu kali (A. Widjaya, Hiryanto, & Handhayani, 2017). Dalam satu set percobaan akan dilakukan k buah percobaan klasifikasi dokumen dengan tiap percobaan menggunakan satu bagian sebagai data testing, (k-1)/2 bagian sebagai labeled documents, dan (k-1)/2 bagian lainnya sebagai unlabeled documents yang akan ditukar setiap percobaan sebanyak k kali. Kumpulan dokumen yang dimiliki terlebih dahulu diacak urutannya sebelum dimasukkan ke dalam sebuah fold. Hal ini dilakukan untuk menghindari pengelompokkan dokumen-dokumen yang berasal dari satu kategori tertentu pada sebuah fold (Lidya, Sitompul, & Efendi, 2015).

2) Confusion Matrix (CM)

Confusion Matrix (CM) adalah alat (tools) visualisasi yang biasa digunakan pada supervised learning (Swastina, 2013). Confusion matrix merupakan sebuah table yang terdiri dari banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi. Tabel ini diperlukan untuk mengukur kinerja suatu model klasifikasi [12]. Tiap kolom pada matirks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian dikelas yang sebenarnya. Contoh data Confusion Matrix (CM) ditampilkan pada Tabel 1.

Tabel 1. Confusion Matrix (CM)

		Prediksi	
		Negative	***Positive***
Aktual	*Negative*	*True Positive*	*True Negative*
	Positive	*True Positive*	*True Negative*

3) Pembobotan Term Frequent-Inverse Document Frequent(TF-IDF)

TF-IDF merupakan suatu algoritma yang biasa digunakan untuk menganalisa hubungan dari sekumpulan dokumen (data text), TF-IDF melakukan pembobotan terhadap dokumen dengan cara menghitung kemunculan dari setiap kata (term) dalam dokumen. Pembobotan dapat diperoleh berdasarkan jumlah kemuculan suati term dalam dokumen Term Frequency (TF) dan jumlah kemunculan term dalam koleksi dokumen Inverse Document Frequency (IDF). Bobot suatu istilah semakin besar jika istilah tersebut semakin sering muncul dalam suatu dokumen dan semakin kecil jika istilah tersebut muncul dalam banyak dokumen (Grossman & Frieder, 2004). Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus (Maarif, 2015). Perhitungan IDF dari sebuah term (kata) dapat dilihat pada persamaan (2).

$$IDF = \log(\frac{d}{dfi}) \quad (2)$$

Dimana, D adalah jumlah dokumen yang berisi term (t); dan dfi adalah jumlah kemunculan (frekuensi) term terhadap D. Untuk menghitung nilai bobot (W) masing-masing dokumen terhadap kata kunci (query), menggunakan persamaan (3).

$$W_{d,t} = TF_{d,t} * IDF_t \quad (3)$$

Dimana, d adalah dokumen ke-d; t adalah term ke-t dari kata kunci; tf adalah *term frequency*/frekuensi kata; W adalah bobot dokumen ke-d terhadap term ke-t. (Andre Widjaya et al., 2017).

4) Consine Similarity (CS)

Consine Similiarity (CS) merupakan sebuah metode yang digunakan untuk menghitung tingkat kesamaan antara dua objek. Secara umum perhitungan metode ini didasarkan pada Vector Space Similiarity antara dua objek (misalkan Dokumen 1 dan Dokumen 2) yang dinyatakan dalam dua buah vector dengan menggunakan keywords (kata kunci) dari sebuah dokumen sebagai ukuran (Nurdiana, Jumadi, & Nursantika, 2018). Metode pengukuran kesesuaian ini memiliki beberapa keuntungan, yaitu adanya normalisasi terhadap panjang dokumen. Hal ini memperkecil pengaruh panjang dokumen. Jarak euclidean(panjang) kedua vektor digunakan sebagaifaktor normalisasi. Hal ini diperlukan karena dokumen yang panjang cenderung mendapatkan nilai yang besar dibandingkan dengan dokumen yang lebih pendek (Pradnyana, 2012). Adapun, perhitungan CS menggunakan persamaan (4).

$$CosSim(d_i, q_i) = \frac{q_i . d_i}{|q_i||d_i} = \frac{\sum_{j=1}^{t}(q_{ij} . d_{ij})}{\sqrt{\sum_{j=1}^{t}(q_{ij})^2} \cdot \sum_{j=1}^{t}(d_{ij})^2} \quad (4)$$

Dimana, qij adalah bobot istilah j pada dokumen i= tfij.idfj; dan dij adalah bobot istilah j pada dokumen i= tfij.idfj

E. Twitter

Twitter merupakan media sosial berjenis microblog yang digunakan oleh banyak orang di seluruh dunia. Media sosial ini didirikan oleh Jack Dorsey pada Maret 2006 di California, AS. Adapun, sistem kerja Twitter; user menyampaikan pesan disebut tweet dengan maksimal karakter berjumlah 140 karakter. Dengan menuliskan tweet tersebut, seorang user dapat mengetahui informasi dari sang penulis tweet. User lain dapat mengetahui isi pesan dari user yang mengirim pesan (tweet) tersebut. User dapat mengundang user lain untuk membaca dan berbalas komentar pada tweet yang dikirimnya (status posting). User dapat pula menyebarkan/membagikan tweet ke user lain yang mengikutinya (following). Pada beberapa tweet, terdapat link untuk mengantarkan user ke alamat sumber dari berita yang ada tersebut. Sehingga, dalam sistem Twitter, user dapat mengetahui perkembangan topik – topik terkini yang sedang banyak dibicarakan melalui tweet – tweet yang beredar disebut trending topic (Adiyana & Hakim, 2015). Twitter termasuk dalam jenis microblog, yang dimana user dibatasi dalam membuat kata dengan tujuan percepatan pertukaran sebuah opini (Sarlan, Nadam, & Basri, 2015).

F. Sampel Data Penelitian

Difteri adalah salah satu jenis penyakit menular yang juga dikenal sebagai penyakit infeksi, sebuah penyakit yang disebabkan oleh sebuah agen biologi (seperti virus, bakteria atau parasit), bukan disebabkan factor fisik (seperti luka bakar dan trauma benturan) atau kimia (seperti keracunan) yang bisa ditularkan atau menular kepada orang lain melalui media tertentu seperti udara, tempat makan dan minuman yang kurang bersih pencuciannya, jarum suntik dan tranfusi darah. Difteri disebabkan oleh kuman Corynebacterium Diphtheriae, suatu bakteri gram positif yang berbentuk polimorf, tidak bergerak dan tidak berbentuk spora. Gejala utama dari penyakit difteri yaitu adanya bentukan pseudomembran yang merupakan hasil kerja dari kuman ini. Pseudomembran sendiri merupakan lapisan tipis berwarna putih keabu-abuan yang timbul terutama di daerah mukosa hidung, mulut sampai tenggorokan (Hartoyo, 2018).

Dalam penelitian ini, sebanyak 177 data dari tweet pada twitter periode 01/09/2017 – 31/12/2017 berlokasi di Indonesia dengan keyword “difteri” telah digunakan untuk dianalisa. Adapun, dataset Twitter dapat dilihat pada Tabel 2.

Tabel 2. Data Twitter

No	Username	Time	Tweet
1	@blogdokter	30-Dec-17	Mengapa orang yg tidak pernah imunisasi difteri tapi bisa tidak kena difteri? 1. Dia sama sekali belum pernah terpapar bakteri difteri. 2. Dia pernah terpapar bakteri difteri tp gejalanya ringan shg tdk dirasakan. 3. Dia mndptkan perlindungan dari lingkungannya yg telah kebal. https://twitter.com/JaelaniNgopi/status/947248169930199041Ã,Â Ã¢â,¬Â¦
2	@kompascom	30-Dec-17	Via @KompasTV : Ahli Imunologi: Vaksin Difteri Diperlukan Orang Dewasa https://www.kompas.tv/content/article/17988/video/berita-kompas-tv/ahli-imunologi-vaksin-difteri-diperlukan-orang-dewasa?utm_source=Kompascom_social&utm_medium=web&utm_campaign=partnerÃ,Â Ã¢â,¬¦
3	@detikcom	30-Dec-17	Ingin Rayakan Tahun Baru dengan Terompet? Begini Caranya Antisipasi Difteri http://detik.id/6ZtQNMÃ,Â via @detikHealthpic.twitter.com/goNvbOPkbl
…	…	…	…
177	@CNNIndonesia	11-Dec-17	Berita terkini tentang penyebaran wabah difteri di Banten. Berikut selengkapnya. http://cnn.id/261822Ã,Â

III. Hasil dan Pembahasan

Metode opinion mining yang telah dilakukan pada penelitian melibatkan proses dan tahapan sebagai berikut :

A. Preprocessing

Tahap Preprocessing adalah tahapan dimana aplikasi melakukan seleksi data yang akan diproses pada setiap dokumen. Dalam percobaan ini, proses preprocessing terdiri meliputi Case Folding, Tokenization, Stopword Removal, Stemming telah diterapkan. Tahap pertama, Case Folding adalah merubah seluruh data menjadi huruf kecil dan karakter selain huruf dihilangkan. Hal ini bertujuan untuk penyeragaman dan lebih mudah dikenali komputer, seperti kata “Dift3r1” menjadi “diftr”. Tahap kedua, Tokenization adalah merubah data menjadi token-token yang berurutan. Dalam percobaan ini, pemecahan kata menjadi token-token yang berurutan dilakukan berdasarkan spasi pada kata. Jadi setiap spasi yang ada pada kata akan menjadi media untuk pemisahan (pemecah) antara satu kata dengan kata lainnya, seperti kata "layanan” menjadi “layan”. Tahap ketiga, Stopword Removal adalah proses penghapusan kata yang dianggap tidak memiliki makna. Dalam percobaan ini, sejumlah kata penghubung seperti “yang”,”di”,”itu” atau yang tidak mempengaruhi konten dokumen secara keseluruhan telah dihapus. Hal ini dilakukan untuk meningkatkan performa klasifikasi, agar dapat berjalan secara efektif, dan Tahap keempat, stemming merupakan proses pengembalian suatu kata ke akarnya. Dalam penelitian ini, algoritma Nazief dan Andriani telah digunakan. Setiap kata pada data akan dicocokan dengan basisdata kamus. Jika ditemukan maka akan dianggap sebagai kata dasar (root words), jika tidak ditemukan maka akan dilakukan penghapusan suffix lalu kemudian di periksa kembali ke basisdata, jika ditemukan maka proses dihentikan dan kata dianggap kata dasar. Jika tidak dilakukan maka akan dilakukan penghapusan prefix lalu kemudian kata akan dicocokan kembali dengan basisdata kamus, jika ditemukan maka data akan dianggap sebagai kata dasar. Adapun, hasil preprocessing data dapat dilihat pada Tabel 3.

Tabel 3. Hasil Preprocessing

No	*Username*	*Time*	*Tweet*
1	@blogdokter	30-Dec-17	['orang', 'yang', 'pernah', 'imunisasi', 'difteri', 'bisa', 'kena', 'difteri', 'sama', 'sekali', 'pernah', 'papar', 'bakteri', 'difteri', 'pernah', 'papar', 'bakteri', 'difteri', 'tp', 'gejala', 'ringan', 'shg', 'tidak', 'rasa', 'mndptkan', 'lindung', 'lingkung', 'yang', 'kebal']
2	@kompascom	30-Dec-17	['via', 'ahli', 'imunologi', 'vaksin', 'difteri', 'perlu', 'orang', 'dewasa']
3	@detikcom	30-Dec-17	['raya', 'tahun', 'baru', 'terompet', 'begini', 'cara', 'antisipasi', 'difteri', 'via']
….	…	…	…
290	@bidikindonesi	05-Sep-17	['masyarakat', 'aceh', 'tinggal', 'serang', 'virus', 'difteri', 'bidik', 'indonesia']

B. Term Frequent – Inverse Document Frequent (TF-IDF)

Pada tahap ini dilakukan perhitungan bobot dari setiap kata yang ada pada dokumen. Dimulai dengan perhitungan jumlah kata dari setiap dokumen menggunakan skema setiap dokumen dan seluruh dokumen lalu kemudian didapatkan bobot hasil dari setiap. Untuk menghitung nilai term frequent berdasarkan jumlah kemunculan term (kata) dalam sebuah dokumen. Untuk menghitung inverse document frequent (IDF) menggunakan persamaan (2). Sedangkan, menghitung bobot TF-IDF menggunakan persamaan (3). Hasil perhitungan TF-IDF dapat dilihat pada Tabel 4.

Tabel 4. Hasil TF-IDF

TF-IDF	a	abad	abai	abis	about	acara	aceh	Actdmii	adha
DOKUMEN 0	0	0	0	0	0	0	0	0	0
DOKUMEN 1	0	0	0	0	0	0	0	0	0
DOKUMEN 2	0	0	0	5.66296	0	0	0	0	0
DOKUMEN 3	0	0	0	0	0	0	0	0	0
….	…	…	…	…	…	…	…	…	…
DOKUMEN 287	0	0	0	0	0	0	0	0	0

C. Consine Similiarity (CS)

Selanjutnya, perhitungan consine similiarity (CS) untuk melihat tingkat kemiripan antara masing-masing dokumen. Pada tahap ini menggunakan data hasil perhitungan TF-IDF untuk menghitung tingkat kemiripan dari setiap dokumen dan menggunakan persamaan (4). Adapun, hasil consine similiarity dapat dilihat pada Tabel 5.

Tabel 5. Hasil Consine Similiarity (CS)

Consine Similiarity	Dokumen1	Dokumen2	Dokumen3	Dokumen4	Dokumen5
Dokumen1	1	1.62E-07	1.32E-07	0.024969	4.48E-07
Dokumen2	1.62E-07	1	0.01051	3.06E-07	2.97E-07
Dokumen3	1.32E-07	0.01051	1	2.49E-07	2.41E-07
Dokumen4	0.024969	3.06E-07	2.49E-07	1	8.48E-07
Dokumen5	4.48E-07	2.97E-07	2.41E-07	8.48E-07	1

D. Klasifikasi Naïve Bayes (NB)

Pada tahap ini data akan diolah dan diklasifikasikan menggunakan algoritma Naïve Bayes (NB). Data yang digunakan berasal dari proses Prepocessing Dan untuk menguji efektifitas dari algoritma, pembagian data menggunakan metode K-Fold Validation. Program akan di-iterasikan sebanyak 10 kali dan setiap data akan di bagi menjadi 29 data testing dan 261 data training. Hasil klasifikasi setiap iterasi dapat dilihat pada Tabel 6.

Tabel 6. Hasil Klasifikasi Per-iterasi

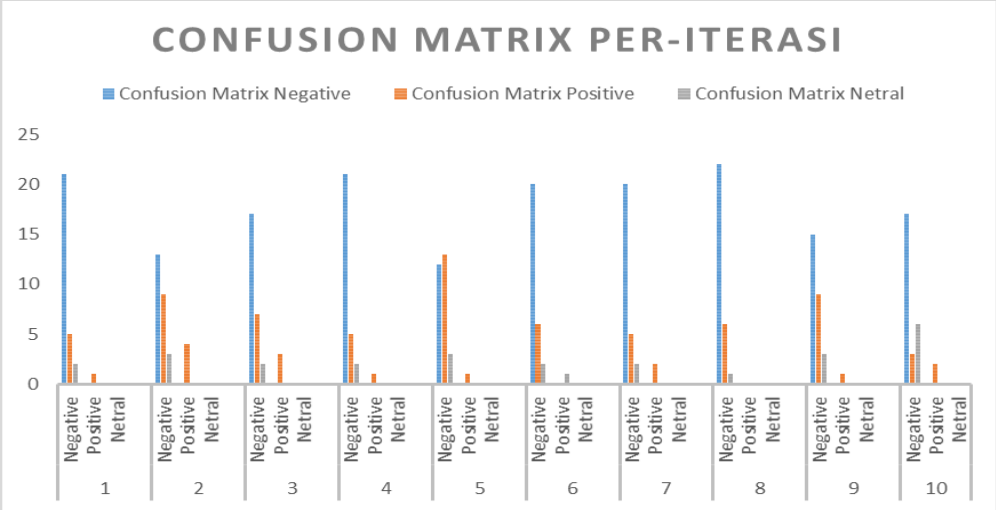
Data	Iterasi									
	1	2	3	4	5	6	7	8	9	10
1	*Neg*	*Neg*	*Pos*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*
2	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Pos*	*Neg*
3	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*
…	…	…	…	…	…	…	…	…	…	…
29	*Neg*	*Neg*	*Neg*	*Pos*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*	*Neg*

[a] Keterangan: Neg = Negative; Pos = Positive

Tabel 6 menunjukkan hasil klasifikasi dari 10 iterasi dan setiap iterasi terdapat 29 data terdapat 16 data positif dan 274 data negatif. Hal ini berarti, opini masyarakat terhadap antisipasi dan pelayanan penyakit difteri terhadap Pemerintah masih kurang.

E. Confusion Matrix (CM)

Untuk pengujian algoritma dilakukan perhitungan confusion matrix (CM) dari tiap-tiap iterasi, selanjutnya berdasarkan CM dilakukan perhitungan akurasi. Hasil perhitungan CM dapat dilihat pada Gambar 2.



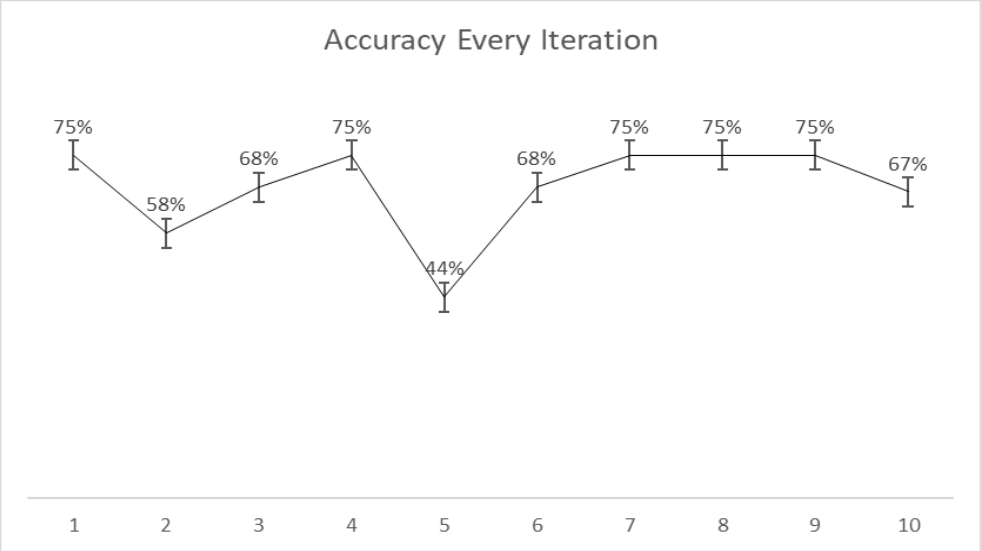
Gambar. 2.Chart Confuson Matrix

Sedangkan, perhitungan akurasi setiap iterasi dengan CM dapat dilihat pada Tabel 8 dan di Visualisaikan pada Gambar 2.

Tabel 7. Akurasi setiap iterasi

Iteration	Accuracy	Iteration	Accuracy
1	75%	6	68%
2	58%	7	75%
3	68%	8	75%
4	75%	9	75%
5	44%	10	67%

Diagram hasil akurasi setiap iterasi ditampilkan pada Gambar 3.



Gambar. 3.Plot akurasi setiap iterasi

Tabel 7 memperlihatkan bahwa nilai akurasi dari tiap-tiap iterasi yang yang dihasilkan memiliki akurasi tertinggi sebesar 75% dan terendah sebesar 44%. Sehingga, rata-rata akurasi dari seluruh klasifikasi adalah sebesar 66%.

IV. Kesimpulan

Penerapan opinion mining (OM)/sentiment analisis (SA) dalam menganalisa opini masyarakat terhadap pelayanan Pemerintah dalam penanganan penyakit difteri berbasis Twitter dengan menggunakan algoritma Naïve Bayes (NB) telah diimplementasikan. Berdasarkan hasil percobaan menghasilkan nilai akurasi CM setiap iterasi tertinggi sebesar 75% dan terendah sebesar 44%. Sehingga, rata-rata akurasi dari seluruh klasifikasi adalah sebesar 66%. Dengan kata lain, 94.5% bernilai negatif, hal ini berarti masyarakat masih belum percaya dengan pelayanan pemerintah mengenai difteri. Namun demikian, penelitian ini hanya memanfaatkan sumber data dari Twitter saja. Kedepan, pemanfaatan dari berbagai media sosial sebagai sumber data, disarankan untuk dilakukan sehingga hasil analisa menjadi lebih akurat dan transparan. Perbaikan algoritma NB dengan menggunakan algoritma optimasi akan menjadi penelitian selanjutnya.

Daftar Pustaka

Adiyana, I., & Hakim, R. B. (2015). *Implementasi Text Mining pada Mesin Pencarian Twitter untuk Menganalisis Topik–Topik Terkait "KPK dan JOKOWI."*

Grossman, D. A., & Frieder, O. (2004). Information Retrieval: Algorithms and Heuristics. In *Information Retrieval Algorithms and Heuristics*. https://doi.org/citeulike-article-id:1832374

Hartoyo, E. (2018). Difteri Pada Anak. *Sari Pediatri*. https://doi.org/10.14238/sp19.5.2018.300-6

Lidya, S. K., Sitompul, O. S., & Efendi, S. (2015). Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM). *Program Studi Teknik Informatika Universitas Atma Jaya Yogyakarta*.

Maarif, A. A. (2015). Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah. *Dokumen Karya Ilmiah | Tugas Akhir | Program Studi Teknik Informatika - S1 | Fakultas Ilmu Komputer | Universitas Dian Nuswantoro Semarang*.

Mihuandayani. (2018). Opinion Mining Pada Komentar Twitter E-KTP Menggunakan Naive Bayes Classifier. *Seminar Nasional Teknologi Informasi Dan Multimedia 2018*, 6.

Nugroho, A. (2018). Analisis Sentimen Pada Media Sosial Twitter Menggunakan Naive Bayes Classifier Dengan Ekstrasi Fitur N-Gram. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *2*(2), 200–209.

Nurdiana, O., Jumadi, J., & Nursantika, D. (2018). Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia. *Jurnal*

Online Informatika. https://doi.org/10.15575/join.v1i1.12

Nurhuda, F., Widya Sihwi, S., & Doewes, A. (2016). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Teknologi & Informasi ITSmart*. https://doi.org/10.20961/its.v2i2.630

Poole, D. L., & Mackworth, A. K. (2010). Artificial intelligence: Foundations of computational agents. In *Artificial Intelligence: Foundations of Computational Agents*. https://doi.org/10.1017/CBO9780511794797

Pradnyana, G. A. (2012). Perancangan dan Implementasi Automated Document Integration dengan Menggunakan Algoritma Complete Linkage Agglomerative Hierarchical Clustering. *Jurnal Ilmu Komputer*, *5*(2).

Pustejovsky, J., & Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Rozi, I. F., Pramono, S. H., & Dahlan, E. A. (2012). Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. *Jurnal EECCIS (Electrics, Electronics, Communications, Controls, Informatics, Systems)*.

Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*.

Sarlan, A., Nadam, C., & Basri, S. (2015). Twitter sentiment analysis. *Conference Proceedings - 6th International Conference on Information Technology and Multimedia at UNITEN: Cultivating Creativity and Enabling Technology Through the Internet of Things, ICIMU 2014*. https://doi.org/10.1109/ICIMU.2014.7066632

Swastina, L. (2013). Penerapan Algoritma C4.5 untuk Penentuan Jurusan Mahasiswa. *Jurnal Gema Aktualita*.

TetraPakIndex. (2017). The Connected Consumer. Retrieved from https://www.tetrapak.com

Triawati, C., Bijaksana, M. A., Indrawati, N., & Saputro, W. A. (2009). No TitlePemodelan Berbasis Konsep Untuk Kategorisasi Artikel Berita. *Semin. Nas. Apl. Teknol. Inf*, 48–53.

Widjaya, A., Hiryanto, L., & Handhayani, T. (2017). Prediksi Masa Studi Mahasiswa dengan Voting Feature Interval 5 pada Aplikasi Konsultasi Akademik Online. *Computatio: Journal of Computer Science and Information Systems*, *1*(1), 25–33.

Widjaya, Andre, Hiryanto, L., Handhayani, T., Studi, P., Informatika, T., Teknologi, F., & Universitas, I. (2017). *Prediksi Masa Studi Mahasiswa Dengan Voting Feature Interval 5 Pada Aplikasi Konsultasi Akademik Online*. *1*, 25–33. https://doi.org/10.1161/CIRCULATIONAHA.117.027355