ALGORITMA NAÏVE BAYES DALAM ANALISIS SENTIMEN UNTUK KLASIFIKASI PADA LAYANAN INTERNET PT.XYZ

Aria Mustofa Hidayat[1], Mohammad Syafrullah[2]

[1, 2)] Program Studi Magister Ilmu Komputer, Program Pascasarjana, Universitas Budi Luhur
Jl. Raya Ciledug, Petukangan Utara, Pesanggrahan, Jakarta Selatan 12260
Telp. (021) 5853753, Fax. (021) 5869225
[1]ariamustofa@gmail.com, [2]mohammad.syafrullah@budiluhur.ac.id

ABSTRAK

Kebiasaan masyarakat untuk mem-posting tweet yang memberikan informasi atau feedback terhadap suatu produk dapat dimanfaatkan sebagai landasan untuk mengetahui sentimen terhadap produk penyedia layanan internet yang ada di Indonesia. Kesulitan untuk mengolah data dan pengklasifikasian sentimen positif dan negatif yang tersedia pada media sosial Twitter serta nilai akurasinya menjadi suatu permasalahan. Pada penelitian ini akan membahas mengenai cara pengolahan data dari Twitter dengan tujuan untuk melakukan pengklasifikasian terhadap sentimen positif dan sentimen negatif dari posting tweet serta mencari akurasi dari metode yang digunakan yaitu Naïve Bayes Classifier. Data tweets yang digunakan yaitu sebanyak 500 data dimana masing-masing data positif sebanyak 250 data dan data negatif sebanyak 250 data. Hasil dari eksperimen yang dilakukan dalam penelitian ini menunjukkan bahwa nilai akurasi klasifikasi dari metode Naïve Bayes Classifier yaitu sebesar 91.00%.

Kata Kunci : Analisis Sentimen, Klasifikasi, Akurasi, *Preprocessing, Naïve Bayes Classifier.*

I. PENDAHULUAN

Twitter merupakan salah satu situs jejaring sosial yang sedang digandrungi akhir-akhir ini. Setelah diluncurkan pada Juli 2006, jumlah pengguna Twitter meningkat sangat pesat. Pada September 2010, diperkirakan jumlah pengguna Twitter yang terdaftar sekitar 160 juta pengguna (Chiang, 2010). Kebiasaan masyarakat untuk mem-posting tweet yang memberikan informasi atau feedback terhadap suatu produk juga dapat dimanfaatkan sebagai landasan untuk mengetahui sentimen terhadap produk penyedia layanan internet yang ada di Indonesia.

Salah satu kegiatan penting dalam distilasi pengetahuan adalah klasifikasi atau kategorisasi teks dengan pendekatan text preprocessing. Dari kelompok pendekatan berbasis numeris, pendekatan berbasis probabilistic Naïve Bayes Classifier memiliki kelebihan antara lain, sederhana, cepat, dan berakurasi tinggi. Metode Naïve Bayes Classifier untuk klasifikasi teks menggunakan atribut kata yang muncul dalam suatu dokumen sebagai dasar kalsifikasinya. Penelitian Rish (2001) menunjukkan bahwa meskipun asumsi independensi antar kata dalam dokumen tidak sepenuhnya dapat dipenuhi, tetapi kinerja Naïve Bayes Classifier dalam klasifikasi relative sangat bagus.

Pada penelitian ini, terdapat beberapa identifikasi masalah antara lain, belum ada pengelolaan sistem informasi yang baik sebagai tolak ukur informasi untuk melakukan evaluasi pada layanan internet first media, belum ada pengklasifikasian sentimen positif dan sentimen negatif pada layanan internet first media dan belum diketahui nilai akurasi klasifikasi algoritma Naïve Bayes Classifier terhadap data uji pada pembentukan sentiment analysis.

Adapun batasan masalah yang akan diteliti pada penelitian ini antara lain, data yang digunakan diambil dari Twitter berdasarkan respon tweets masyarakat dari penyedia layanan internet first media yang dibatasi 500 respon tweets, metode dan algoritma yang digunakan untuk pada penelitian ini yaitu metode dan algoritma Naïve Bayes Classifier, postingan tweets yang digunakan adalah tweets dalam bahasa Indonesia, data sentimen yang diklasifikasikan kedalam dua kelas yaitu opini positif dan opini negatif, proses Stopword dan Stemming hanya berlaku pada kata-kata berbahasa Indonesia.

Berdasarkan uraian diatas, maka pada penelitian ini akan membahas mengenai cara pengolahan data dari Twitter dengan tujuan untuk melakukan pengklasifikasian terhadap sentimen positif dan sentimen negatif pada layanan internet first media dari tweets posting dan mencari nilai akurasi dari metode yang diusulkan yaitu Naïve Bayes Classifier serta mengimplementasikan hasil nilai akurasi dari metode tersebut.

II. LANDASAN TEORI

A. Tinjauan Studi

Banyak penelitian sebelumnya yang telah dilakukan oleh peneliti lain dengan menggunakan metode pengklasifikasi Naïve Bayes Classifier dalam klasifikasi teks analisis sentimen. Pada penelitian ini, digunakan beberapa referensi dari buku dan beberapa sumber ilmiah untuk menjelaskan pengklasifikasian Naïve Bayes Classifier.

Pada penelitian ini akan mengolah data kemudian mengklasifikasi sentimen positif dan negatif dari media Twitter dan menguji tingkat akurasi dengan menggunakan metode

Naïve Bayes Classifier terhadap data uji pada pembentukan sentiment analysis.

Penelitian yang berkaitan dengan metode Naïve Bayes Classifier :

Tabel 3: Tinjauan Studi

Peneliti	Tahun	Judul	Metode	Hasil
Kalyan Netti dan Dr. Y Radhika	2015	A Novel Method for Minimizing Loss of Accuracy in Naive Bayes Classifier	Naïve Bayes Classifier	Hasil dari implementasi metode Naïve Bayes Classifier menghasilkan akurasi sebesar 79.09%.
John Dodd	2014	Twitter Sentiment Analysis	Naïve Bayes Classifier, Decision Tree, Random Forest and Support Vector Machine	Hasil dari pengujian dengan perbandingan beberapa metode, dan Random Forest merupakan hasil yang terbaik yaitu 58.11% untuk kelas positif dan 41.89% untuk kelas negatif.
Henny Leidiana	2013	Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor	K-Nearest Neighbor	Hasil dari pengujian dengan menggunakan algoritma K-Nearest Neighbor dengan tingkat akurasi sebesar 81.46%.
Boy Utomo Manalu	2014	Analisis Sentimen pada Twitter Menggunakan Text Mining	Naïve Bayes Classifier dan N-gram	Proses klasifikasi semakin akurat jika data latih yang digunakan dalam pembelajaran berjumlah banyak dan seleksi fitur menggunakan N-gram kata dapat meningkatkan kemampuan analisis sentimen pada Tweet.
Amir Hamzah	2012	Klasifikasi Teks dengan Naïve Bayes Classifier untuk Pengelompokan Teks Berita dan Abstrak Akademis	Naïve Bayes Classifier	Penelitian menggunakan data 1000 dokumen berita dan 450 dokumen akademik, pada dokumen berita akurasi mencapai 84.00% dan pada dokumen akademik akurasi mencapai 82.00%.

B. Landasan Teori

Data mining merupakan disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data [1] Data mining sering juga disebut knowledge discovery in database (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan[2].

Klasifikasi adalah sebuah proses untuk mencari model atau fungsi yang menjelaskan dan membedakan kelas atau konsep dari data, dengan tujuan untuk menggunakan model dan melakukan prediksi dari kelas suatu objek dimana tidak diketahui label dari kelas tersebut. Model yang ada berasal dari analasis dari kumpulan training data (objek data dimana kelas dari label diketahui).

Confusion Matrix berisi membahas informasi mengenai aktual dan prediksi klasifikasi yang dilakukan dengan sistem klasifikasi[3]. Kinerja sistem tersebut pada umumnya dievaluasi menggunakan data dalam matriks.

K-Fold Cross Validation adalah metode umum yang digunakan untuk mengevaluasi kinerja *classifier*. Dalam pendekatan cross validation, setiap *record* digunakan beberapa kali dalam jumlah yang sama untuk training dan tepat sekali untuk testing. Metode ini mempartisi data ke dalam dua sub set

data yang berukuran sama. Pilih salah satu sebagai data training dan satu lagi untuk testing, kemudian dilakukan pertukaran fungsi dari subset sedemikian sehingga subset yang sebelumnya sebagai training set menjadi test set demikian sebelumnya. Selama proses, salah satu dari partisi dipilih untuk training, sedangkan sisanya untuk testing. Prosedur ini diulangi k kali sedemikian sehingga setiap partisi digunakan untuk testing tepat satu kali. Total error ditentukan dengan menjumlahkan error untuk semua k proses tersebut.

Dalam melakukan text mining, teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau dataset mentah disebut juga dengan proses text preprocessing. Text preprocessing berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur.

Naive Bayes memungkinkan klasifikasi berdasarkan asumsi kondisi tersendiri antara prediksi attributes diberikan class. Untuk itu Naive Bayes Classifier adalah klasifikasi yang benar-benar kompeten, bekerja cukup baik dalam tugas-tugas klasifikasi sehingga banyak peneliti yang mencoba untuk meningkatkan performa Naive Bayes.

PHP adalah bahasa pemrograman yang digunakan sebagai bahasa script server-side dalam pengembangan web yang disisipkan pada dokumen HTML. Penggunaan PHP memungkinkan web dapat dibuat dinamis sehingga maintenance situs web tersebut menjadi lebih mudah dan efisien. PHP merupakan software Open Source yang disebarkan dan dilisensikan secara gratis serta dapat di-download secara bebas dari situs resminya. PHP memiliki banyak kelebihan tang tidak dimiliki oleh script sejenis.

MySQL adalah *Relational Database Management System* (RDBMS) yang cepat dan akurat. Sebuah basis data dapat membuat pengguna untuk menyimpan, mencari, mengurutkan dan mendapatkan data dengan sangat efisien. Server MySQL mengendalikan akses ke dalam data untuk memastikan bahwa para pengguna dapat bekerja dalam waktu yang bersamaan, untuk mendukung akses secara cepat dan memastikan hanya pengguna yang telah terotorisasilah yang mendapatkan hak akses.

Xampp adalah suatu bundel web server yang popular digunakan khususnya pada sistem operasi Windows karena kemudahan installasinya. Bundel program open source tersebut berisi dari server web Apache, interpreter PHP, dan basis data MySQL[4].

ISO 9126 adalah standar terhadap kualitas perangkat lunak yang diakui secara internasional. Terpenuhinya item-item pada ISO 9126 pada sebuah perangkat lunak tidak serta-merta memberikan sertifikat ISO terhadap perangkat lunak tersebut karena standar ISO juga harus dipenuhi dari sisi manajemen pembuat perangkat lunak tersebut, dengan kata lain jika manajemennya tidak memenuhi standar ISO maka hasil kerjanya pun tidak dapat diberikan sertifikat standar ISO.

III. METODE DAN LANGKAH PENELITIAN

Pada penelitian ini menggunakan metode penelitian eksperimen, dapat diartikan sebagai pendekatan penelitian kuantitatif yang paling penuh, yang berarti memenuhi semua persyaratan untuk menguji hubungan sebab akibat. Penelitian eksperimen merupakan pendekatan penelitian cukup khas. Kekhasan tersebut diperlihatkan oleh dua hal, hal pertama yaitu penelitian eksperimen menguji secara langsung pengaruh suatu variabel terhadap variabel lain, kedua menguji hipotesis hubungan sebab akibat [5].

Metode penelitian yang digunakan pada penelitian ini adalah penelitian eksperimen dengan melalui beberapa tahapan sebagai berikut :

1. Pengumpulan data

Penelitian ini diawali dengan pengumpulan data yang diambil dari tweets pada www.twitter.com dengan menggunakan fitur twitter API dan perangkat lunak R kemudian memasukkan kata kunci terkait dengan akun official first media.

Gambar 18: Data Tweet Positif

2. Pengolahan data awal

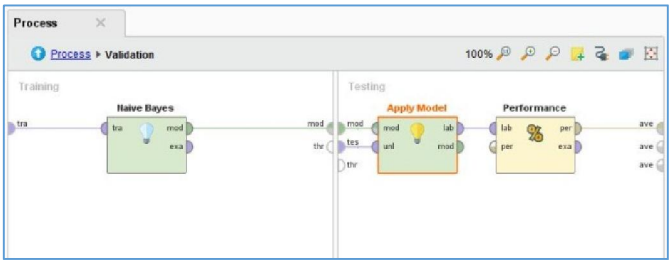
Data tweet yang akan digunakan diolah terlebih dahulu, karena masih terdapat noise. Data tersebut diolah dengan melakukan text preprocessing, yang selanjutnya diklasifikasikan dengan menggunakan metode naïve bayes classifier.

Tabel 4: Text Preprocessing Stemming

Sebelum proses *Stemming*	*"Alhamdulillah akhirnya xyz nyampe juga di ciketing Bisa ngerasain juga pake internet yang wushhh"*
Sesudah proses *Stemming*	*alhamdulillah akhir xyz nyampe juga di ciketing bisa rasa juga pake internet yang wushhh*

3. Eksperimen dan pengujian metode

Data yang sudah diolah, kemudian digunakan untuk eksperimen dan pengujian dengan sebuah metode, dan aplikasi yang digunakan adalah RapidMiner 7 untuk menghasilkan nilai akurasi dan untuk pengujian metode akan dibuat aplikasi menggunakan bahasa pemrograman PHP dan HTML.



Gambar 19: Pengujian Metode Naïve Bayes

4. Evaluasi dan validasi hasil pengujian

Setelah melakukan pengujian terhadap data tweet, maka akan muncul hasil berupa nilai keakuratan yang kemudian dianalisa, dievaluasi dan divalidasi. Untuk mengetahui nilai akurasi diukur dengan menggunakan Confusion Matrix, sedangkan untuk menguji validasi menggunakan teknik 10-Fold Cross Validation.

Tabel 5: Confusion Matrix

		Prediksi	
		Positif	Negatif
Aktual	Positif	234	29
	Negatif	16	221

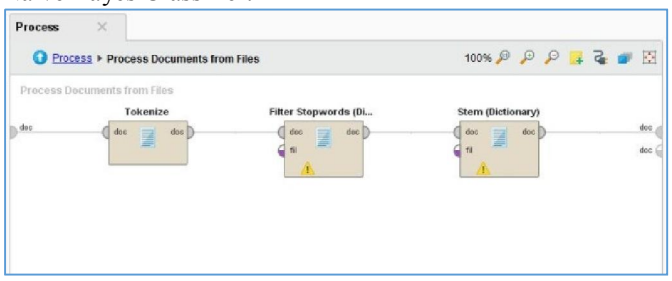
Sampel adalah sebagian dari jumlah dan karakteristik yang dimiliki oleh populasi tersebut[6]. Dalam pengambilan sampel sebaiknya menggunakan cara-cara yang lebih dapat dipertanggungjawabkan secara ilmiah.

Pada penelitian ini, penulis menggunakan teknik random sampling atau pengambilan sampel secara acak. Untuk menentukan besarnya jumlah responden atau sampel, peneliti menggunakan rumus Slovin [7] yaitu :

$$n = \frac{N}{1 + Ne^2}$$

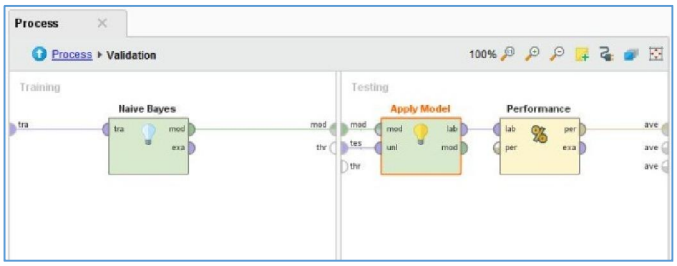
IV. HASIL PENELITIAN

Pada bab ini akan diuraikan mengenai hasil-hasil penelitian, aplikasi dan pengukuran kinerja model. Pengembangan aplikasi akan dibahas dalam pengujian aplikasi untuk menunjukkan bahwa hasil dari aplikasi yang dibuat sudah sesuai dengan yang diharapkan. Untuk pengukuran kinerja model akan dijelaskan hasilnya yang merupakan analisa sentimen tweets first media dengan menggunakan metode Naïve Bayes Classifier.



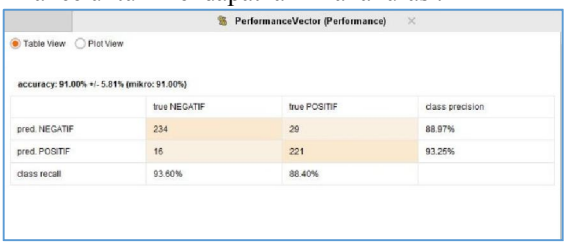
Gambar 20: Desain Preprocessing

Pada Gambar 3: merupakan *preprocessing* dari proses dokumen dimana terdapat tiga tampilan proses, yaitu proses *tokenize*, proses *stopword* dan proses *stemming*.



Gambar 21: Desain Model Naïve Bayes

Pada Gambar 4: adalah proses evaluasi dan validasi dari metode Naïve Bayes Classifier. Pada proses validasi terdapat dua bagian dan tiga tampilan proses, untuk bagian pertama yaitu bagian training dimana proses training dilakukan dengan menggunakan metode Naïve Bayes Classifier, untuk bagian kedua yaitu testing dilakukan dengan pengujian model dan performance untuk mendapatkan nilai akurasi.

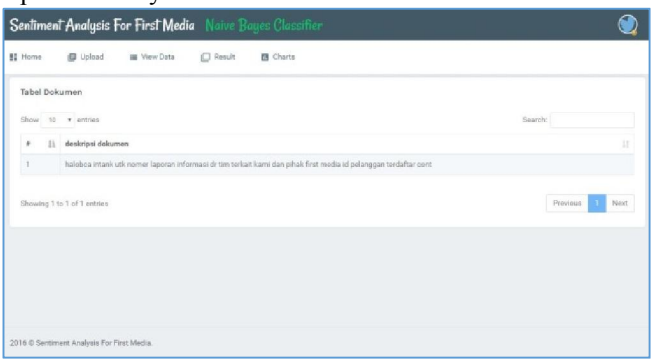


Gambar 22: Hasil Prediksi dan Akurasi Metode Naïve Bayes

Pada Gambar 5: yaitu menunjukkan hasil akurasi setelah dilakukan pengujian dengan menggunakan metode Naïve Bayes Classifier yang menghasilkan nilai akurasi sebesar **91.00%.**

A. Desain dan Implementasi

Pada penelitian ini dibuat rancangan aplikasi berbasis web dengan menggunakan bahasa pemrograman PHP dan MySQL dengan algoritma terpilih yaitu Naïve Bayes Classifier. Berikut adalah penjelasan tampilan rancangan aplikasi dan implementasinya.



Gambar 23: Menu View Data Uji

Pada Gambar 6: adalah menu view data (lihat data uji) yang berfungsi untuk melihat data uji yang sudah di-upload.

Gambar 24: Menu Charts

Menu charts adalah menu yang menunjukkan hasil analisis sentimen berupa grafik, terdapat dua grafik pada menu charts, yaitu grafik pie dan grafik batang. Pada grafik tersebut dapat terlihat persentase positif dan persentase negatif yang dihasilkan dari pengolahan data yang sudah diproses sebelumnya. Dari persentase tersebut, dapat dijadikan acuan dan bahan evaluasi untuk first media ataupun user.

B. Pengujian Blackbox dan ISO 9126

Pengujian Blackbox fokus pada persyaratan fungsional perangkat lunak. Pengujian aplikasi analisis sentimen ini menggunakan data uji berupa data input yang telah dibuat.

Tabel 6: Pengujian Menu Lihat Data Uji

Kasus dan Hasil Uji (Data Benar)			
Data yang dimasukkan	**Yang diharapkan**	**Pengamatan**	**Kesimpulan**
Klik tombol *view data* pada menu *view data* dan klik tombol lihat data uji	Jika berhasil maka tampilan data uji pada menu *view data* akan muncul	Sesuai dengan harapan	(√) diterima ( ) ditolak
Kasus dan Hasil Uji (Data Salah)			
Data yang dimasukkan	**Yang diharapkan**	**Pengamatan**	**Kesimpulan**
Klik tombol *view data* pada menu *view data* dan klik tombol lihat data uji	Jika tampilan data uji pada menu *view data* **tidak** muncul	Sesuai dengan harapan	(√) diterima ( ) ditolak

Pada pengujian kualitas ISO 9126 ini terdiri dari dua bagian, yaitu tingkat kualitas masing-masing aspek yang mengadaptasi empat karakteristik ISO 9126 dan tingkat kualitas secara keseluruhan dari empat karakteristik ISO 9126. Dari 15 responden yang mengisi angket kuesioner ini, tanggapan dari masing-masing responden terhadap indikator kualitas perangkat lunak dengan mengadaptasi ISO 9126.

Tabel 7: Tanggapan Responden Berdasarkan Aspek Functionality

Kriteria Jawaban	Bobot	Suitability	Accuracy	Interoperability	Compliance	Total
Sangat setuju	5	1	2	2	1	30
Setuju	4	13	11	7	17	192
Ragu-ragu	3	3	3	9	1	48
Tidak setuju	2	2	3	1	-	12
Sangat tidak setuju	1	-	-	-	-	-
Jumlah responden		19	19	19	19	76
Skor aktual		70	69	67	76	282
Skor ideal		95	95	95	95	380

V. KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Berdasarkan hasil dari implementasi dan pengujian pada bab-bab sebelumnya, maka dapat ditarik kesimpulan sebagai berikut :

1. Aplikasi analisis sentimen dapat mempermudah dalam proses pengolahan data pada produk layanan internet first media.
2. Dapat melakukan pengklasifikasian sentimen positif dan sentimen negatif terhadap produk layanan *internet* first media.
3. Klasifikasi dengan menggunakan metode *Naïve Bayes Classifier* mendapatkan hasil yang akurat dalam menentukan *tweets* kedalam sentimen positif dan sentimen negatif.

5.2 SARAN

Dari hasil implementasi pengujian metode dan pengujian aplikasi pada penelitian ini, peneliti menyadari masih banyak kekurangan dalam pengembangan aplikasi ini. Adapun saran-saran yang dapat peneliti berikan antara lain :

1. Pengklasifikasian sentimen bisa dikembangkan dengan menambahkan sentimen netral.

Aplikasi analisis sentimen dapat dikembangkan dengan menggunakan metode yang berbeda atau komparasi metode yang berbeda

DAFTAR PUSTAKA

[1] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques.* San Fransisco: Diane Cerra.

[2] Santoso, B. (2007). *Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

[3] Kohavi, R., & Provost, F. (1998). Machine Learning. 271-274.

[4] Nugroho. (2004). *Latihan Membuat Aplikasi Web PHP dan MySQL dengan Dreamweaver MX.* Yogyakarta: Gaya Media.

[5] Sukmadinata, N. S. (2009). *Metode Penelitian Pendidikan.* Bandung: Rosdakarya.

[6] Sugiyono. (2012). *Metode Penelitian Pendidikan.* Bandung: Alfabeta.

[7] Prasetyo, B. (2005). *Metode Penelitian Kuantitatif Teori dan Aplikasi.* Jakarta: PT. Raja Grafindo Persada