KLASIFIKASI OPINI : PENGGUNA SMARTPHONE PADA TWITTER DI INDONESIA

Ega Febri Dharmawan[1)], Eka Dyar Wahyuni[2)], Amalia Anjani Arifiyanti[3)]
E-mail : [1)]egafebridh@gmail.com, [2)]ekadyar.w@gmail.com,
[3)]amalia_anjani.fik@upn.jatim.ac.id

[1,2,3]Sistem Informasi, Fakultas Ilmu Komputer, UPN Veteran Jawa Timur

Abstrak

Twitter merupakan salah satu sosial media yang memiliki pengguna paling banyak di dunia. Di indonesia, sebanyak 19,5 juta orang aktif menggunakan *twitter* tahun 2019. Dengan banyaknya pengguna, maka *tweet* yang dihasilkan perharinya juga sangat banyak. Ini adalah alasan mengapa *twitter* merupakan sosial media yang tepat di dalam melakukan klasifikasi opini. Dengan klasifikasi opini, banyak hal menarik yang didapatkan. Seperti mengetahui bagaimana respon masyarakat tehadap sebuah permasalahan hingga menganalisa tanggapan tersebut. Persaingan tiap perusahaan *smartphone* untuk meningkatkan kualitasnya dan mendapatkan pelanggan perlu adanya analisa, salah satunya adalah dengan menggunakan klasifikasi opini. Pembuatan sistem klasifikasi opini pengguna *smartphone* pada *twitter* di indonesia ini dibangun menggunakan python dan framework django. Penelitian dilakukan dengan melakukan studi literatur, lalu analisis kebutuhan sistem dan data, perancangan model, pembangunan model menggunakan algoritma Naive Bayes, evaluasi model, perancangan sistem, pembangunan sistem dan implementasi sistem. Sistem yang dibuat mampu menghasilkan perbandingan opini dari *tweet* pengguna smartphone yang dicari dalam *Twitter*. Sistem ini dibangun dengan menggunakan model opini yang dibuat untuk memisahkan *tweet* opini dengan *tweet* fakta. Model opini memiliki dataset dengan jumlah 500 *tweet* yang diambil dalam waktu 1 januari 2018 hingga 31 desember 2018. Pengujian *hamming distance* juga dilakukan pada dataset dan menghasilkan nilai sebesar 99.4%. Pengambilan dataset diambil dengan menggunakan *keyword* tentang *smartphone* seperti samsung, xiaomi, advan, vivo dan oppo. Pengujian dilakukan dengan menggunakan beberapa skenario dengan parameter algoritma (Multinomial Naive Bayes, Gaussian Naive Bayes, dan Bernoulli Naive Bayes) dan jumlah *test size*. Hasil terbaik untuk model opini adalah akurasi sebesar 88.6%.

Kata Kunci: *Opinion, Klasifikasi, Smartphone, Twitter, Opinion Classification*

1. PENDAHULUAN

Kemudahan yang di tawarkan oleh internet saat ini membuat penggunanya bertambah pesat setiap tahunnya. Dengan internet kita dapat melakukan banyak hal mulai dari mencari informasi, menambah pengetahuan, dan bahkan menjalankan bisnis. Bedasarkan data yang telah diberikan oleh Asosiasi Penyelenggara jasa Internet Indonesia (APJII), di tahun 2019 sebanyak 64,8% masyarakat Indonesia sudah menggunakan internet [1]. Popularitas jaringan media sosial pun ikut melesat karena adanya internet. Jaringan media sosial mengubah cara masyarakat dalam berkomunikasi dan berinteraksi setiap harinya. Banyak yang memanfaatkan sosial media untuk memperluas jaringan dan mencari informasi. Media sosial juga dapat membantu dalam perkembangan bisnis suatu perusahaan. Salah satu media sosial yang paling populer di Indonesia adalah *Twitter*. *Twitter* adalah layanan *micro-blogging* yang membantu penggunanya mengekspresikan sendiri tentang berbagai topik. Penggunanya dapat memposting 280 karakter pesan teks yang disebut *tweet* (*Twitter* telah menggandakan ukuran *tweet* dari 140 karakter pada 2017). Jutaan *tweet* dihasilkan setiap hari, sehinga *tweet* yang ditulis oleh pengguna dibatasi oleh ukuran pesan teks [2]. Popularitas *Twitter* di Indonesia sudah sangat besar.

Bedasarkan data dari 2 Kementrian Komunikasi dan Informatika [3], lebih dari 19,5 Juta orang indonesia aktif menggunakan *twitter*. *Twitter* memiliki beragam fitur yang membuat penggunanya mudah untuk berkomunikasi dan berinteraksi dengan pengguna lainya. Pengguna dapat membuat akun lalu memposting *tweet* yang berisi foto, video, atau teks yang akan di tampilkan di *timeline twitter* pengguna, lalu dapat direspon oleh pengguna lain. *Tweet* tersebut dapat di like dan di *retweet* atau dibagikan di *timeline* kita. *Tweet* dapat ditemukan dengan mencari di mesin pencarian dengan menggunakan *keyword* atau *hashtag* tertentu. Dengan semua fasilitas ini kita dapat menemukan banyak informasi online yang menjadi alasan mengapa *twitter* merupakan sumber data yang tepat dalam melakukan penelitian ini, Klasifikasi pada opini ini sangat penting karena terdapat pengetahuan yang bisa menjadi dasar dalam membuat keputusan bisnis untuk perusahaan. Seperti bagaimana respon pengguna *smartphone* terhadap merek *smartphone* tertentu yang nantinya dapat berguna bagi perusahaan untuk mengevaluasi produk tersebut dan melihat pasar *smartphone* di Indonesia.

Bedasarkan dari penelitian yang di lakukan oleh Anmol Nayak dan Dr. S. Natarajan dari PES University Bangalore, India yang berjudul "Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Opinion Analysis of Twitter feeds", membuktikan bahwa Naive Bayes memiliki presentase akurasi terbesar dibanding Random Forest dan Support Vektor Machine. Penelitian yang bertujuan untuk membandingkan algoritma terbaik pada klasifikasi opini ini menggunakan dua jenis kelas opini yaitu opini dan fakta. Kesimpulan dari penelitian ini mengatakan bahwa Naive Bayes memiliki akurasi sebesar 89%, lalu SVM sebesar 88% dan Random Forest sebesar 85%. Bedasarkan hasil penelitian tersebut maka diputuskan untuk menggunakan metode Naive Bayes pada penelitian ini.[4] Di dalam penelitian ini akan digunakan beberapa jenis metode dari Naive Bayes yaitu Multinomial Naive Bayes, Gaussian Naive Bayes dan Bernoulli Naive Bayes. Hal ini dilakukan sebagai variasi skenario di dalam pembuatan model pada penelitian ini, setelah ditemukan akurasi terbaik maka akan diimplementasikan kedalam sistem. Perbandingan metode pada naive bayes pernah dilakukan sebelumnya oleh M. Ali Fauzi pada penelitiannya yang berjudul "Automatic Complaint Classification System Using Classifier Ensembles" pada tahun 2018. Dengan membandingkan banyak metode penelitian seperti Multinomial NB, Gaussian NB dan Bernoulli NB pada data dokumen di dalam aplikasi sambat online. Hasilnya Multinomial memperoleh presentase sebesar 80%, sedangkan Gaussian sebesar 65.5% dan Bernoulli sebesar 67.7%. [5].

2. METODOLOGI

2.1 Studi Literatur

Studi Literatur yang dilakukan yaitu dengan mencari dan mereview berbagai buku, peneitian, skripsi dan jurnal yang membahas mengenai klasifikasi opini pada twitter, penggunaan metode *naive bayes* pada *machine learning*, dan beberapa tinjauan pustaka mengenai penelitian sebelumnya yang relevan.

2.2 Pembangunan Model Klasifikasi

Di dalam pembuatan model klasifikasi opini terdapat beberapa proses seperti pengumpulan data yang diambil dari twitter, penyaringan data *tweet* untuk menghilangkan data yang tidak diperlukan di dalam pembuatan model klasifikasi Opini, pelabelan tiap *tweet* untuk membagun model, pembagian data set dan data training, pembobotan dengan TF-IDF, klasifikasi dengan metode *naive bayes*, dan melakukan evaluasi model klasifikasi untuk mengukur akurasi dari model yang dibangun, dan pembahasan hasil model. Proses ini dilakukan dengan menggunakan library *scikit learn*. Terdapat 3 variasi Algoritma Naive Bayes yang akan digunakan di dalam penelitian ini, yaitu :

a) Multinomial Naive Bayes

Multinomial sendiri mengimplementasikan algoritma naive Bayes untuk data yang terdistribusi secara multinomial, dan merupakan salah satu dari dua varian naive Bayes klasik yang digunakan dalam klasifikasi teks di mana data 14 biasanya diwakili sebagai jumlah vektor kata. Distribusi dibatasi oleh vektor x untuk setiap kelas y, di mana jumlah fitur dalam klasifikasi teks dan ukuran kosa kata merupakan probabilitas P(x|y) dari fitur i yang muncul dalam sampel milik kelas y. Penghitungan frekuensi relatif pada multinomial yaitu :

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Gambar 1. Rumus Multinomial Naive Bayes

b). Bernoulli Naive Bayes

Bernoulli mengimplementasikan pelatihan naive Bayes dan algoritma klasifikasi untuk data yang didistribusikan sesuai dengan distribusi Bernoulli multivariat. Ada beberapa fitur yang masing-masing diasumsikan sebagai variabel binary-valued. Oleh karena itu, kelas ini membutuhkan sampel untuk direpresentasikan sebagai vektor fitur bernilai biner; jika menyerahkan jenis data lainnya, turunan Bernoulli dapat membuat binari inputnya. Aturan keputusan untuk Bernoulli naif Bayes didasarkan pada :

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

Gambar 2. Rumus Bernoulli Naive Bayes

c). Gaussian Naive Bayes

Gaussian mengimplementasikan algoritma Gaussian Naive Bayes untuk klasifikasi. Rumus Algoritma Gaussian :

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Gambar 3. Rumus Gaussian Naive Bayes

2.3 Evaluasi Model

Tahapan yang terakhir di dalam pembagunan model opini adalah evaluasi model klasifikasi dengan menggunakan *confusion matrix*. Evaluasi digunakan untuk mengetahui seberapa akurat model yang sudah dibuat Metode yang digunakan untuk evaluasi pada penelitian ini adalah confusion matrix. Menurut Manning, Confusion matrix merupakan salah satu tools penting dalam metode evaluasi yang digunakan pada mesin pembelajaran yang biasanya 16 memuat dua kategori atau lebih [6]. Hasil akurasi tersebut didapatkan dengan membandingkan data yang sudah di beri label benar atau salah pada model dengan hasil prediksi label yang nantinya akan digunakan untuk diimplementasikan di dalam sistem. Tabel 1 adalah gambaran dari confusion matrix.

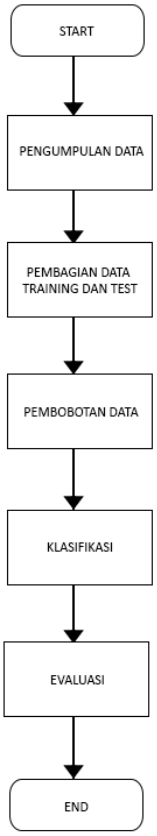
Tabel 1. Pengujian Confusion Matrix

		Actual Class	
		Class 1	Class 2
Predicted Class	Class 1	True Positive	False Negative
	Class 2	True Negative	False Positive

Model juga akan diuji dengan menggunakan hamming distance untuk mengukur hasil akurasi dengan ketepatan data yang sudah diberikan label (dataset).

3. HASIL DAN PEMBAHASAN

3.1 Pembangunan Model Klasifikasi



Gambar 4. Flowchart proses sistem

Terdapat beberapa proses di dalam pembangunan model klasifikasi Opini. Tahapan tersebut antara lain adalah sebagai berikut.

a). Pengumpulan Data

Di dalam mengumpulkan data pada pembuatan model opini dari twitter, data dikupulkan dengan menggunakan library twitterscraper dengan memasukkan keyword merek smartphone seperti samsung, advan, vivo, oppo, dan xiaomi. Dengan batasan tanggal 1 januari 2018 hingga 31 desember 2018 dengan hanya tweet berbahasa indonesia yang diambil. Proses pengumpulan data tweet dilakukan hingga data berjumlah 500 tweet.

Tabel 2. Contoh Hasil Pengumpulan Tweet

No.	Text
1	Oppo: Kami yakin Find X Tak akan Mengecewakan http://dlvr.it/Qc1DLC pic.twitter.com/dbiviiu0Ko
2	Peluncuran Oppo Bertabur Bintang, Raisa Hingga Vanesha Prescilla. #willsznet #detikinet http://tinyurl.com/y8562ych
3	#PopulerB1 1: Meluncur di Indonesia, Ini Harga Oppo Find X http://brt.st/5VhY pic.twitter.com/PUCBaKCkMV
4	begonya gue trade hp ama adek, dari oppo f1s 64gb ke samsung j7 2016 16gb -_-

b). Pelabelan Data

Pelabelan data yang dilakukan dengan *semisupervised*, data akan diberikan label secara manual dan akan diuji dengan menggunakan *hamming distance.* Dari dataset yang sudah dibuat pada tahapan pengumpulan data yang berjumlah 500 tweet, didapat jumlah tweet opini sebesar 250 tweet dan jumlah tweet fakta 250 tweet.

Tabel 3. Contoh Hasil Pelabelan Tweet

No.	Text	Label
1	Oppo: Kami yakin Find X Tak akan Mengecewakan http://dlvr.it/Qc1DLC pic.twitter.com/dbiviiu0Ko	Fakta
2	Peluncuran Oppo Bertabur Bintang, Raisa Hingga Vanesha Prescilla. #willsznet #detikinet http://tinyurl.com/y8562ych	Fakta
3	#PopulerB1 1: Meluncur di Indonesia, Ini Harga Oppo Find X http://brt.st/5VhY pic.twitter.com/PUCBaKCkMV	Fakta
4	begonya gue trade hp ama adek, dari oppo f1s 64gb ke samsung j7 2016 16gb -_-	Opini

c) Pembagian Dan Pembobotan Data

Pembagian data latih dan data uji menggunakan metode Hold Out Method. Data yang digunakan adalah dataset berjumlah 500 tweet. Di dalam penelitian ini digunakan 3 metode sebagai klaisfikasi. Diantaranya Multinomial Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes. Pada proses ini menggunakan library dari scikit-learn dengan fungsi TfidfVectorizer(). Untuk data X_train akan divektorisasi dengan menggunakan fungsi fit_transform() karena data yang ada di dalamnya masih berbentuk text. Sedangkan untuk X_test akan divektorisasi dengan menggunakan fungsi transform(). Untuk metode klasifikasi yang akan diterapkan hanya salah satu dari ketiga variasi Naive Bayes diatas.

e). Klasifikasi

Model yang sudah disimpan akan di load kembali dengan menggunakan library joblib. Setelah model di load, lalu akan dilakukan proses klasifikasi dimana sistem akan memprediksi label (opini atau fakta) dari tweet yang ada pada dataset tersebut.

3.2 Pengujian Model Klasifikasi

Pada penelitian ini akan dibuat 10 skenario dengan variasi algoritma dan test size yang berbeda. Tabel 4 menggambarkan detail dari pengujian skenario pada penelitian ini.

Tabel 4. Pengujian Akurasi Model

Algoritma	Test_Size	Nilai Akurasi
Multinomial Naive Bayes	0.2	88 %
Multinomial Naive Bayes	0.3	88.6 %
Multinomial Naive Bayes	0.4	87.5 %
Multinomial Naive Bayes	0.5	84.8 %
Gaussian Naive Bayes	0.2	46 %
Gaussian Naive Bayes	0.3	46.6 %
Gaussian Naive Bayes	0.4	49.5 %
Bernoulli Naive Bayes	0.2	88 %
Bernoulli Naive Bayes	0.3	86 %
Bernoulli Naive Bayes	0.4	87 %

Meskipun menggunakan confusion matrix, pada penelitian ini hanya akan mempertimbangakn besar akurasi saja. Bedasarkan hasil diatas didapat hasil terbaik dengan menggunakan algoritma multinomial Naive Bayes dengan test_size sebesar 0.2. Model ini akan diimplementasikan kedalam sistem. Setelah model terbaik ditemukan, dataset yang diberikan label secara manual akan diuji dengan menggunakan *hamming distance* untuk mengetahui ketepatan data dengan hasil prediksi.

Tabel 5. Pengujian Hamming Distance

Nama Dataset	Manual Label		Predicted Label		*Hamming Distance*
	Opini	Fakta	Opini	Fakta	
Dataset Opini	250	250	253	247	99.4 %

Setelah melalui pengujian *hamming distance* didapatkan nilai sebesar 99.4%. Ini membuktikan bahwa data yang sudah diberikan label dapat digunakan didaam penelitian ini atau label pada dataset tidak bersifat subjektif dan mampu dimengerti oleh model.

4. KESIMPULAN DAN SARAN

Kesimpulan dari penelitian ini adalah model yang dibuat dapat berkerja dengan baik. Dengan menggunakan algoritma multinomial Naive Bayes, didapatkan model dengan akurasi yang cukup tinggi yaitu 88.6%. Dataset yang dibuat untuk membuat model ini juga terbukti dapat menghasilkan nilai *hamming distance* yang tinggi sebesar 99.4%, yang membuktikan bahwa hasil prediksi model sudah sangat akurat dengan hasil pelabelan manual.

Saran untuk penelitian selanjutnya adalah menggunakan tahapan preproccessing di dalam pembuatan model. Tahapan ini berfungsi untuk membersihkan data agar data dapat lebih mudah dipelajari oleh model. Di dalam pengujian model, penggunaan confusion matrix dapat mempertimbangkan nilai selain akurasi, seperti *precision, F1 score* dan *recall.*

5. DAFTAR RUJUKAN

[1] Pratomo. Y, 2019. APJII: Jumlah Pengguna Internet di Indonesia Tembus 171 Juta Jiwa. https://tekno.kompas.com/read/2019/05/16/03260037/apjii-jumlah-pengguna-internet-di-indonesia-tembus-171-juta-jiwa.(Diakses pada 16 Agustus 2019)

[2] Juei H. W, 2018. Measuring the Spreading of News on Twitter. International Journal of Computer Applications.

[3] Kominfo, 2019. Indonesia Peringkat Lima Pengguna Twitter. https://kominfo.go.id/content/detail/2366/indonesia-peringkat-lima-pengguna-twitter/0/sorotan_media (Diakses pada 16 Agustus 2019)

[4] Anmol N. & Natarajan S. 2015 .“International Journal of Advanced Studies in Computer Science and Engineering”. IJASCSE Volume 5, Issue 1

[5] Fauzi, M. A. 2018. Automatic Complaint Classification System Using Classifier Ensembles. Telfor Journal, Vol. 10, No. 2, 2018.

[6] Trajdos P. & Kurzynski. M, 2018). Weighting Scheme for a Pairwise Multi-label Classifier Based on the Fuzzy Confusion Matrix. Pattern Recognition Letters.