Komparasi Akurasi Naïve Bayes dan Support Vector Machine (SVM) untuk Rekomendasi Produk in Fashion Dress

Nancy Ria Silvani Huaturuk[1], Rima Dias Rahmadani[2], Dwi Januarita AK[3]
[1,2,3]Program Studi Informatika, Fakultas Teknik Industri dan Informtika, Institut Teknologi Telkom Purwokerto
[1,2,3]5314, Purwokerto Kulon, Jawa Tengah, Indonesia
nancyhtrk@gmail.com, [2] Rima@ittelkom-pwt.ac.id, dwijanuarita@ittelkom-pwt.ac.id

Abstrak - Data mining atau penemuan pengetahuan adalah prosedur menggunakan teknik statistik dan berbasis pengetahuan untuk menganalisis data dengan pola tambang yang memiliki makna dari kumpulan data yang luas dan mengubahnya menjadi informasi bermanfaat. Pada data mining terdapat beberapa metode dalam data mining salah satunya adalah klasifikasi. Klasifikasi merupakan salah satu teknik data mining yang digunakan untuk membangun suatu model dari sampel data yang belum terklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas-kelas yang sejenis. Pada penelitian ini digunakan metode klasifikasi Naive Bayes dan SVM untuk melakukan klasifikasi pada data rekomendasi produk dress. Naïve Bayes digunakan karena memiliki kelebihan diantaranya adalah algoritma sederhana tapi memiliki akurasi yang tinggi. Dan SVM juga memiliki kelebihan yaitu svm juga memiliki tingkat akurasi yang tinggi dan dapat bekerja sangat baik pada data dengan banyak dimensi dan menghindari kesulitan dari permasalahan dimensionalitas. *Tools* Matlab digunakan untuk melakukan proses klasifikasi pada dataset Dresses_attribute_sales yang diambil malalui UCI *Repository.*Hasil penelitian ini mendapatkan akurasi pada Naive Bayes sebesar 74% dan pada SVM sebesar 66%.

Kata kunci- *Naive Bayes, SVM, Klasifikasi, Dress*

Abstract– Data mining or knowledge discovery is a procedure using statistical and knowledge-based techniques to analyze data with a mine pattern that has the meaning of a vast data set and turn it into useful information. In data mining there are several methods in data mining one of which is classification. Classification is one of the data mining techniques used to construct a model of unclassified data samples to be used to classify new sample data into similar classes. In this study used the classification method Naive Bayes and SVM to do the classification on the data recommendation dress products. Naïve Bayes is used because it has advantages among them is simple algorithm but has a high accuracy. And SVM also has advantages that svm also has a high degree of accuracy and can work very well on the data with many dimensions and avoid the difficulty of dimensionality problems. The Matlab tool is used to perform the classification process on the dataset of Dresses_attribute_sales taken by UCI Repository. The results of this study get accuracy on Naive Bayes by 74% and on SVM by 66%.

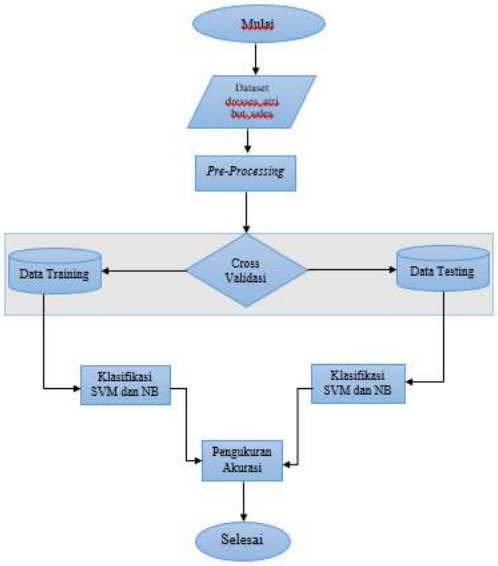
Keywords - Dataset, Naive Bayes, SVM, Classification, Dress

I. PENDAHULUAN

Data mining atau penemuan pengetahuan adalah prosedur menggunakan teknik statistik dan berbasis pengetahuan untuk menganalisis data dengan pola tambang yang memiliki makna dari kumpulan data yang luas dan mengubahnya menjadi informasi bermanfaat [1]. Menurut Mirkin data mining didefinisikan sebagai suatu proses untuk mencari pola dari sekumpulan data yang terdapat di dalam database untuk kemudian dianalisis sehingga menghasilkan suatu informasi tertentu untuk dapat dimanfaatkan pada proses selanjutnya [2]. Salah satu pendekatan metode statistika yang dapat melakukan pengkategorian adalah klasifikasi. Klasifikasi merupakan salah satu teknik data mining yang digunakan untuk membangun suatu model dari sampel data yang belum terklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas-kelas yang sejenis [3]. Klasifikasi termasuk ke dalam *supervised learning* karena menggunakan sekumpulan data untuk dianalisis terlebih dahulu, kemudian pola dari hasil analisis tersebut digunakan untuk pengklasifikasian data uji. Proses klasifikasi data terdiri dari pembelajaran dan klasifikasi. Pada pembelajaran data training dianalisis menggunakan algoritma klasifikasi, selanjutnya pada klasifikasi digunakan data testing untuk memastikan tingkat akurasi dari rule klasifikasi yang digunakan [4].Terdapat beberapa algoritma pada metode

klasifikasi yaitu Naive bayes dan Support Vector Machine (SVM). Naïve Bayes adalah salah satu algoritma teknik data mining yang menerapkan teori Bayes dalam klasifikasi dimana algoritma ini mengasumsikan bahwa atribut suatu objek bersifat independen atau bebas [5]. Metode Naïve Bayes telah banyak digunakan dalam penelitian mengenai text mining, beberapa kelebihan Naïve Bayes diantaranya adalah algoritma sederhana tapi memiliki akurasi yang tinggi [6]. Teknik SVM berakar pada teori pembelajaran statistik dan telah menunjukkan hasil empiris yang menjanjikan dalam berbagai aplikasi praktis dari pengenalan digit tulisan tangan sampai kategorisasi teks. SVM juga bekerja sangat baik pada data dengan banyak dimensi dan menghindari kesulitan dari permasalahan dimensionalitas [7]. Berdasarkan penelitian sebelumnya, naïve bayes memiliki kelebihan yang dapat menangani data dengan jumlah yang besar dan tidak mempengaruhi *irrelevant atribute. Naïve Bayes* juga memiliki kelebihan dalam kesederhanaannya [7]. SVM memiliki kelebihan dimana tingkat akurasi yang lebih baik dibandingkan Naïve Bayes [8]. Berdasarkan penelitian terkait yang menyatakan bahwa hasil akurasi SVM lebih baik dibandingkan dengan Naïve Bayes, maka peneliti melakukan perbandingan algoritma SVM dan NB dengan studi kasus yang berbeda yaitu rekomendasi produk in *fashion dress* menggunakan dataset Dresses_Attribute_Sales dari UCI Machine Learning [1]. Penelitian ini diharapkan mampu memberikan rekomendasi algoritma yang sesuai dengan kebutuhan data untuk dapat diterapkan pada sistem klasifikasi untuk rekomendasi fashion dress masa kini.

II. METODE PENELITIAN

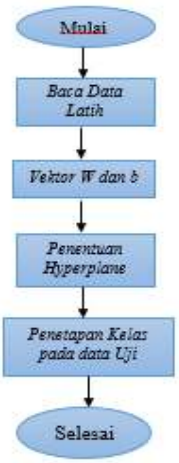


Gambar 2.1 Diagram Alur Perancangan Model

Untuk perancangan ini menggunakan metode Naïve Bayes dan SVM dengan tools Matlab. Parameter yang digunakan adalah atribut berdasarkan dataset dresses_atribut_sales. Pada Gambar 2.1, digambarkan Diagram perancangan model algoritma Naïve Bayes dan SVM dengan penjelasan alur sebagai berikut :

1. Dataset yang digunakan berasal dari dresses_atribut_sales pada tahun 2014 dan dataset berjumlah 500 data melalui UCI Repository
2. Pada tahap Melakukan Pre-proccessing dataset akan divalidasi kebersihan data dari noise, data yang hilang (missing) dan yang tidak konsisten dalam hal ini dilakukan tahap *replacing missing value* yaitu mengganti nilai yang hilang dengan hasil dari nilai *mean* dari sebuah atribut.
3. Tahap selanjutnya adalah K-fold cross-validation, dimana dataset akan dibagi menjadi dua yaitu data latih dan data uji. Pembagian akan dilakukan secara acak berdasarkan nilai iterasi.
4. Setelah memiliki data uji kemudian akan dilakukan klasifikasi menggunakan algoritma SVM linier dan Naive Bayes. Pengklasifikasian data dilakukan terpisah setiap algoritma.
5. Pengukuran akurasi dilakukan dengan klasifikasi yang dibagi menjadi dua kelas, recommended dan tidak recommended. Hasil klasifikasi setiap metode akan diuji keakurasiannya menggunakan persamaan akurasi. Kemudian perbandingan kedua metode dilakukan setelah hasil akurasi kedua metode keluar. Matlab digunakan untuk melakukan pengukuran akurasi.
6. Selesai.

Gambar 2.2 Flowchart Klasifikasi Naïve Bayes



Gambar 2.3 Flowchart Klasifikasi SVM

Pada Gambar 2.2 menunjukkan tahapan algoritma Naive Bayes yang digunakan pada penelitian ini. Dimulai dengan membaca data latih Dresses_Attribute_Sales yang sudah dibagi melalui k-fold validation kemudian dilanjutkan dengan proses pemodelan dengan mencari rata-rata (mean) dan standar deviasi dari setiap atribut yang ada pada dataset dresses. Selanjutnya memasuki proses solusi menggunakan data uji.

Pada Gambar 2.3 menggambarkan alur tahapan algoritma SVM dengan melakukan klasifikasi. Dimulai dengan pembacaan data latih yang sudah dibagi melalui k-fold validation lalu pembuatan model dengan data latih. Model akan digunakan untuk menentukan hyperplane. Selanjutnya penetapan kelas lalu selesai.

III. HASIL PENELITIAN

A. Nominal to Numeric

Pada penelitian ini dilakukan pengubahan dataset *nominal to numerical* sehingga atribut Dresses_Attribute_Sales yang bersifat kategorik diubah menjadi data angka. Proses pengubahan data *nominal to numerical* dilakukan dengan cara menginput angka berdasarkan isi dari record atribut terbanyak sampai yang paling sedikit dan dimulai dari angka 1.

B. Preprocessing

Untuk mengganti nilai yang hilang pada dataset dilakukan *pre-processing*. Data set ini memiliki 2 missing value pada atribut data. Price, 2 missing value pada atribut Season, 3 missing value pada atribut NeckLine, 2 missing value pada atribut SleeveLength, 87 missing value pada atribut Waistline, 128 missing value pada atribut Material, 266 missing value pada atribut FabricType, 236 missing value pada atribut Decoration, 109 missing value pada atribut Pattern Type. Pada proses ini dilakukan normalisasi untuk menghilangkan *misisng value* dengan cara menghitung satu per satu rata-rata dari masing-masing atribut yang memiliki *missing value* untuk mengetahui nilai rata-rata yang dimiliki oleh setiap atribut tersebut.

Pada perhitungan manual untuk mengganti nilai/ data yang hilang ditemukan hasil perhitungan rata-ratanya yang dihitung berdasarkan persamaan (4.1) karena terdapat 9 atribut yang memiliki missing value secara berurut adalah 2, 2, 2, 2, 1, 2, 1, 3, 1. Hasil dari perhitungan rata-rata atribut tersebut diinput ke dalam atribut yang hilang.

Berikut ini contoh salah satu atribut yang dilakukan perhitungan manual missing value dengan mengikut persamaan berikut :

$$\bar{x} \frac{\sum_{i=1}^{N} Xi}{N}$$

$\bar{x}$ = Rata-rata
N = Jumlah Seluruh Data
Xi = Data ke i

Persamaan diatas selanjutnya diterapkan pada nilai-nilai yang hilang pada 9 atribut yaitu pada atribut Price, Season, NeckLine, SleeveLength, Waistline, Material, FabricType, Decoration, Pattern Type.

Tabel 3.1 Nilai Pada Atribut Price

Price						
X*i*	1	2	3	4	5	?
N	174	252	30	21	21	2
X*i*N	174	504	90	84	105	

$\bar{x}$= (174 + 504 + 90 +84 + 105)/500

$\bar{x}$ = 957/500 = 1.914

Hasil dari perhitungan dengan nilai rata-rata atribut price adalah 1.914, angka tersebut akan dibulatkan menjadi 2. Jadi nilai missing value pada atribut price akan diganti dengan angka 2.

C. K-fold Cross Validation

Setelah dataset bersih dari missing value selanjutnya adalah membagi data menjadi dua bagian yaitu training dan testing. Proses ini dilakukan dengan Cross validation menggunakan K-fold, dimana K adalah jumlah iterasi dalam pembagian dataset.

D. Klasifikasi Dengan Naive Bayes

Klasifikasi kedua algoritma dilakukan secara terpisah untuk dihitung akurasi pada masing-masing algoritma. Naive Bayes memiliki persamaan yang berbeda tergantung jenis data yang digunakan. Pada kasus ini akan menggunakan Naive Bayes dengan Gaussian. Karena data yang dimiliki bersifat kontinu bukan polinomial. Berikut persamaan Naive Bayes dengan menggunakan Gaussian.

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

Telah dimiliki data latih sebanyak 450 data yang terdiri dari :

a. 189 data dengan kelas Recommended (1)
b. 261 data dengan kelas tidak Recommended (0)

Dataset secara default memiliki 9 atribut dan 2 label atau kelas. Kemudian kedua kelas/label dihitung secara terpisah untuk mencari rata-rata (mean) dan standar deviasi pada atribut masing-masing kelas menggunakan persamaan berikut :

Mean :

$$\mu = \frac{\sum_{i=1}^{N} Xi}{N}$$

Standar Deviasi :

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(xi - \mu)^2\right]$$

Pada Tabel 4.11 dibawah in diketahui adalah hasil dari pehitungan mean dan standar deviasi dari setiap atribut pada setiap kelas/label recommended dan tidak recommended yang telah diketahui sebagai berikut :

Tabel 4.11 Hasil Mean dan Standar Deviasi Kelas Recommended (1)

Atribut	*Mean* (μ)	Standar Deviasi (μ)
1	2,682539683	2,009223917
2	2,031746032	1,129416917
3	3,65820105	1,945179241
4	1,978835979	0,983684303
5	2,285714286	0,985456554
6	2,206349206	2,087134324
7	1,957671958	1,303965669
8	1,248677249	0,490940095
9	2,650793651	2,402295424
10	2,380952381	2,45661442
11	5,137566138	3,733419573
12	2,126984127	1,485662652

Hasil Mean dan Standart Deviasi yang diketahui pada kelas Tidak *Recommended* (0) dapat lihat pada Tabel 4,12 dibawah ini.

Tabel 4.12 Hasil Mean dan Standar Deviasi Kelas Tidak Recommended (0)

Atribut	*Mean* (μ)	Standar Deviasi (μ)
1	2,819923372	2,468677938
2	1,823754789	0,84556017
3	3,398084291	2,067206441
4	2,14559387	1,075008825
5	2,107279693	1,083079954
6	2,283524904	2,313022629
7	2,264367816	1,484222952
8	1,22605364	0,437047943
9	2,904214559	3,300702352
10	2,061302682	1,649031367
11	4,942528736	3,41272389
12	2,287356322	1,886641578

Kemudian langkah selanjutnya akan dihitung menggunakan persamaan Naive Bayes dengan persamaan Naive Bayes classifier dengan menggunakan persamaan sebagai berikut

$$P(y) = \underset{y \in \{1,\ldots,K\}}{\operatorname{argmax}} \prod_{i=1}^{n} P(x_i \mid y_k).$$

Hasil perhitungan manual pada klasifikasi rekomendasi produk dress menggunakan algoritma naive bayes adalah :

P (y = Recommended | x) = 2.02955E-09

P (y = Tidak Recommended | x) = 9.08236E-10

Berdasarkan hasil diatas pada perhitungan algoritma Naive Bayes tergolong pada kelas Tidak recommended. Karena nilai yang dihasilkan lebih besar pada kelas tidak recommended dibandingkan dengan kelas recommended.

Pada Perhitungan SVM data latih yang digunakan adalah diambil secara acak dikarena data yang digunakan pada penelitian ini memiliki banyak atribut. Berikut ini hasil dari eliminasi nilai W dari setiap atribut dan b dari nilai bias.

Tabel 4.16 Nilai Bobot SVM

Atribut	**Bobot**
Style	-4.6143
Price	1.2963
Rating	-0.2870
Size	1.1215
Bias (*b*)	0.44454

Setelah tahap diatas selesai dilakukan, selanjutnya adalah menguji kelas pada data uji dengan persamaan berikut ini.

f(x) = sign (w. 〖 x〗 _1 b)……..(4.10)

Persamaan diatas dilakukan untuk menentukan kelas dengan menghitung bobot dan bias dari data uji yang telah dipakai. Berikut ini perhitungannya.

= sign (-2.243 * -2 + 0.44454)

= sign (-1.79846)

= - 1

Dari hasil perhitungan diatas dapat disimpulkan bahwa kelas untuk nilai -2 pada atribut Style termasuk ke dalam kelas -1 karena memiliki nilai f(x) < 0. Kemudian dilakukan perhitungan untuk menentukan pemisah atau *hyperplane* dengan menggunakan *quadrat programming*.

Untuk hasil klasifikasi dengan algoritma Naive Bayes pada Matlab dengan percobaan cross validation 0.1-0.9 dalam pembagiannya. Hasil cross validation mempengaruhi hasil akurasi dan berikut tabel hasil dari akurasinya berdasarkan cross validation :

Tabel 4.17 Hasil Akurasi Naive Bayes

Pembagian	Akurasi
0.1	**74%**
0.2	59%
0.3	61.3%
0.4	55%
0.5	54.4%
0.6	47.7%
0.7	48.3%
0.8	49.5%
0.9	50.5%

Kemudian berikut hasil akurasi dari klasifkasi data set menggunakan SVM. Pada SVM juga diterapkan validasi silang mulai dari 0.1 hingga 0.9. Hasil dari klasifikasi SVM sebagai berikut :

Tabel 4.19 Hasil Akurasi Naive Bayes

Pembagian	Akurasi
0.1	**66%**
0.2	61%
0.3	60%
0.4	59.5%
0.5	57.6%
0.6	58.7%
0.7	58%
0.8	55%
0.9	51.6%

IV. PEMBAHASAN

Pengunaan algoritma Naïve Bayes digunakan dengan menggunakan Naïve Bayes dengan Gaussian hal ini disebabkan data yang diguankan telah diubah bentuk menjadi data numerik dan bukan data kategori yang perhitungannya dilakukan dengan menggunakan tetapan rata-rata (mean) dan standar deviasi setiap atribut pada setiap atribut yang terdapat pada kedua kelas atau label yang digunakan. Penggunaan Metode SVM dalam melakukan perhitungan manual dilakukan dengan menggunakan eliminasi dan subtitusi yang digunakan untuk mengetahui nila bobot atau W dan juga nilai bias yang akan digunakan untuk menghitung hyperlane. Nilai yang dihasilkan pada perhitungan hyperlane akan menghasilkan kelas pada data uji (testing) yang telah ditentukan.

Hasil penelitian yang telah dilakukan menunjukkan jika terdapatnya perbedaan hasil akurasi yang memiliki tingkat perbedaan yang sangat signifikan hal ini diketahui jika hasil akurasi yang dihasilkan dnegna menggunakan Naïve Bayes pada dataset in *fashion dress* memiliki nilai akurasi yang tinggi yaitu sebesar 74% dan nilai akurasi yang dihasilkan oleh metode yang kedua yang diujikan pada penelitian ini yaitu metode SVM yang diketahui berdasarkan penerapan dan pengujian yang telah dilakukan diketahui jika nilai hasil akurasi yang ditemukan adalah sebesar 66%.

Kecilnya nilai akurasi yang dihasilkan pada metode SVM diketahui disebabkan oleh karena jenis data dan perlakuan khusus pada masing-masing data juga akan menentukan perhitungan dan cara pengeksekusian serta hasil pengujian metode pada data yang digunakan. Penggunaan metode SVM pada penelitian disebabkan oleh karena tingginya tingkat kinerja metode SVM pada proses klasifikasi data yang memiliki dimensi yang tinggi namun berdasarkan hasil penelitian yang dilakukan maka diketahui jika pada dataset yang digunakan pada penelitian ini SVM tidak menghasilkan nilai akurasi yang lebih tinggi dan lebih besar serta menunjukkan jarak hasil akurasi yang ditemukan cukup jauh dari hasil akurasi yang dihasilkan berdasarkan penggunaan metode Naïve Bayes yaitu sebesar 8% hal ini disebabkan oleh karena tidak dilakukannya tahapan normalisasi data yang mempengaruhi nilai akurasi yang dihasilkan oleh perhitungan, penerapan dan pengujian metode SVM.

Berdasarkan penggunaan algoritma Naïve Bayes juga diketahui jika hasil akurasi yang dihasilkan tinggi disebabkan oleh karena algoritma Naïve Bayes melakukan penilaian suatu data (data uji dengan tidak memiliki label/kelas) dengan melakukan perhitungan hasil yang lebih besar. Perhitungan manual pada proses klasifikasi data uji dapat mengubah nilai data uji yang sebenarnya adalah data yang tergolong recommended (1) menjadi data tidak recommended (0) dengan menggunakan teorema bayes pada algoritma Naïve Bayes menggunakan Gaussian. Berdasarkan penelitian ini yang telah dilakukan diketahui juga jika algoritma Naïve Bayes memiliki kelemahan dalam melakukan perhitungan manual untuk menentukan nilai rata-rata dan standar deviasi hal ini dikarenakan algoritma Naïve Nayes menghitung nilai rata-rata (mean) dan nilai standart deviasi sebagai nilai yang lebih dominan dalam perhitungan manual pada saat melakukan klasifikasi. Pada Naïve Bayes dilakukan pengujian terpisah untuk melakukan klasifikasi karena adanya kedua kelas/label hal ini juga mempengaruhi tinggi rendahnya nilai akurasi yang dihasilkan sehingga pada penelitian ini diketahui jika nilai akurasi yang dihasilkan pada algoritma Naïve Bayes tidak terlalu besar yaitu sebesar 74%. Namun diketahui bahwa penelitian ini membuktikan metode klasifikasi Naïve Bayes dan metode SVM dapat mengklasifikasikan dengan benar berdasarkan kelas recommended (1) dan tidak recommended (0) yang telah ditentukan sebelumnya.

V. PENUTUP

A. Kesimpulan

Pada penelitian ini dapat disimpulkan bahwa akurasi yang dihasilkan klasifikasi SVM cenderung lebih rendah dibandingkan SVM. Kesalahan klasifikasi pada SVM dapat terjadi dikarenakan dataset yang digunakan tidak dilakukan normalisasi terlebih dahulu sebelum dilakukan proses klasifikasi SVM. Penggunaan algoritma juga sangat bergantung pada dataset yang digunakan karena dapat mempengaruhi hasil dari tingkat akurasi masing-masing algoritma. Pada Naive Bayes juga terdapat kesalahan klasifikasi dikarenakan tetapan standar deviasi dan rata-rata mempengaruhi proses klasifikasi.

Naive Bayes mendapatkan hasil akurasi lebih tinggi dari SVM karena dalam tahapan klasifikasiannya Naive Bayes memproses satu persatu data atribut. Berbeda dengan SVM yang melakukan klasfikasi secara general sehingga cakupan SVM lebih luas. Hasil yang diperoleh Naive Bayes sebesar 74% sedangkan SVM memiliki nilai akurasi yaitu 66%.

B. Saran

1. Memahami terlebih dahulu dataset yang akan digunakan pada penelitian karena data dapat mempengaruhi hasil akurasi.
2. Memilih metode yang tepat untuk melakukan tahap pre-processing.
3. Pada penelitian selanjutnya diharapkan untuk melakukan percobaan Non-linier SVM.

DAFTAR PUSTAKA

[1] D. Jasim, " Data Mining Approach and Its Application to Dresses Sales Recommendation,", pp. 1-19, 2016.
[2] B. Mirkin, "Data Analysis, Mathematical Statistics, Machine Learning, Data Mining: Similarities and Differences," Int. Conf. Adv. Comput. Sci. Inf
[3] Syst., Vol. 2, pp. 1–8, 2015. M. Ramageri, “Data Mining Techniques and Applications,” Indian J. Comput. Sci. Eng., Vol. 1, No. 4, pp. 301–305, 2010.
[4] Sartika, D. Sensuse, “Perbandingan Algoritma Klasifikasi Naive bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian,” Jurnal Publikasi, Vol. 1, No. 2, pp. 151-161, 2017.
[5] S. James Luke, "Data Mining of Automatically Promotion Tweet for Products and Services Using Naïve Bayes Algorithm to Increase Twitter Engagement Followers at PT. Bobobobo,"
[6] Procedia Computer Science, vol. 59, no.1, pp. 254-261, 2015.
[7] I. Rish, "An Empirical study of The Naïve Bayes Classifier," International Joint Conference on Artificial Intelligence, 2015.
[8] A. Tripathy, "Classification of Sentimental Reviews Using Machine Learning Techniques," Procedia Computer Science, 2015.