Optimasi Skripsi Mahasiswa Teknik Informatika Menggunakan Klasifikasi Algoritma Naive Bayes Dan Support Vector Machine.

HS Sulistyowati, Ir., MM.
Teknik Informatika, STMIK Bani Saleh, hs.sulistyowati@gmail.com

Hudi Kusuma Bharata, ST, MKom
Teknik Informatika, STMIK Bani Saleh,hudi@stmik.banisaleh.ac.id

ABSTRAK

Skripsi merupakan salah satu syarat yang harus dipenuhi oleh mahasiswa untuk menyelesaikan pendidikan di perguruan tinggi . Pada program studi Teknik Informatika STMIK Bani Saleh, penyusunan skripsi oleh mahasiswa dapat dipilih dari 14 (empat belas) topik/kategori yang mewakili ilmu pengembangan teknik informatika. Penelitian ini bertujuan untuk melakukan klasifikasi terhadap skripsi mahasiswa Jurusan Teknik Informatika berdasarkan pada abstrak dalam bahasa Indonesia. Hal ini dilakukan karena abstrak mengandung isi yang merinci dan menjelaskan topik penelitian dari skripsi mahasiswa. Klasifikasi dilakukan untuk mengorganisasikan teks-teks dengan isi sama akan dikelompokkan ke dalam satu kategori tertentu, dan merupakan salah satu klasifikasi dari tex mining. Metode klasifikasi yang digunakan dalam penelitian ini adalah klasifikasi teks dengan algorima *Naïve Bayes Classifier* dan *Support Vector Machine (SVM)*.

Proses klasifikasi diawali dengan Tahap 1: Studi Pustaka, Tahap 2: dilanjutkan dengan pengumpulan Data set, kemudian Tahap 3: Analisa yang meliputi Klasifikasi dan Pelabelan data secara manual, preprocessing data dan ekstraksi fitur serta pengambilan data Training dan Penentuan model. Metode pembobotan kata yang digunakan adalah *Term Frequency- Inverse Document Frequency (TF-IDF)*.Klasifikasi menggunakan metode klasifikasi *Naïve-Bayes Classifier (NBC)*, dan *Support Vector Machine (SVM)* dan pemrograman dengan bahasa phyton serta dengan sampel sebanyak 80 abstrak srikpsi. Dari sampel tersebut, 70 % dijadikan sebagai data training dan 30 % sebagai data testing . Nilai akurasi dari kedua metode tersebut menunjukkan hasil yang berbeda, yaitu akurasi 79,166 % untuk klasifikasi dengan metode NBC dan akurasi 83,33 % dengan metode SVM.

Hasil dari penelitian ini dapat dilakukan untuk melakukan pemetaan klasifikasi topik skripsi yang sudah dikerjakan pada periode sebelumnya, sehingga mahasiswa yang akan mengerjakan skripsi dapat diarahkan pada topik yang belum banyak dikerjakan atau melanjutkan topik tertentu yang sudah pernah dikerjakan. Dengan demikian pemilihan topik skripsi di jurusan Teknik Informatika menjadi lebih beragam, kualitas dan kuantitas skripsi dapat lebih optimal.

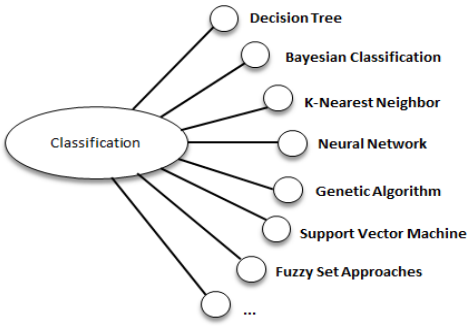
Kata Kunci: Naïve-Bayes Classifier, Support Vector Machine, Nilai Akurasi

1. Pendahuluan

Sebagai tugas akhir untuk memperoleh gelar kesajarnaan Strata-1, mahasiswa wajib membuat skripsi. Pada saat ini banyak judul dan topik skripsi yang sudah dibuat oleh mahasiswa khususnya jurusan Teknik Informatika Sekolag Tinggi Manajemen dan Ilmu Komputer (STMIK) Bani Saleh. Namun dari skripsi tersebut belum terklasifikasi secara maksimal, sesuai dengan bidang peminatan dan pengembangan ilmu informatika yang terus berubah mengikuti perkembangan zaman.Untukitu dalam penelitian ini bertujuan untuk melakukan klasifikasi/pengelompokan topik skripsi mahasiswa yang ada di jurusan Teknik Informatika. Pengelompokan skripsi dilakukan dengan melihat isi abstraksi dari tiap judul skripsi yang telah dibuat oleh mahasiswa. Abstrak dipilih karena dapat mendefinisikan istilah secara lebih ringkas dan jelas pada setiap bidang penelitian skripsi, sehingga dapat digunakan untuk mengklasifikasikan topik skripsi mahasiswa di prodi Teknik Informatika.

Terdapat banyak teknik klasifikasi yang dikenal.*Text Mining* adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dokumen dimana *Text Mining* merupakan variasi data mining yang berusaha menemukan pola menarik dari sekumpulan data tekstual yang berjumlah besar (Kurniawan, Effendi & Sitompul, 2012).

Penggunaan algoritma tertentu tergantung pada jenis input yang tersedia. Kemampuan untuk memahami dan menganalisa penggunaan masing-masing algoritma sangat penting dalam meningkatkan akurasi hasil klasifikasi. (Vanjari, 2013). Beberapa teknik klasifikasi yang dikenal dapat dilihat pada gambar dibawah ini:



Gambar 1. Teknik Klasifikasi

Adapun metode yang digunakan pada penelitian ini adalah metode *Naïve bayes classifier* dan *Support Vector Machine*.Metode *Naïve Bayes Classifier* (NBC), adalah salah satu metode yang dapat mengklasifikasikan teks dan memerlukan algoritma yang sederhana namun memiliki nilai akurasi yang tinggi. Naive *Bayes Classifier* merupakan metode klasifikasi yang berdasar pada teorema *Bayes*.Metode klasifikasi ini cocok digunakan ketika jumlah masukan yang sangat besar.Klasifikasi ini lebih disukai karena kecepatan dan kesederhanaannya. Meskipun klasifikasi ini bisa dibilang klasifikasi yang sederhana, namun hasil yang diperoleh dari klasifikasi ini sering mencapai performa yang sebanding dengan algoritme lain seperti *Decision tree*, dan *Neural Network classifier*. Klasifikasi ini selain memperlihatkan tingginya akurasi juga cepat dalam memproses data dalam jumlah yang besar (C.Aggarwal, 2015).

Secara umum teorema *Bayes* dapat dinotasikan pada persamaan berikut:

$$P(A|B) = \frac{P(A|B).P(A)}{P(B)}$$

Pada *Naïve Bayes Classifier* setiap kalimat direpresentasikan dalam pasangan atribut (a_1, a_2, a_3, … a_n) dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya, sedangkan V adalah himpunan kelas. Pada saat klasifikasi, metode ini akan menghasilkan kelas yang paling tinggi probabilitasnya (V_{MAP}). Berikut rumus V_{MAP} :

$$Vmap = \underset{v_{j\epsilon}V}{argmax}\ P(a_1, a_2, a_3 \dots a_n|V_j)(P(V_j)$$

Naïve Bayes Classifier menyederhanakan hal ini dengan mengasumsikan bahwa dalam setiap kategori, setiap atribut bebas bersyarat satu sama lain. Sebagai berikut:

$$P(a_1, a_2, a_3 \dots a_n|V_j) = \prod_i P(a_i|v_j)$$

Apabila persamaan $Vmap$ disubtitusikan ke persamaan $P(a_1, a_2, a_3 \dots a_n|V_j)$, maka akan menghasilkan persamaan berikut:

$$Vmap = \underset{v_{j\epsilon}V}{argmax}\ P(V_j)\ x \prod_i P\ (a_i|V_j)$$

$P(v_j)$ dan probabilitas kata a_i untuk setiap kategori $P(a_i|v_j)$ dihitung pada saat *training*, yang diruuskan sebagai berikut (Rodiansyah & Winarko, 2013):

$$P(v_j) = \frac{docs_j}{training}$$

$$P(a_i|v_j) = \frac{n_i + 1}{n + kosakata}$$

Keterangan:

- $docs_j$ = jumlah data pada kelas *j*
- *training* = jumlah dokumen yang digunakan dalam proses *training*
- n_i= jumlah kemunculan kata a_i pada kelas v_j
- *n* = jumlah kosakata yang muncul pada kelas v_j
- kosakata = jumlah kata unik pada semua data *training*

Sedangkan metode SVM, adalah seperangkat metode pembelajaran terbimbing yang menganalisis data dan mengenali pola dengan menggunakan analisis regresi.Menurut Santosa, (2007) *Support Vector Machine* (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. SVM memiliki prinsip dasar *linier classifier* yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun SVM telah dikembangkan agar dapat bekerja pada problem non-linier dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi. Pada ruang berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan jarak (margin) antara kelas data. Menurut Santosa (2007) hyperplane klasifikasi linier SVM dinotasikan

$$f(x) = w^T x + b$$

Dan menurut Vapnik dan Cortes (1995) diperoleh persamaan:

$$\left[(w^T x_i) + b\right] \geq 1 \text{ untuk } y_i = +1$$

$$\left[(w^T x_i) + b\right] \leq 1 \text{ untuk } y_i = -1$$

Dengan x_i = himpunan data training I = 1,2,, ……n den y_i = label kelas dari x_i.

Proses klasifikasi menggunakan SVM dimulai dengan mengubah text menjadi data vector. *Vector* dalam penelitian ini memiliki dua komponen yaitu dimensi (word id) dan bobot.Bobot ini sering dikombinasikan ke dalam sebuah nilai TF-IDF (*Term*

Frequency- Inverse Document Frequency), secara sederhana dengan mengalikan keduanya bersama-sama.

Inverse Document Frequency (IDF) adalah jumlah total dokumen atas hitungan yang berisi istilah tersebut. Jadi, jika ada 50 dokumen dalam koleksi, dan dua di antaranya terdapat istilah yang menjadi query, IDF akan menjadi 50 / 2 = 25. Untuk menjadi akurat, kita harus memasukkan query dalam perhitungan IDF, jadi jika dalam koleksi ada 50 dokumen, dan 2 berisi istilah dari query, perhitungan yang sebenarnya akan (50 +1) /(2 +1) = 51 / 3. Diambil log dari IDF untuk memberikan beberapa penghalusan. Jika sebuah istilah A direpresentasikan dalam x buah dokumen, dan istilah B sejumlah 2x kali, maka istilah A adalah istilah yang lebih spesifik yang harus memberikan hasil yang lebih baik, tetapi belum tentu dua kali lebih baik. Kelembutan dari log adalah pemecahan perbedaan-perbedaan ini. Dokumen dapat dinyatakan sebagai list dari term.Kita kemudian menormalisasi tiap komponen dengan panjang dari vector sehingga bobot tersebut dinyatakan dalam 1 unit panjang.

Contoh format data input untuk klasifikasi SVM dalam penelitian ini adalah +1 1:0.049 45:0.0294. Dengan masukan yang pertama +1 atau -1 menyatakan dua kelas (atau 0 untuk data yang akan diklasifikasi). Angka kedua menyatakan dimensi (row id) dan angka ketiga (setelah karakter ":") menyatakan bobot dari term tersebut, tiap term dalam sebuah dokumen dipisahkan dengan spasi.

SVM mencoba untuk menemukan garis yang terbaik membagi dua kelas, dan kemudian mengklasifikasikan dokumen uji berdasarkan di sisi mana dari garis tersebut itu muncul. Intuisi yang mendorong SVM sebagai penentu garis terbaik yang memisahkan kedua kelas adalah yang memiliki margin terbesar diantaranya dan titik pelatihan sampel terdekat di kedua sisinya. Oleh karena itu, *support vector* adalah vektor yang menentukan margin ,dimana vektor ini yang paling dekat dengan garis pemisah.

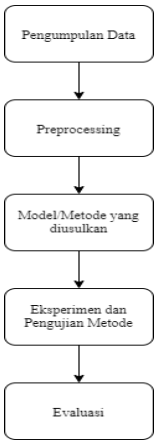
Dari kedua metode yang digunakan untuk penelitian diatas diatas diperlukan data teks berupa abstrak dari skripsi mahasiswa, dan dilakukan pengolahan*preprocessing* dan *stemming*, yaitu melalui tahapan penyeragaman huruf menjadi huruf kecil *(case folding)*, penghapusan tanda baca serta angka *(stop character removal)*, penguraian kalimat menjadi kata *(tokenizing)*, penghapusan kata penghubung *(stopword removal)* dan juga penguraian kata menjadi kata dasar *(stemming)*.

Setelah tahapan *preprocessing* dan *stemming* selesai, akan dilakukan tahapan ekstraksi fitur dan pembobotan. Tahapan ekstraksi fitur yang akan digunakan yaitu fitur *unigram* dan pembobotan kata akan menggunakan *Term Frequency Inverse Document Frequency (TF-IDF)*. Proses selanjutnya yaitu proses klasifikasi dengan membandingkan hasil dari metode *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine (SVM)*, dengan bantuan pemrograman dengan bahasa pemrograman phyton. Hasil klasifikasi dari kedua metode ini selanjutnya akan dibandingkan terhadap klasifikasi topik skripsi dari sampel yang diambil, sehingga diperoleh gambaran topik skripsi mana yang sudah dikerjakan mahasiswa.

2. Metodologi

1. Tahap Penelitian:

Penelitian dilakukan melalui beberapa tahapan yang meliputi langkah - langkah sebagai berikut: pengumpulan data, *preprocessing*/proses pendahuluan, menentukan model, eksperimen dan pengujian model dan evaluasi, seperti dalam gambar berikut:



Gambar 2. Tahap Penelitian

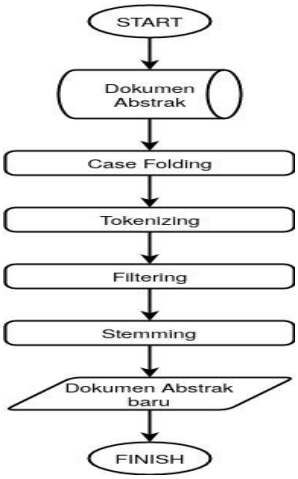
2. Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa softcopy Skripsi mahasiswa prodi Teknik Informatika periode 2016 – 2018, dan diambil abstraksi dari tiap skripsi dalam bentuk bahasa indonesia. Jumlah data abstraksi yang diperlukan adalah 80 (delapan puluh) judul skripsi.Sebagian data dijadikan sebagai data latih *(Training)*sebesar 70 % dan sebagian sebagai data uji *(Testing)*sebesar 30 %. Pengumpulan data menggunakan metode *crawling*, yaitu menyalin halaman untuk diproses oleh *search engine* dan mengindeks halaman yang diunduh sehingga pencarian lebih efisien.

3. Preprocessing Data

Proses *preprocessing* merupakan tahapan awal, yaitu mengurangi atribut yang kurang berpengaruh terhadap proses klasifikasi. Tahapan

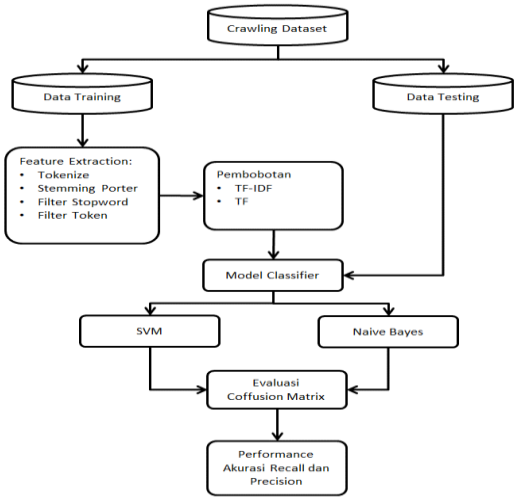
preprocessing, merupakan tahapan pengumpulan data teks sudah di bersihkan dari tag-tag, doc, html dan sejenisnya. *Text processing* bertujuan mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap di gunakan untuk pemrosesan data selanjutnya.Tahapan *preprocessing*dapat dilihat pada gambar berikut:



Gambar 3. Preprosesing Data

4. Model/Metode Klasifikasi.

Metode Klasifikasi dengan dua metode, yaitu algoritma klasifikasi *Naïve Bayes* (NBC) dan *Support Vector Machine* (SVM). Seperti bagan berikut:



Gambar 4. Model Klasifikasi

5. Evaluasi

Evaluasi hasil klasifikasi kedua metode diukur dengan *Confusion matriks*, yaitu matriks yang dibagi dalam empat bagian berisi prediksi kelas dan hasil klasifikasi yang sebenarnya. Evaluasi dilakukan dengan membandingkan nilai akurasi pengujian dari kedua metode tersebut dan dapat diukur dengan rumus sebagai berikut:

Tabel 1.*Confussion Matriks*

		True Class	
		Positive	**Negative**
Predicted Class	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Keterangan

TP :kelas diprediksi positif ,faktanya positif.
TN: kelas diprediksi negatif ,faktanya negatif.
FP: kelas diprediksi positif ,faktanya negatif.
FN: kelas diprediksi negatif , faktanya positif.

Rumus nilai akurasi:

$$Akurasi = \frac{TN + TP}{P + N} = \frac{TN + TP}{TP + FP + TN + FN}$$

Akurasi : 0.90 – 1.00 = Excellent
0.80 – 0.90 = Good
0.70 – 0.80 = Fair
0.60 – 0.70 = Poor
0.50 – 0.60 = Failure

Hasil dan Pembahasan

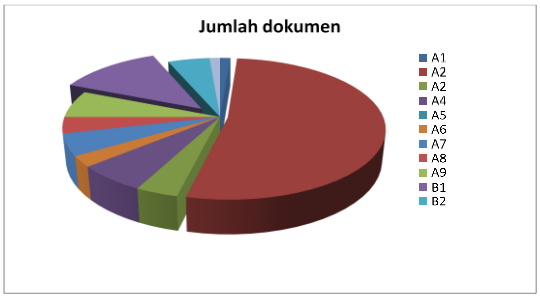
1. Hasil Klasifikasi Manual dan Pelabelan

Secara garis besar terdapat dua dasar pengelompokan skripsi mahasiswa jurusan Teknik Informatika di Sekolah Tinggi Manajemen dan Ilmu Komputer (STMIK) Bani Saleh. Berdasar Buku Panduan Skripsi, kedua kelompok tersebut adalah kelompok Rekayasa Perangkat Lunak (RPL) yang terdiri dari 9 ruang lingkup/topik, serta kelompok Rekayasa Sistem dan Jaringan Komputer yang terdiri atas 5 topik.

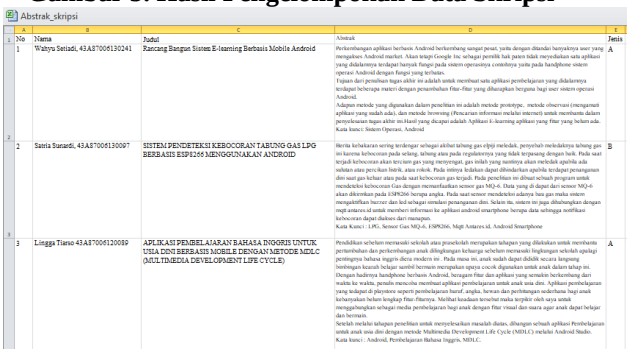
Dari 80 sampel abstrak skripsi mahasiswa teknik Informatika dari tahun 2016 sampai 2018 yang diambil secara acak, diperoleh hasil pengelompokan secara manual sebagai berikut:

a. Kelompok A : Rekayasa Perangkat Lunak (RPL) yang terbagi dalam 9 topik yaitu: (A1) Database, (A2) *Enterprise Systems Design and Development*,(A3) *Enterprise Resources Planning*, (A4) *Data mining*, (A5) *High Performance Computing*, (A6) *Neural Networks and Machine Learning*, (A7) *Computer Graphics and Animation*, (A8) *Software Engineering methodology* dan (A9) *System Integration.*

b. Kelompok B : Rekayasa Sistem dan Jaringan Komputer (JK) adalah : (B1) *Systems Integration, (B2) Networking System Methods, (B3) Multimedia System, (B4) Next Generation Network,* dan (B5) *Sensor Network..*



Gambar 5. Hasil Pengelompokan Data Skripsi



Gambar 6. Pelabelan Data Skripsi

Hasil klasifikasi dan pelabelan data dari 80 sampel abstrak dapat dilihat pada tabel berikut:

Tabel 2. Pengelompokan Data skripsi

Kelompok	Ruang lingkup	Jumlah dokumen	Persen (%)
RPL (A)	A1	1	1.25
	A2	42	52.5
	A2	3	3.75
	A4	5	6.25
	A5	0	0
	A6	2	2.5
	A7	4	5
	A8	3	3.75
	A9	5	6.25
JK (B)	B1	10	12.5
	B2	4	5
	B3	0	0
	B4	1	1.25
	B5	0	0
	Total	80	100

Untuk kelompok RPL, dengan simbol dengan huruf A1, A2, A3, …A9 dan kelompok Jaringa Komputer dengan simbol B1, B2, B3, B4 dan B5, maka jumlah terbesar dengan skripsi yang dikerjakan mahasiswa adalah kelompok A2 (RPL dengan topik *Enterprise Systems Design and Development* sebesar 52.5 %.

Selanjutnya probabilitas mahasiswa mengerjakan skripsi dengan kelompok RPL sebesar 81,25 % atau 0.8125 dan mahasiswa mengerjakan skripsi dengan peminatan Rekayasa System dan Jaringan Komputer sebesar 18.75 % atau 0.1875.

2. Data *Training* dan Data *Testing*

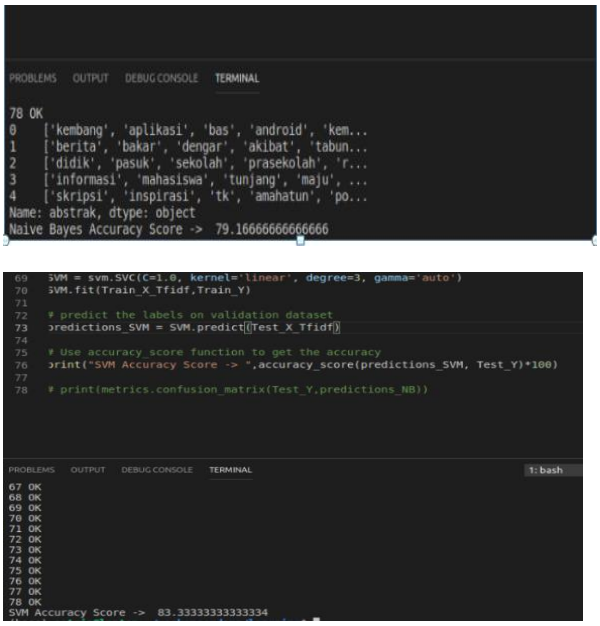
Analisa klasifikasi melalui dua proses, yaitu proses *training* dan proses *testing*,sehingga diperlukan data *training* dan data *testing*. Data *training* digunakan oleh algoritma klasifikasi untuk membentuk sebuah *classifier*, model ini digunakan untuk prediksi kelas data baru yang belum pernah ada,. Data *testing* digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar. Data yang digunakan untuk data *training* dan data *testing* adalah abstrak skripsi berbahasa Indonesia yang telah diberi label secara manual, dengan perbandingan 70% (56 sampel abstrak) untuk data training dan 30% (24 sampel abstrak) untuk data testing.

3. Klasifikasi Naïve Bayes dan Support Vector Machine.

Dokumen abstrak yang sudah melalui tahapan preprocessing dilakukan pengklasifikasian, dalamdengan algoritma Naïve Bayes Classifier (NBC) dan Support Vector Machine, dengan menggunakan bahasa pemrograman Phyton. Karena terbatasnya dokumen yang ada klasifikasi hanya pengolahan klasifikasi dikelompokkan dalam label A= Rekayasa Perangkat Lunak (RPL) dan kelompok B= Rekayasa System dan Jaringan Komputer (JK). Sedangkan untuk tiap kategori dari tiap kelompok RPL dan JK tidak dapat dilakukan, karena kurangnya data dari tiap kategori.Dalam hal ini karena kurangnya skripsi yang cukup memenuhi untuk tiap kategori.

Hasil klasifikasi menunjukkan nilai akurasi sebagai berikut:

Naïve Bayes accuracy score ***: 79,166 %***
SVM Accuracy score ***: 83,333 %***



Gambar 7. Hasil Pemrograman akurasi dengan Phyton

4. Confusion Matriks

Confussion matrik untuk metode SVM diperoleh TP= 20 dan FP = 4, dengan akurasi = 20/24 * 100 % = 83,33 %.

	Positive	Negative
Positive	20	0
Negative	4	0

Sedangkan confussion matrik untuk Naïve Bayes Classifier, TP= 19 dan FP = 6, dengan akurasi sebesar = 19/24* 100 % = 79, 166 %.

	Positive	Negative
Positive	19	0
Negative	5	0

KESIMPULAN

1. Hasil penelitian klasifikasi skripsi pada jurusan Teknik Informatika STMIK Bani Saleh dengan menggunakan pelabelan secara manual menunjukan bahwa skripsi dengan kelompok Rekayasa Perangkat Lunak (RPL) sebesar 81,25 %, sedangkan kelompok Rekayasa System dan Jaringan Komputer sebesar 18.75 %.
2. Hasil klasifikasi dari kedua kelompok skripsi menggunakan algirtma Naïve Bayes memberikan hasil akurasi sebesar 79,166 %, sedangkan tingkat akurasi dengan metode support vector machine memberikan hasil akurasi yang lebih baik , yaitu sebesar 83,33 %.Hal ini menunjukkan bahwa nilai akurasi metode Support Machine memberikan hasil yang lebih baik dari pada metode Naïve Bayes.

DAFTAR PUSTAKA

1. Kurniawan, B., Effendi, S., Sitompu, O.S. 2012. Klasifikasi Konten Berita Dengan Metode Text Mining.Universitas Sumatera Utara, Medan.
2. Liu, B. Sentiment Analysis and Subjectivity. Synthesis Lectures on Human Language Technologies. USA: editor: Graeme Hirst Morgan & Claypool Publishers. 2012.
3. Nur, M. Y., & Santika, D. D. Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine. Konferensi Nasional Sistem dan Informatika, (Vol. 009). Bali. 2011.
4. Prasetyo, Heri. Data Mining Mengolah Data Menjadi Informasi. Yogyakarta: Andi Offset. 2014.
5. Palanisamy, P., Yadav, V., & Elchuri, H. Serendio: Simple and Practical lexicon based approach to Sentiment Analysis. 2013.
6. Triawati, C. 2009. Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia.Institut Teknologi Telkom, Bandung.
7. Vanjari, S.P., & Thombre, V.D., Classification Techniques: A Survey, International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, 2013.