PENERAPAN TEKNIK DATA MINING UNTUK MENGELOMPOKKAN E-MAIL

Ratih Puspasari[1)]

[1)] *Manajemen Informatika Universitas Potensi Utama*
Jl .K.L. Yos Sudarso Km. 6,5 No. 3A Tanjung Mulia Medan
Email : puspasariratih21@yahoo.com[1)]

ABSTRAK

Penelitian ini dilakukan di Universitas Potensi Utama Medan dengan menitik beratkan kepada proses pengelompokan *email* dengan menggunakan teknik *Artificial Intelligent Rough Set* yang bertujuan untuk pengambilan keputusan apakah *email* yang ada di STMIK Potensi Utama Medan aman atau tidak dari virus atau ancaman dari luar. Data di ambil dari email di bagian LPPM Universitas Potensi Utama. Data yang telah dikumpulkan kemudian dianalisis dan dipelajari serta dirumuskan sehingga menghasilkan sistem pengambilan keputusan. Sistem pengambilan keputusan tersebut dibantu dengan teknik *Artificial Intelligent Rought Set.* Dari hasil penelitian ini ditemukan bahwa teknik *Artificial Intelligent Rought Set* merupakan sebuah teknik yang dapat diandalkan untuk pengambilan keputusan karena dapat menganalisis data dalam skala besar serta dapat membantu pengambilan keputusan yang tepat waktu.

Kata kunci : *data mining, email, artificial intelligent, roughtsSet, pengambilan keputusan.*

1. Pendahuluan

Saat ini email mempunyai peranan yang sangat besar dalam berkomunikasi dengan lebih cepat dan lebih murah dibandingkan dengan metoda komunikasi yang lebih tradisional. Oleh sebab itu email merupakan media komunikasi yang paling banyak digunakan, baik untuk kebutuhan perorangan maupun untuk kebutuhan organisasi, bahkan alamat email sudah merupakan atribut dari identitas seseorang ataupun organisasi.

Dengan relasi yang semakin luas memungkinkan si pemilik alamat email menerima banyak sekali email setiap hari dan terus bertambah tanpa kita sempat membacanya satu persatu. Padahal dari email-email tersebut memungkinkan banyak sekali informasi-informasi berharga yang harus kita ketahui dan banyak pula email-email yang harus hati-hati kita sikapi misalnya email-email yang berisi virus, email yang berisi ancaman atau serangan atau email-email yang berisi rahasia berharga yang tidak bisa diukur nilai kerugiannya dengan uang apabila informasi ini bocor. Bahkan mungkin saja kita pernah menerima email yang tidak perlu dibaca dan hanya memenuhi database email yang kita miliki karena email-email tersebut dikirimkan oleh *spammer.*

Untuk mendapatkan informasi-informasi berharga yang tersembunyi itu diperlukan suatu cara agar dapat menemukan informasi berharga yang tersembunyi dalam tumpukan email yang ada dalam database. Penelitian ini menitik beratkan pada isi dari E-mail yang di terima pada Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) STMIK Potensi Utama. Adapun isi E-mail yang diterima yaitu :

- E-mail yang berisi spam contohnya email yang berupa iklan
- E-mail yang berisi rahasia contohnya email yang berupa pengiriman dan penerimaan bahan makalah.
- E-mail bersifat biasa contonya email yang berupa informasi dari luar contoh undangan seminar.
- E-mail yang bersifat pribadi

Adapun identifikasi masalah adalah sebagai berikut :

1. Bagaimana menerapkan *Rough Set* untuk mengetahui jumlah pesan-pesan yang dikategorikan sebagai *spam* atau bukan *spam* ?.
2. Bagaimana mengevaluasi kinerja dari *Rough Set* dalam memfilter *email spam*?

Adapun tujuan penelitian ini dilakukan adalah :

1. Untuk mengetahui jumlah e-mail yang masuk berdasarkan kategori *spam* atau bukan *spam.*
2. Sebagai alat penunjang keputusan apakah e-mail LPPM aman atau tidak.

2. Pembahasan

Electronic mail adalah salah satu sarana komunikasi yang cukup handal, perbandingannya dengan mail adalah waktu pengirimannya yang sangat cepat. Electronic mail atau disingkat e-mail bukanlah pelayanan "end to end", karena mesin pengirim dan penerima tidak perlu berkomunikasi secara langsung. Proses penyampaian electronic mail dapat dianalogikan dengan penyampaian surat oleh
Kantor Pos dan Giro. Proses ini disebut "*store and forward*".

Sistem kerja E-mail yang banyak digunakan saat ini banyak menggunakan Unix E-mail model. Model ini mampu untuk mengirim dan menerima E-mail baik di lokal perusahaan (intranet) maupun juga dalam lingkungan global (internet). Unix E-mail Model membagi dalam 3 fungsi yaitu :

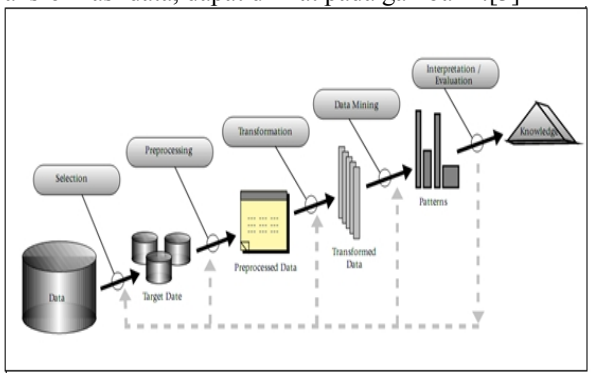
1. *Message Transfer Agent* (*MTA)*
2. *Message Delivery Agent (MDA)*
3. *Message User Agent (MUA)*

MTA, MDA dan MUA beserta E-mail database

disimpan dalam sebuah server E-mail. Server ini dapat diakses secara lokal maupun via internet.[1]

Data mining merupakan proses yang menggunakan teknik statistic, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Algoritma asosiasi merupakan suatu bentuk algoritma dalam data mining yang memberikan informasi hubungan antar item data di database.[2]

Data mining merupakan bagian dari proses yang disebut penemuan KDD-pengetahuan dalam database. Proses ini pada dasarnya terdiri langkah-langkah yang dilakukan sebelum melakukan data mining,seperti pemilihan data,pembersihan data, pra-pengolahan,dan transformasi data, dapat dilihat pada gambar 1.[3]



Gambar 1. Langkah-langkah KDD

Ada beberapa langkah untuk penyelesaian dengan memanfaatkan Metode Rough Set yaitu: Information System (IS). Dalam rough set, sebuah set data di representasikan sebagai sebuah tabel, dimana baris dalam tabel merepresentasikan objek dan kolom-kolom merepresentasikan atribut dari objek-objek tersebut.Tabel tersebut disebut dengan information system yang dapat digambarkan sebagai di mana U adalah set terhingga yang tidak kosong dari objek yang disebut dengan universe dan A setter hingga tidak kosong dari atribut dimana:.

$$IS = \{U,A\} \quad \ldots\ldots\ldots (1)$$

Untuk tiap $a \in A$. Set *Va* disebut value set dari a.
U={e1, e2,…, em}merupakan sekumpulan example dan A={a1, a2, …, an}yang merupakan attribute kondisi secara berurutan.[3] Dari persamaan (1) didapt sebuah Information Systems yang sederhana diberikan dalam tabel 1

Tabel 1. *Information System*

Attribute	*Subject*	*Content Type*	*Dec*
E1	Penelitian	text	Penting
E2	Tugas	text	Pribadi
E3	Iklan	Tag HTML	Spam
E4	Penelitian	Tag HTML	Spam
E5	Iklan	Text	Spam
E6	Tugas	text	Pribadi
E7	Penelitian	text	Penting
E8	Penelitian	text	Penting
E9	Iklan	Tag HTML	Spam
E10	Iklan	Tag HTML	Spam

Tabel 1 memperlihatkan sebuah Information Systems yang sederhana. Dalam Information System, tiap-tiap baris merepresentasikan objek sedangkan column merepresentasikan attribute. Ianya terdiri dari m objek, seperti E1, E2,…, Em,

Dari permasalahan yang timbul yaitu penulis melakukan perhitungan manual bagaimana teknik *rough set* pada proses data untuk mendapatkan pola penentuan berapa jumlah email yang *spam* dan bukan *spam.* Dari hasil pengumpulan data dapat diinformasikan sebagai berikut :

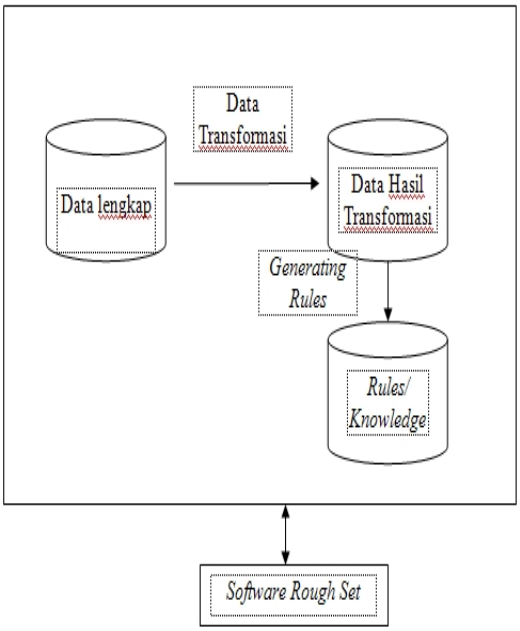
1. Menghitung jumlah email dilakukan berdasarkan pengelompokkan komponen email.
2. Pengelompokkan komponen email dan ketentuannya, komponen email dikelompokkan dalam 3 kelompok yaitu :
   a. Spam
   b. Pribadi
   c. Penting

Pengelompokkan komponen seperti dapat dilihat pada tabel 2.

Tabel 2. *Pengelompokkan Komponen*

No	Subjek	Content Type	Jenis
1	Penelitian	text	Penting
2	Tugas	text	Pribadi
3	Iklan	Tag HTML	Spam

Arsitektur dari sistem pendukung keputusan dalam analisa pada pengelompokkan email terdapat pada rancangan umum dapat dilihat pada gambar 1.



Gambar 2. *Rancangan umum SPK dalam proses pengelompokkan email*

Pada gambar 2 memperlihatkan rancangan umum dari sistem pendukung keputusan dalam pengelompokkan email yang terdiri dari tiga bagian utama yaitu *database*, Model dan Dialog.

Database yang akan diolah bersumber dari data hasil pengelompokkan seperti terlihat pada tabel 3.

Tabel 3. *Database Hasil Pengelompokkan Email*

No	Subjek	Content Type	Jenis (D)
1	Penelitian	text	Penting
2	Tugas	text	Pribadi
3	Iklan	Tag HTML	Spam
4	Penelitian	Tag HTML	Spam
5	Iklan	Text	Spam
6	Tugas	text	Pribadi
7	Penelitian	text	Penting
8	Penelitian	text	Penting
9	Iklan	Tag HTML	Spam
10	Iklan	Tag HTML	Spam

Dari hasil pengelompokan data diatas maka dilakukan *Indiscernibility* yaitu mengelompokkan sekumpulan *object* yang mempunyai nilai *decision* yang sama, seperti pada tabel 4.

Tabel 4. *Indiscernibility*

Class	*Subjek*	*Content Type*	*Dec*	*Indiscernibility Realtion*
EC1	Penelitian	text	Penting	E5, E7, E8
EC2	Tugas	text	Pribadi	E9, E10
EC3	Iklan	Tag HTML	Spam	E1, E3, E4
EC4	Iklan	Text	Spam	E2
EC5	Penelitian	Tag HTML	Spam	E6

Setelah melakukan *Indiscernibility* maka akan dilakukan *Equivalence Class* untuk mengelompokkan sekumpulan *object* yang mempunyai nilai atribut yang sama. Hasil dari *equivalence class* dari tabel 5.

Tabel 5. *Equivalence Class*

Class	Subjek	Content Type	Dec	num_obj
EC1	Penelitian	text	Penting	3
EC2	Tugas	text	Pribadi	2
EC3	Iklan	Tag HTML	Spam	3
EC4	Iklan	Text	Spam	1
EC5	Penelitian	Tag HTML	Spam	1

Langkah selanjutnya melakukan penyederhanaan *equivalence class* pada tabel 5 ke representasi *numeric* yaitu dengan cara mengelompokkan data seperti pada tabel 6.

Tabel 6. *Representasi Numeric*

Subject	*Content Type*	*Dec*
Iklan = 1	Tag HTML = 4	Penting = 6
Penelitian = 2	Text = 5	Pribadi = 7
Tugas = 3		Spam = 8

Dari pengelompokkan representasi *numeric* diatas maka didapatlah sebuah tabel *equivalence class* seperti yang terlihat pada tabel 7.

Tabel 7. *Representasi Numeric Equivalence Class*

Class	*Subject*	*Content Type*	*Dec*	*Num_obj*
EC1	2	5	6	3
EC2	3	5	7	1
EC3	1	4	8	3
EC4	1	5	8	1
EC5	2	4	8	2

Setelah itu dilakukan *Discernibility Matrix* yang bertujuan untuk pengelompokkan sejumlah atribut dimana yang dikelompokkan hanyalah *atribut conditional* saja. Contoh *Discernibility Matrix* dapat dilihat pada tabel 8.

Tabel 8. *Descernibility Matrix*

	EC1	EC2	EC3	EC4	EC5
EC1	X	A	AB	A	B
EC2	A	X	AB	A	AB
EC3	AB	AB	X	B	A
EC4	A	A	B	X	AB
EC5	B	AB	A	AB	X

Setelah dilakukan *Discernibility Matrix* langkah selanjutnya adalah *Discernibility Matrix Modulo D* untuk mengelompokkan sejumlah *atribut conditional* dan berbeda pula *decisionnya*. Contoh *Discernibility Matrix Modulo D* dapat dilihat pada tabel 9.

Tabel 9. *Discernibility Matrix Modulo D*

	EC1	EC2	EC3	EC4	EC5
EC1	X	A	AB	A	B
EC2	A	X	AB	AB	AB
EC3	AB	AB	X	X	X
EC4	A	A	X	X	X
EC5	B	AB	X	X	X

Dari hasil *Discernibility Matrix Modulo D* maka dilakukan proses *Reduct Calculation* dengan menggunakan *prime impilcant fungsi Boolean.* Kumpulan dari semua *prime implicant* mendeterminasikan *sets of reduct. Reduct calculation* yang dihasilkan dapat dilihat pada tabel 10.

Tabel 10. *Reduct*

Class	*CNF of Boolean Function*	*Prime Implicant*	*Reducts*
E1	A^(A v B)^A^B	A^B	{A,B}
E2	A^(A v B)^(A v B)^(A v B)	A	{A}
E3	(A v B)^(A v B)	A v B	{A,B}
E4	A^A	A	{A}
E5	B^(A v B)	B	{B}

Setelah didapatkan hasil dari *Reduct calculation* langkah terakhir adalah *Generating Rules* untuk menghasilkan *rules/knowledge* berdasarkan *equivalence class* dan *reduct.* ini akan menghasilkan *rules/knowledge* yang dapat digunakan dalam sebuah pengambilan keputusan. *Knowledge* yang digali dari *equivalence class* berdasarkan hasil *reduct* pada tabel 10 maka *rules* yang terbentuk adalah seperti pada tabel 11.

Tabel 11. *Contoh Generating Rules*

Class	A	B	Dec	*Reducts*
EC1	2	5	6	{A,B}
EC2	3	5	7	{A}
EC3	1	4	8	{A,B}
EC4	1	5	8	{A}
EC5	2	4	8	{B}

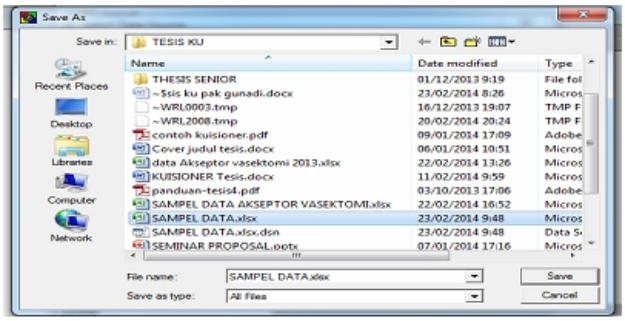
Jadi *rule* yang terbentuk dari *reduct* yang telah dilakukan adalah sebagai berikut :

1. a2b5 → d6
2. a3 → d7
3. a1b4 → d8
4. a1 → d8
5. b2 → d8

Rule yang dihasilkan adalah sebagai berikut :

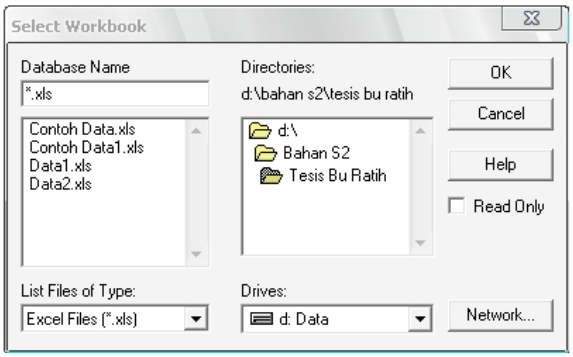
1. If A = 2, or B = 5 then Dec = 6
   Jika A = "Penelitian" , atau B = "Text" maka keputusannya "Penting"
2. If A = 3 then Dec = 7
   Jika A = "Tugas" maka keputusannya "Pribadi"
3. If A = 1, or B = 4 then Dec = 8
   Jika A = "Iklan", atau B = "Tag HTML" maka keputusannya "*Spam*"
4. If A = 1 then Dec = 8
   Jika A = "Iklan" maka keputusannya "*Spam*"
5. If B = 4 then Dec = 8
   Jika A = "Tag HTML" maka keputusannya "*Spam*"

Setelah menganalisa sistem maka dilakukan pengujian terhadap analisa teknik *rough set* yang dilakukan secara manual, pengujian selanjutnya dilakukan dengan *tool rosetta*, dan diselaraskan dengan pembuktian dari analisa metode terhadap permasalahan yang ada pada bab sebelumnya.



Gambar 3. *Directori Penyimpanan Data Source*

Gambar 3 menggambarkan tentang lokasi penyimpanan File Rosetta. Setelah melakukan proses penyimpanan.



Gambar 4. *Pemilihan File Decision System*

Pada penelitian ini, penulis menggunakan"Contoh Data1.xls" sebagai data processing. Untuk tahapan selanjutnya mencardi mana letakdata Source yang teladi Create pada tahapan sebelumnya.

Gambar 5. *Sheet of Decision System*

dimana data yang diolah adalah sampel data yang diambil dari hasil pengelompokkan email dapat dilihat pada gambar 6.

Decision System

	Class	Subjek	Content Type	Dec
1	C1	2	5	6
2	C2	3	5	7
3	C3	1	4	8
4	C4	2	4	8
5	C5	1	5	8
6	C6	3	5	7
7	C7	2	5	6
8	C8	2	5	6
9	C9	1	4	8
10	C10	1	4	8

Gambar 6. *Tampilan Decision System*

Hasil *reduct* menggunakan *tools rosetta* dapat dilihat pada gambar 7.

No name

	Reduct	Support	Length
1	{Class}	1	1
2	{Subjek, Content Type}	1	2

Gambar 7. *Tampilan Reduct*

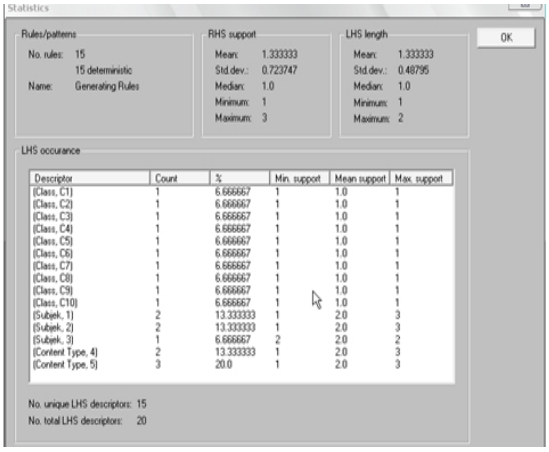
Hasil *Generating Rules* menggunakan *tools rosetta* dapat dilihat pada gambar 8.

Generating Rules

	Rule	LHS Support	RHS Support	RHS Accuracy	LHS Coverage	RHS Coverage	RHS Stability	LHS Length	RHS Length
1	Class(C1) => Dec(6)	1	1	1.0	0.1	0.333333	1.0	1	1
2	Class(C2) => Dec(7)	1	1	1.0	0.1	0.5	1.0	1	1
3	Class(C3) => Dec(8)	1	1	1.0	0.1	0.2	1.0	1	1
4	Class(C4) => Dec(8)	1	1	1.0	0.1	0.2	1.0	1	1
5	Class(C5) => Dec(8)	1	1	1.0	0.1	0.2	1.0	1	1
6	Class(C6) => Dec(7)	1	1	1.0	0.1	0.5	1.0	1	1
7	Class(C7) => Dec(6)	1	1	1.0	0.1	0.333333	1.0	1	1
8	Class(C8) => Dec(6)	1	1	1.0	0.1	0.333333	1.0	1	1
9	Class(C9) => Dec(8)	1	1	1.0	0.1	0.2	1.0	1	1
10	Class(C10) => Dec(8)	1	1	1.0	0.1	0.2	1.0	1	1
11	Subjek(2) AND Content Type(5) => Dec(6)	3	3	1.0	0.3	1.0	1.0	2	1
12	Subjek(3) AND Content Type(5) => Dec(7)	2	2	1.0	0.2	1.0	1.0	2	1
13	Subjek(1) AND Content Type(4) => Dec(8)	3	3	1.0	0.3	0.6	1.0	2	1
14	Subjek(2) AND Content Type(4) => Dec(8)	1	1	1.0	0.1	0.2	1.0	2	1
15	Subjek(1) AND Content Type(5) => Dec(8)	1	1	1.0	0.1	0.2	1.0	2	1

Gambar 8. *Tampilan Generating Rules*

Tampilan *Statistic* yang memberikan informasi berapa jumlah *Rules/patterns,* RHS *support,* LHS *length* dan LHS *occurance.* Seperti gambar 9.



Gambar 9. *Tampilan Statistic*

Dari pengolahan data yang dilakukan dengan menggunakan program *Rosseta* maka didapatkan hasil sebagai berikut :

1. *Rules* sebanyak 15 (15 *deterministic*)
2. *Righ hand set (RHS Support)*
   a. *Mean* : 1.333333
   b. *Std.dev* : 0.48795
   c. *Median* : 1.0
   d. *Minimum* : 1
   e. *Maximum* : 3
3. *Left hand set (LHS Length)*
   a. *Mean* : 1.333333
   b. *Std.dev* : 0.232182
   c. *Median* : 1.0
   d. *Minimum* : 1
   e. *Maximum* : 2
4. *Unique LHS descriptors* : 15
5. *Total LHS descriptors* : 20

Dari dua hasil pengujian yang telah dilakukan yaitu proses secara manual dan menggunakan *software Rosetta* dapat kita ambil sebuah kesimpulan bahwa hasil pengujian sangat baik karena *rule* yang dihasilkan hampir sama dan jumlah email yang dihasilkan yaitu :

a. Untuk kategori ”Penting” sebanyak 1 item
b. Untuk kategori ”Pribadi” sebanyak 1 item
c. Untuk kategori ”*Spam*” sebanyak 3 item

3. Kesimpulan

Hasil pengolahan dan analisa data yang dilakukan dalam penelitian ini. Terdapat 3 *class* yaitu “Pribadi”, ”Penting” dan *“Spam”,* sehingga *rule-rule* yang dihasilkan oleh proses *Data mining* dengan menggunakan program *Rough Set* dapat membantu dalam pengembangan sistem pendukung keputusan untuk pengelompokan email yang ada di Universitas Potensi Utama.

Daftar Pustaka

[1] Christina Chung “Applying Data Mining to Data Security”
[2] Emha Taufik, Luthfi, 2009, Penerapan Data Mining Algoritma Asosiasi Untuk Meningkatkan Penjualan, jurnal, Yogyakarta, STMIK AMIKOM Yogyakarta
[3] Dr. H. SarjonDefit, S.Kom,MSc. (2012).“RoughtSetTheoryAnd DataMining.”Modul.

ISSN : 2302-3805

Biodata Penulis

Ratih Puspasari, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi STMIK Potensi Utama, lulus tahun 2007. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Magister Komputer UPI YPTK Padang, lulus tahun 2010.Saat ini menjadi Dosen di Universitas Potensi Utama.