E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 02, Nomor 02, Desember 2017 : 94 - 98

DETEKSI *E-MAIL* DAN *SPAM* MENGGUNAKAN *FUZZY ASSOCIATION RULE MINING*

Fahrur Rozi[1)], Rikie Kartadie[2)]
[1) 2)]Jurusan Pendidikan Teknologi Informasi, STKIP PGRI Tulungagung
Jl Mayor Sujadi Timur no.7. Tulungagung
e-mail: rozi.fahrur04@gmail.com [1)], rikie.kartadie@gmail.com[2)]

ABSTRAK

Munculnya suatu e-mail komersial yang tidak diharapkan atau yang lebih sering disebut dengan spam sangat menggangu pengguna e-mail karena dapat menambah penggunaan bandwith koneksi internet, serta akan menjadi suatu sampah yang menumpuk sehingga mengurangi kapasitas penyimpanan. Berdasarkan permasalahan tersebut penelitian bertujuan mengembangkan suatu metode hybird yang menggabungkan antara logika fuzzy dan association rule untuk mendeteksi antara e-mail dan spam. Sehingga dengan adanya metode hybird ini diharapkan e-mail yang diterima dapat diseleksi seakurat mungkin terhadap munculnya spam. Penelitian yang berjudul Deteksi E-mail dan Spam dengan menggunakan fuzzy association rule mining ini akan terdiri dari beberapa tahap yaitu preprocsessing e-mail dan spam, ekstraksi kandidat cluster, dan pembentukan cluster tree. Tahap preprocessing merupakan tahap dimana e-mail dan spam akan diambil kata kuncinya dengan melakukan penghilangan stopwords, stemming, dan seleksi term. Tahap ekstraksi kandidat cluster terdiri dari beberapa tahap lagi yaitu, penentuan fungsi keanggotaan fuzzy dan pembentukan kandidat cluster dengan association rule. Selanjutnya adalah tahap pembentukan cluster tree yang merupakan tahap pendeteksian e-mail dan spam dengan cara mengelompokkannya ke cluster yang sesuai.

***Kata Kunci**: Artificial Neural Network, Cuaca, Fuzzy Inference System, Moving Average, Prediksi*

ABSTRACT

The emergence of an unsolicited commercial e-mail or more often called spam is very annoying to e-mail users because it can increase the bandwidth usage of internet connection, and will become a garbage that accumulate, thus reducing the storage capacity. Based on the problem, the research aims to develop a hybird method that combines fuzzy logic and association rule to detect between e-mail and spam. So with the hybird method is expected to receive e-mail can be selected as accurately as possible against the emergence of spam. The research entitled E-mail and Spam Detection using fuzzy association rule mining will consist of several stages of preprocessing e-mail and spam, cluster candidate extraction, and cluster tree formation. The preprocessing stage is the stage where e-mail and spam keywords will be retrieved by performing stopwords, stemming, and term selection. The cluster candidate extraction stage consists of several stages, namely, determination of fuzzy membership function and cluster candidate formation with association rule. Next is the stage of cluster tree formation which is the stage of detecting e-mail and spam by grouping it to the appropriate cluster.

***Keywords**: Artificial Neural Network, Fuzzy Inference System, Moving Average, Prediction, Weather.*

I. PENDAHULUAN

Penggunaan media elektronik dalam penyampaian surat – menyurat atau yang sering disebut e-mail sudah tidak asing lagi dalam penerapannya dikehidupan sehari – hari. Penggunaan e-mail sebagai alat komunikasi tidak terlepas dari murah dan cepat dalam penyampaiannya ke penerima. Namun dengan semakin banyaknya penggunaan e-mail semakin banyak pula penyalahgunaan e-mail dalam bidang komunikasi. Salah satunya adalah munculnya suatu e-mail komersial yang tidak diharapkan atau yang lebih sering disebut dengan spam. Munculnya spam sangat menggangu pengguna e-mail karena dapat menambah penggunaan bandwith koneksi internet, serta akan menjadi suatu sampah yang menumpuk sehingga mengurangi kapasitas penyimpanan. Beberapa spam digunakan untuk menyampaikan suatu pesan iklan yang berisi suatu konten pornografi atau suatu media penyebar virus. Untuk itu dibutuhkan suatu media yang dapat mendeteksi dan menyaring spam, sehingga dapat memisahkan antara spam dan e-mail.

Penelitian untuk mendeteksi keberadaan spam telah banyak dikembangkan. Beberapa diantaranya adalah penelitian mengenai spam campaign detection [1], Appavu dkk membagi spam email menjadi beberapa campaign yang kemudian memberikan skor sesuai kriteria dari masing – masing campaign. Penelitian dengan menggunakan fuzzy yang terintegrasi dengan Wordnet telah dikembangkan [2]. Penggunaan fuzzy mampu memberikan solusi ketika suatu term penting yang dapat menjadi kata kunci dalam deteksi email dan spam jarang muncul dalam suatu koleksi e-mail dan spam dan memiliki frekuensi yang kecil, sehingga dengan fuzzy term

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 02, Nomor 02, Desember 2017 : 94 - 98

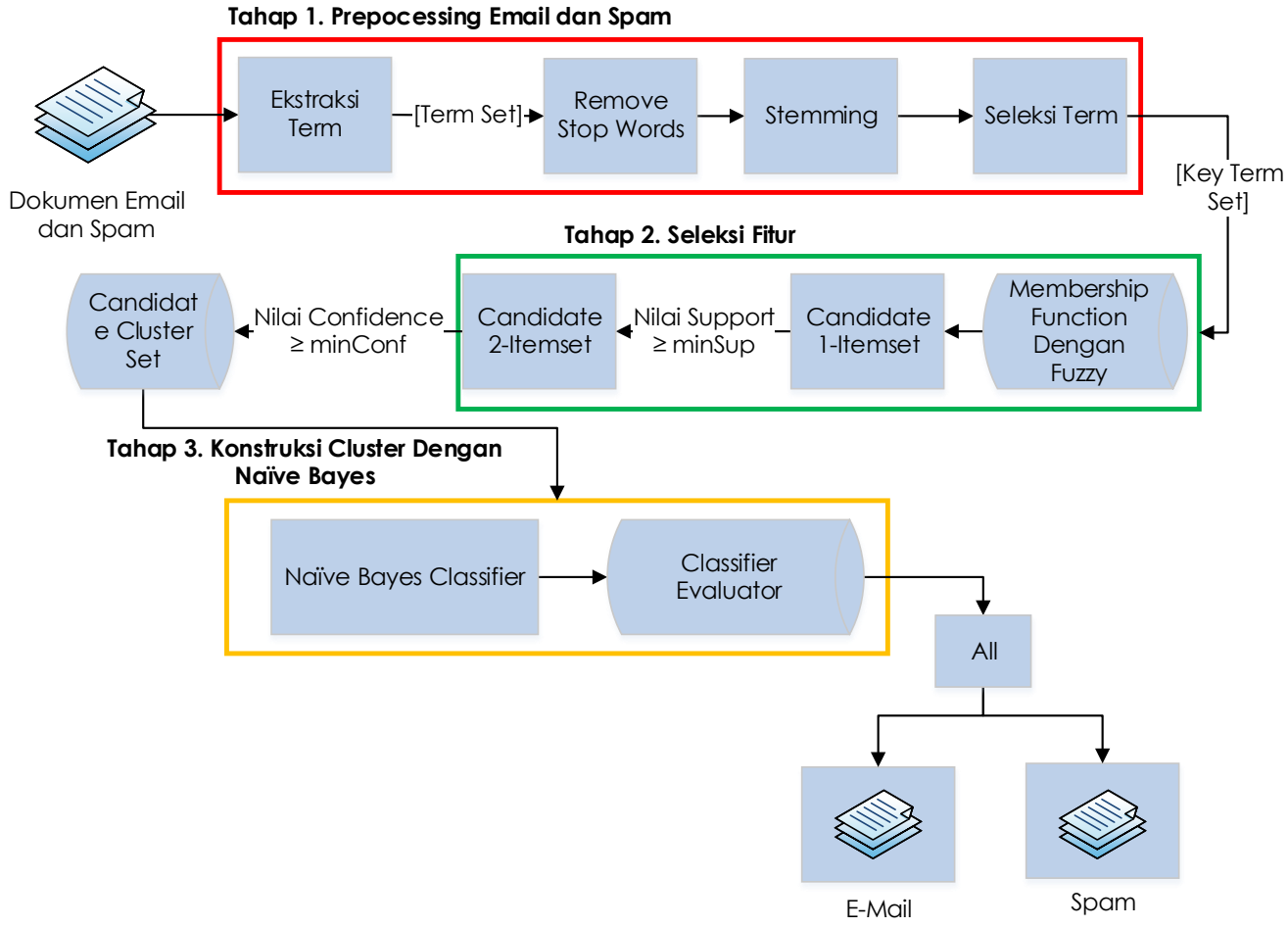
dibangkitkan untuk digunakan sebagai kata kunci. Selain dengan menggunakan logika fuzzy clustering dengan menggunakan association rule mining [3]. Penelitian [3] membangkitkan terlebih dahulu sebuah rule untuk dapat mendeteksi antara e-mail dan spam.

Kelebihan yang dimiliki association rule mining dalam clustering e-mail dan spam untuk menentukan hubungan yang terjadi antar term, serta kelebihan dari fuzzy dalam clustering e-mail dan spam dengan mengintegrasikan variabel linguistic term ke dalam fuzzy set, dapat dikembangkan dengan mengkombinasikan antara fuzzy dengan association rule mining. Penggabungan antara fuzzy dan association rule mining [4] yaitu Fuzzy Frequent Itemset-Based Hierarchical Clustering (F2IHC) mampu meningkatkan tingkat akurasi serta menghasilkan cluster yang overlapping dalam clustering dokumen. Penelitian yang lain dengan menggunakan fuzzy association rule mining juga dilakukan oleh [5], dalam penelitiannya Rozi dkk, menggunakan fuzzy tipe-2 dan association rule mining dengan mengintegrasikan terhadap Wordnet untuk pengelompokan dokumen. Penggunaan Adapative Neuro Fuzzy Inference System juga digunakan untuk peramalan cuaca [6][7], metode yang digunakan menggabungkan antara ANFIS dengan metode siklis dan moving average.

Berdasarkan kelebihan yang dimiliki logika fuzzy dan association rule mining dalam mendeteksi serta mengelompokkan suatu e-mail dan spam dengan menentukan hubungan yang terjadi antar term melalui association rule serta mengintegrasikan variabel linguistic term ke dalam fuzzy set, maka dapat dikembangkan suatu metode hybird antara logika fuzzy and association rule mining untuk deteksi e-mail dan spam.

II. METODE

Rancangan sistem dalam penelitian ini terdiri atas tiga bagian utama yaitu : *preprocessing e-mail* dan *spam* , seleksi fitur, konstruksi cluster dengan naïve bayes. Gambaran umum arsistektur sistem ditunjukkan pada Gambar 1.



Gambar 1. Arsistektur Sistem

a. Prepocessing Email dan Spam

Pada tahap *preprocessing* terdapat *input* berupa koleksi *e-mail* dan *spam* D, daftar kata stop – word,

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 02, Nomor 02, Desember 2017 : 94 - 98

minimum tf – idf (α). Tahap yang dilakukan dalam *preprocessing*, yaitu : ektraksi *term*, penghilangan *stopwords*, *stemming*, dan seleksi *term*. Dari hasil *preprocessing* akan didapatkan koleksi *keyterm*s serta frekuensinya

b. Seleksi Fitur
Pada tahap seleksi fitur terdapat empat jenis input yang digunakan, yaitu : koleksi *key terms*, fungsi keanggotaan *fuzzy*, *minimum support* , *minimum confidence*. Terdapat empat proses yang harus dilalui untuk mendapatkan kandidat *cluster*, diantaranya : menghitung nilai membership function dengan *fuzzy* set tipe-2, menemukan *candidate*-1 *itemset*, menemukan *candidate*-2 *itemset*, dan seleksi kandidat *cluster.*

c. Konstruksi Cluster dengan Naïve Bayes
Pada tahap ini terdapat 2 jenis sub bagian utama, yaitu : Naïve Bayes Classifier dan Classifier Evaluator. Pada tahap Naïve Bayes Classifier akan dilakukan mengenai perhitungan probabilistik berdasarkan kata kunci yang didapatkan dari seleksi fitur. Sementara pada tahap Classifier Evaluator adalah penentuan kecenderungan dari e-mail apakah termasuk SPAM atau non SPAM.

III. HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan mengenai hasil uji coba serta evaluasi dari metode yang diusulkan dalam penelitian ini. Metode dalam penelitian ini diaplikasikan dengan didukung oleh hardware dan software dengan spesifikasi Processor Intel® Core™2 Duo T5750@2.00Ghz, memori 1014 MB, sistem operasi Windows 7, dan menggunakan Java Netbeans 6.9.1 dengan jdk1.6.0_18.

A. Dataset

Penelitian ini menggunakan 3 jenis dataset. Penjelasan mengenai dataset tersebut dijelaskan sebagai berikut :
Dataset training Non SPAM : merupakan koleksi dataset non spam yang berjumlah 400 data yang digunakan untuk mendapatkan keyterm dari file non spam.
Dataset training SPAM : merupakan koleksi dataset spam yang berjumlah 100 data yang digunakan untuk mendapatkan keyterm dari file spam.
Dataset uji : merupakan koleksi dataset yang digunakan untuk melakukan pengujian terhadap file yang masuk. Data uji ini terdiri dari 200 data yang terdiri dari 160 data Non SPAM dan 40 dataset SPAM.

B. Pengujian

Setiap pengujian yang dilakukan akan dicari nilai accuracy, precision, recall, dan F-Measure dengan menggunakan data uji sejumlah 160 data non spam dengan 40 data spam.

1. Pengujian dengan Data Uji 1

Pada pengujian ini akan dilakukan pencarian nilai accuracy, precision, recall, dan F-Measure dengan menggunakan 100 data email non spam dan 25 data email spam berdasarkan keyword dari ekstraksi fitur kata kunci dari data SPAM yang sama.
Berdasarkan pengujian yang dilakukan dengan menggunakan 500 kata kunci dari HAM (Non SPAM) dan SPAM dari hasil ekstraksi dengan menggunakan Fuzzy Association Rule Mining diadapatkan nilai accuracy sebesar 0.984, precision sebesar 0.9253, recall sebesar 1, dan F-Measure sebesar 0.961.

2. Pengujian dengan Data Uji 2

Pada pengujian ini akan dilakukan pencarian nilai accuracy, precision, recall, dan F-Measure dengan menggunakan 200 data email non spam dan 50 data email spam berdasarkan keyword dari ekstraksi fitur kata kunci dari data SPAM yang sama.
Berdasarkan pengujian yang dilakukan dengan menggunakan 500 kata kunci dari HAM (Non SPAM) dan SPAM dari hasil ekstraksi dengan menggunakan Fuzzy Association Rule Mining diadapatkan nilai accuracy sebesar 0.968, precision sebesar 0.875, recall sebesar 0.98, dan F-Measure sebesar 0.924.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 02, Nomor 02, Desember 2017 : 94 - 98

3. Pengujian dengan Data Uji 3

Pada pengujian ini akan dilakukan pencarian nilai accuracy, precision, recall, dan F-Measure dengan menggunakan 300 data email non spam dan 75 data email spam berdasarkan keyword dari ekstraksi fitur kata kunci dari data SPAM yang sama.

Berdasarkan pengujian yang dilakukan dengan menggunakan 500 kata kunci dari HAM (Non SPAM) dan SPAM dari hasil ekstraksi dengan menggunakan Fuzzy Association Rule Mining diadapatkan nilai accuracy sebesar 0.968, precision sebesar 0.871, recall sebesar 0.987, dan F-Measure sebesar 0.925.

4. Pengujian dengan Data Uji 4

Pada pengujian ini akan dilakukan pencarian nilai accuracy, precision, recall, dan F-Measure dengan menggunakan 400 data email non spam dan 100 data email spam berdasarkan keyword dari ekstraksi fitur kata kunci dari data SPAM yang sama.

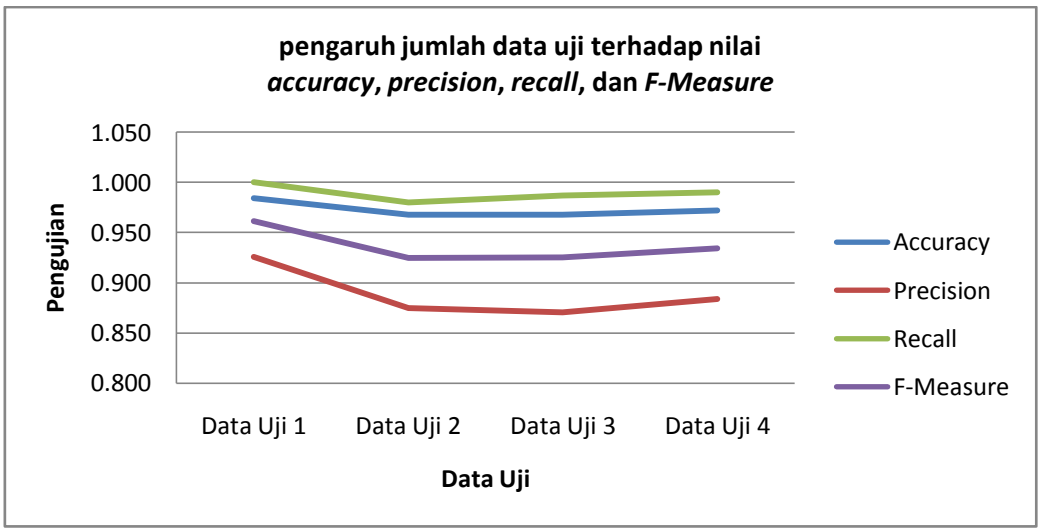
Berdasarkan pengujian yang dilakukan dengan menggunakan 500 kata kunci dari HAM (Non SPAM) dan SPAM dari hasil ekstraksi dengan menggunakan Fuzzy Association Rule Mining diadapatkan nilai accuracy sebesar 0.972, precision sebesar 0.884, recall sebesar 0.99, dan F-Measure sebesar 0.934.

Pengujian terhadap data uji yang berbeda dapat ditunjukkan didalam tabel 5.1 yang merupakan tabel pengaruh jumlah data uji terhadap nilai accuracy, precision, recall, dan F-Measure.

Tabel I. Tabel pengaruh jumlah data uji terhadap nilai accuracy, precision, recall, dan F-Measure

Data Uji	Data Uji 1	Data Uji 2	Data Uji 3	Data Uji 4
Accuracy	0.984	0.968	0.968	0.972
Precision	0.926	0.875	0.871	0.884
Recall	1.000	0.980	0.987	0.990
F-Measure	0.962	0.925	0.925	0.934

Berdasarkan tabel 5.1 dapat digambarkan ke dalam grafik pengaruh jumlah data uji terhadap nilai accuracy, precision, recall, dan F-Measure yang terdapat pada gambar 5.1



Gambar 5.1 Pengaruh jumlah data uji terhadap nilai accuracy, precision, recall, dan F-Measure

Berdasarkan pengujian yang dilakukan terlihat digambar 5.1 bahwa penggunaan data uji 1 memiliki nilai yang paling tinggi diantara yang lain. Tingginya nilai pengujian dari data uji 1 terjadi karena data uji 1 memiliki jumlah data yang sedikit sehingga noise yang terjadi dalam kesalahan pendeteksian e-mail semakin kecil.

E-ISSN : 2540 - 8984

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Volume 02, Nomor 02, Desember 2017 : 94 - 98

IV. KESIMPULAN

Berdasarkan pengujian yang dilakukan, penggunaan metode Fuzzy Association Rule Mining dalam pendeteksian SPAM dan non SPAM pada e-mail adalah baik, hal ini terlihat pada pengujian dimana nilai accuracy, precision, recall, dan F-Measure cukup tinggi yang mendekati nilai 1.

DAFTAR PUSTAKA

[1] S. Appavu, Arravind, Athiappan, Bharatiraja, M. Pandian, and R. Rajaram, "Association Rule Mining for Suspicious Email Detection: A Data Mining Approach," *IEEE*, pp. 317–324, 2007.

[2] F. Rozi, C. Fatichah, and D. Purwitasari, "Ekstraksi Kata Kunci Berdasarkan Hipernim Menggunakan Fuzzy Association Rule Mining untuk Pengelompokan Dokumen," *J. Ilm. Teknol. Inf.*, vol. 13, no. 2, pp. 190–197, 2015.

[3] F. Rozi and R. Kartadie, "Clustering Dokumen dengan Semantic Word Holonim dan Fuzzy Association Rule Mining," *Semnasteknomedia Online*, vol. 5, no. 1, pp. 13–18, 2017.

[4] C. Chen, F. S. C. Tseng, and T. Liang, "Mining fuzzy frequent itemsets for hierarchical document clustering," *Inf. Process. Manag.*, vol. 46, no. 2, pp. 193–211, 2010.

[5] F. Rozi and R. Kartadie, "Sinonim untuk ekstraksi kata kunci pada pengelompokan dokumen menggunakan fuzzy association rule mining," *Semnasteknomedia Online*, pp. 7–12, 2016.

[6] F. Rozi and F. Sukmana, "Metode siklis dan adaptive neuro fuzzy inference system untuk peramalan cuaca," *J. Ilm. Penelit. dan Pembelajaran Inform.*, vol. 1, no. 1, pp. 7–13, 2016.

[7] F. Rozi and F. Sukmana, "Penggunaan moving average dengan metode hybrid artificial neural network dan fuzzy inference system untuk prediksi cuaca," *J. Ilm. Penelit. dan Pembelajaran Inform.*, vol. 1, no. 2, pp. 38–42, 2016.