JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNOLOGI INFORMASI

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

USULAN TUGAS AKHIR

1. IDENTITAS PENGUSUL

NAMA	**: Agus Tri Wibowo**
NRP	**: 5110100156**
DOSEN WALI	**: Imam Kuswardayan, S.Kom., M.Kom.**
DOSEN PEMBIMBING	**: 1. Ahmad Saikhu, S.Si., M.T.**
	2. Rully Soelaiman, Ir., M.Kom.

2. JUDUL TUGAS AKHIR

"Implementasi Algoritma Deteksi *Spam* yang Tersisipi Informasi Citra dengan Metode SVM dan *Random Forest* "

3. LATAR BELAKANG

Email spam merupakan *email* yang tidak diinginkan yang masuk pada *inbox email* kita. *Email* ini biasanya berisi pesan atau citra yang berupa iklan produk atau bahkan tipuan. *Email spam* masih merupakan masalah yang masih melandai dunia *internet* sampai saat ini. *Email spam* biasanya dikirimkan secara massal oleh *botnet* yang dikendalikan *spammer*. Menurut studi yang telah dilakukan oleh The Radicati Group, Inc., pada tahun 2013 *email spam* telah mencakup 84% dari keseluruhan *traffic email* per harinya [1]. Tentu saja hal ini menimbulkan banyak kerugian baik itu bagi pengguna *email* maupun penyedia layanan *email*. *Email spam* merugikan bagi pengguna *email* karena *email spam* telah membuang-buang waktu pengguna *email* ketika mengecek *email* mereka, bahakan tidak sedikit pula yang menjadi korban penipuan karena *email spam* yang diterimanya. Sedangkan bagi penyedia layanan *email*, *email spam* telah memakan sumber daya *bandwidth* yang terbatas dan media penyimpanan *email* yang telah disediakan.

Sudah begitu banyak juga metode yang diajukan dari berbagi riset yang dilakukan oleh kalangan akademisi dan perusahaan untuk memfilter *email spam*. Salah

satu metode yang populer adalah *Bayesian Filtering* yang telah digunakan oleh *SpamAssassin* sebuah perangkat lunak untuk *email spam filtering* [2].

Seiring dengan berkembangnya teknologi *email spam filtering*, para *spammer* juga mengembangkan metode yang digunakan untuk menyebarkan *email spam* mereka salah satunya yaitu dengan menyamarkan pesan yang biasanya menggunakan teks biasa dengan menggunakan medium teks yang ada pada citra sehingga hal ini sangat menyulitkan perangkat *spam filtering* yang telah dijalankan oleh penyedia layanan *email*. Untuk mengatasi hal tersebut telah digunakan *Optical Character Recognition (OCR)* untuk mengenali teks yang ditanamkan pada citra, akan tetapi *spammer* kemudian menggunakan CAPTCHA (*Completely Automated Public Turing Test to Tell Computer and Human Apart*), dengan ini *spammer* bisa men-*distort*, menambahkan latar yang berwarna-warni atau ber-*noise* sehingga hanya manusia saja yang bisa membaca teks pada citra [3] [4]. Oleh karena itu dikembangkan teknik *email spam filtering* yang berdasarkan fitur *low-level* yang terkandung di dalam citra pada *email spam*.

Pada tugas akhir ini, citra pada *email* yang berbentuk citra akan akan diklasifikasikan berdasarkan fitur tekstur yang terkandung pada citra itu dengan menggunakan metode klasifikasi SVM (*Support Vector Machine*) sebagai citra *spam* atau *ham* (citra bukan *spam*), kemudian akan digunakan juga metode klasifikasi *Random Forest* sebagai pembanding.

4. RUMUSAN MASALAH

Rumusan masalah yang diangkat dalam tugas akhir ini adalah sebagai berikut:

1. Memahami fitur tekstur yang penting untuk klasifikasi citra pada *email*.
2. Memahami penerapan metode klasifikasi SVM dan *Random Forest*.
3. Menerapkan metode klasifikasi SVM dan *Random Forest.*
4. Menyusun uji coba menggunakan SVM dan *Random Forest*.

5. BATASAN MASALAH

Permasalahan yang dibahas dalam tugas akhir ini memiliki beberapa batasan, yaitu sebagai berikut:

1. Ekstraksi fitur tekstur akan menggunakan perangkat lunak MaZda [5] [6] [7].
2. Implementasi menggunakan bahasa pemrograman Java dengan pustaka dari perangkat lunak Weka [8].
3. Implementasi menggunakan metode klasifikasi SVM dan *Random Forest* sebagai pembanding.
4. *Dataset* yang digunakan adalah Dredze *dataset* [9] dan *Image Spam Hunter dataset* [10].

6. TUJUAN PEMBUATAN TUGAS AKHIR

Tugas akhir ini bertujuan untuk:

1. Mengetahui penerapan klasifikasi dengan metode SVM dan *Random Forest* untuk mendeteksi citra *spam*.
2. Mengimplementasikan klasifikasi dengan metode SVM dan *Random Forest* untuk mendeteksi citra *spam*.
3. Mengevaluasi kinerja SVM dibandingkan dengan *Random Forest* dengan melakukan uji coba.

7. MANFAAT TUGAS AKHIR

Tugas akhir ini dikerjakan dengan harapan mendapatkan metode yang cepat dan efisien untuk mendeteksi citra *spam*.

8. TINJAUAN PUSTAKA

Biasanya *email spam filtering* dilakukan dengan memeriksa isi, *header* atau keduanya. Pada klasifikasi *email spam,* digunakan *machine learning* untuk beberapa fitur yang diekstrak dari isi *email* kemudian mengklasifikasikan *email* itu apakah *email* itu *spam* atau *ham* (*email* bukan *spam*). *Machine Learning* pada *email spam filtering* dibagi menjadi dua [11], yaitu: "*Non Content based (Header-based) spam filtering*" dan "*Content-based spam filtering*". Pada *Content-based, email* yang datang akan diperiksa isinya untuk dicari *keyword* atau suatu *feature* yang biasanya dipakai oleh *spammer*. Ada juga cara lain yang menggunakan *pattern recognition* untuk mengenali *email spam* yang mempunyai *pattern* tertentu. Sebuah *email* bisa saja berisi teks, citra atau bahkan keduanya. Pada tugas akhir ini akan berfokus pada deteksi *email spam* yang memuat informasi citra.

Email spam akan dideteksi berdasarkan fitur tekstur yang diekstrak dari informasi citra yang disisipkan pada *email*. Seperti yang telah diajukan oleh Andrzej Materka [12], 6 fitur tekstur penting yang akan digunakan untuk mendeteksi citra *spam* adalah: *image histogram, image gradient, run-length matrix, co-occurrence matrix, autoregressive, wavelet-transform.* Perangkat lunak MaZda digunakan untuk mengekstrak fitur-fitur tersebut. Sedangkan SVM digunakan untuk mengklasifikasikan *email spam* dan *Random Forest* akan digunakan sebagai metode klasifikasi pembanding.

Support Vector Machine (SVM) merupakan salah satu *supervised learning* yaitu data yang dijadikan *input* sudah mempunyai label masing-masing. Label ini menunjukkan tiap data itu termasuk dalam suatu kelas. SVM biasanya digunakan untuk klasifikasi biner. Cara kerja metode ini ialah dengan mencari *hyperplane* yang digunakan untuk memisahkan *d*-dimensional data menjadi 2 kelas. Akan tetapi, di dalam praktek dalam dunia nyata, data sering kali tidak bisa dipisahkan secara linear. Sehingga diperkenalkan SVM dengan *"kernel induced feature space"* yang mana data yang akan diklasifikasikan akan dikonversi ke ruang dimensi yang lebih tinggi dimana data dapat dipisahkan [13].

Random Forest merupakan metode *bagging* yaitu metode yang membangkitkan

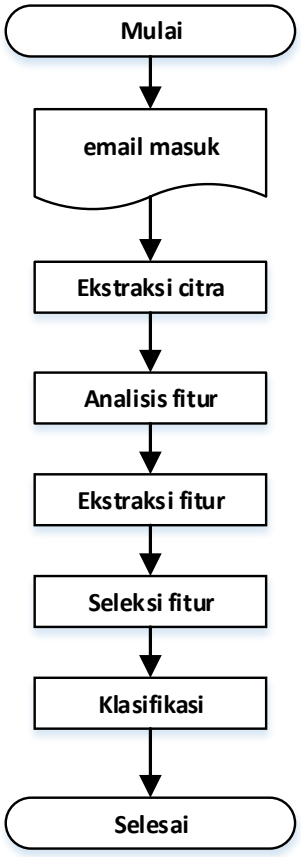
banyak *tree* dari data *sample* yang mana pembuatan satu *tree* pada saat *training* tidak bergantung pada *tree* sebelumnya kemudian keputusan diambil berdasarkan *voting* terbanyak [14].

9. RINGKASAN ISI TUGAS AKHIR

Tekstur merupakan fitur citra yang mengandung banyak informasi. Pada umumnya tekstur adalah *pattern* visual yang kompleks yang terdiri dari entitas yang mempunyai karakteristik kecerahan, warna, kemiringan, ukuran dan lain sebagainya. Pemilihan fitur tekstur citra untuk deteksi citra *spam* didasari oleh fakta bahwa citra yang dihasilkan oleh komputer mempunyai kualitas tekstur yang berbeda dibanding yang dihasilkan oleh non-komputer.

Fitur tekstur yang akan digunakan pada deteksi citra *spam* ini adalah *image histogram, image gradient, run-length matrix, co-occurrence matrix, autoregressive, wavelet-transform* karena fitur-fitur ini dianggap sebagai fitur yang paling penting untuk analisis tekstur pada citra [12].

Proses pendeteksian citra *spam* pada *email* dapat dilihat pada Gambar 1.



Gambar 1. Diagram alir proses pendeteksian citra *spam* pada *email*

Keterangan diagram:

1. **Email masuk**

Email yang berisi citra dijadikan *input*.

2. **Ekstraksi citra**
   Ekstraksi citra dari *email*.
3. **Analisis fitur**
   Melakukan analisis tekstur pada citra yang telah diekstrak.
4. **Ekstraksi fitur**
   Menghitung fitur vektor dari tekstur citra dari parameter tekstur citra. Perangkat lunak MaZda digunakan untuk analisis dan ekstraksi fitur.
5. **Seleksi fitur**
   Melakukan *pre-processing* pada fitur yang telah dihitung untuk menentukan fitur-fitur yang signifikan. Pada proses ini PCA (*Principal Component Analysis*) digunakan untuk menentukan fitur-fitur tersebut.
6. **Klasifikasi**
   Fitur citra yang telah dipilih digunakan untuk mengklasifikasikan citra.

10.METODOLOGI

a. Penyusunan proposal tugas akhir

Proposal tugas akhir ditulis untuk mengajukan ide atas pengerjaan tugas akhir. Proposal tugas akhir juga mengandung proyeksi hasil dari ide tugas akhir yang diajukan.

b. Studi literatur

Pada proses ini dilakukan studi lebih lanjut terhadap konsep-konsep yang terdapat pada jurnal, buku, artikel, dan literatur lain yang menunjang. Studi dilakukan untuk mendalami konsep algoritma SVM dan *Random Forest* dalam menyelesaikan permasalahan yang muncul pada proses pengerjaan tugas akhir ini.

c. Implementasi perangkat lunak

Implementasi merupakan tahap untuk membangun sistem tersebut. Metode klasifikasi yang akan diimplementasikan adalah SVM dan *Random Forest* sebagai pembanding. Implementasi diproses menggunakan bahasa pemrograman Java dengan pustaka dari Weka.

d. Pengujian dan evaluasi

Pengujian dilakukan dengan membandingkan kinerja dari klasifikasi dengan metode SVM dan metode *Random Forest* dalam hal kinerja, kecepatan dan kualitas solusi dengan beberapa percobaan.

e. Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam tugas akhir ini serta hasil dari implementasi aplikasi perangkat lunak yang telah dibuat. Sistematika penulisan buku tugas akhir secara garis besar antara lain:

1. Pendahuluan
   a. Latar Belakang
   b. Rumusan Masalah
   c. Batasan Tugas Akhir
   d. Tujuan
   e. Metodologi
   f. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

11. JADWAL KEGIATAN

Jadwal perencanaan tugas akhir ini dijelaskan pada Tabel 1.

Tabel 1. Jadwal kegiatan pengerjaan tugas akhir

Tahapan	2014																			
	Februari				Maret				April				Mei					Juni		
Penyusunan Proposal																				
Studi Literatur																				
Perancangan sistem																				
Implementasi																				
Pengujian dan evaluasi																				
Penyusunan buku																				

12. DAFTAR PUSTAKA

[1] THE RADICATI GROUP, INC., "Email Statistics Report, 2009-2013," THE RADICATI GROUP, INC., Palo Alto, 2013.

[2] Apache Software Foundation, "What is SpamAssassin?," 13 Februari 2009. [Online]. Available: http://wiki.apache.org/spamassassin/SpamAssassin. [Diakses 18 Februari 2013].

[3] N. Aye dan M. W. Win, "Identification of Image Spam by Using Histogram," *International Journal of Science and Research (IJSR)* , vol. 2, no. 11, p. 310, 2013.

[4] B. Al-Duwairi, I. Khater dan O. Al-Jarrah, "Detection Image Spam Using Image Texture Features," *International Journal for Information Security Research (IJISR),* vol. 2, no. 3/4, p. 344, 2012.

[5] M. Strzelecki, P. Szczypinski, A. Materka dan A. Klepaczko, "A software tool for automatic classification and segmentation of 2D/3D medical images," *Nuclear Instruments & Methods In Physics Research A,* vol. 702, pp. 137-140, 2013.

[6] P. Szczypinski, M. Strzelecki dan A. Materka, "MaZda - a Software for Texture Analysis," *Proc. of ISITC 2007*, no. November 23-23, pp. 245-249, 2007.

[7] P. Szczypinski, M. Strzelecki, A Materka dan A. Klepaczko, "MaZda-A software package for image texture analysis," *Computer Methods and Programs in Biomedicine,* vol. 94(1), pp. 66-76, 2009.

[8] I. H. Witten, E. Frank dan M. A. Hall, DATA MINING: Practical Machine Learning Tools and Techniques 3rd edition, Burlington: Morgan kaufmann, 2010.

[9] R. M. Dredze dan A. Elias-Bachrach, "Learning Fast Classifers for Image Spam," dalam *Proc. CEAS 2007*, Mountain View, 2007.

[10] Y. Gao, M. Yang dan X. Zhao, "Image Spam Hunter," *Acoustics, Speech and Signal Processing ICASSP 2008*, pp. 1765, 1768, 2008.

[11] B. Agrawal, N. Kumar dan M. Molle, "Controlling spam Emails at Routers," *2005 IEEE International Conference* , vol. 3, pp. 1588-1592, 2005.

[12] A. Materka, Texture Analysis Methods – A Review, Brussels: Institute of Electronics, Technical University of Lodz, 1998.

[13] D. Boswell, "Introduction to Support Vector Machines," 6 Agustus 2002. [Online]. Available: http://www.work.caltech.edu/~boswell/IntroToSVM.pdf. [Diakses 25 Februari 2014].

[14] A. Law dan M. Wiener, "Classification and Regression by randomForest," Desember 2002. [Online]. Available: ftp://131.252.97.79/Transfer/Treg/WFRE_Articles/Liaw_02_Classification%20and%20regression%20by%20randomForest.pdf. [Diakses 25 Februari 2014].