PERBANDINGAN ANTARA NAIVE BAYES DAN SINGLE-LAYER PERCEPTRON UNTUK PENDETEKSIAN SPAM EMAIL

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh:

Muhammad Zaki Tamimy
NIM : 09111402011

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2018

LEMBAR PENGESAHAN TUGAS AKHIR

PEMBANDINGAN ANTARA NAÏVE BAYES DAN SINGLE-LAYER PERCEPTRON UNTUK PENDETEKSIAN SPAM EMAIL

Oleh :

Muhammad Zaki Tamimy
NIM: 09111402011

Palembang, 31 Juli 2018

Pembimbing 1	Pembimbing 2
Samsuryadi, M. Kom, Ph.D	Danny Matthew Saputra, S.T.,M.Sc.
NIP. 197102041997021003	NIP. 198505102015041002

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M.T.
NIP. 197706012009121004

TANDA LULUS UJIAN SIDANG TUGAS AKHIR

Pada hari Selasa tanggal 31 Juli 2018 telah dilaksanakan ujian sidang tugas akhir oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

N a m a : **Muhammad Zaki Tamimy**
N I M : **09111402011**
J u d u l : ***Perbandingan Antara Naive Bayes dan Single-Layer Perceptron Untuk Pendeteksian Spam Email.***

1. **Pembimbing I**

   Syamsuryadi, M.Kom., Ph.D. .....
   NIP. 197102041997021003

2. **Pembimbing II**

   Danny M. Saputra, S.T., M.Sc. .....
   NIP. 198505102015041002

3. **Penguji I**

   Rizki Kurniati, M. T. .....
   NIP. 1671045207910003

4. **Penguji I**

   Rifkie Primartha, M. T. .....
   NIP. 197706012009121004

Mengetahui,
Ketua Jurusan Teknik Informatika

Rifkie Primartha, M. T.
NIP. 197706012009121004

HALAMAN PERNYATAAN PLAGIAT

Yang bertanda tangan di bawah ini :

Nama	: Muhammad Zaki Tamimy
NIM	: 09111402011
Program Studi	: Teknik Informatika
Judul Skripsi	: Pembandingan Antara Naive Bayes dan Single-Layer Perceptron untuk Pendeteksian Spam Email
Hasil Pengecekan Software *iThenticate/Turnitin*	: 16 %

Menyatakan bahwa Laporan Projek saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan projek ini, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian, pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapapun.

Palembang, 31 Juli 2018

(Muhammad Zaki Tamimy)
NIM. 09111402011

Motto:

"Nothing ventured, nothing gained!!!"

–Ben Franklin

I dedicated this paper to:

- My Mom and Dad
- My Family
- Beloved friends
- Respected teachers
- Informatics Engineering Unsri
- Sriwijaya University

PEMBANDINGAN ANTARA NAÏVE BAYES DAN SINGLE-LAYER PERCEPTRON UNTUK PENDETEKSIAN SPAM EMAIL

Oleh :

Muhammad Zaki Tamimy

09111402011

ABSTRAK

Permasalahan Spam pada email perlu diatasi supaya tidak menggangu kenyamanan pengguna email. Metode paling umum untuk pendeteksian *anti-spam* bekerja dengan memisahkan atau mengelompokkan isi *email* kemudian mempelajari sejumlah email spam yang telah terdeteksi. Metode pendeteksian spam *email* yang digunakan *Naïve Bayes Classifier* (NB) dan *Single-Layer Perceptron Classifier* (SLP) dengan masing-masing performa akurasi sebesar 88,75% dan 81,25%. Metode NB lebih tinggi sebesar 7,50% dari metode SLP dalam pendeteksian spam.

Kata kunci: Naïve Bayes Classifier, Spam Email, Single-Layer Classifier.

COMPARISON BETWEEN NAÏVE BAYES AND SINGLE-LAYER PERCEPTRON FOR SPAM EMAIL DETECTION

By :

Muhammad Zaki Tamimy

09111402011

ABSTRACT

Spam issues in e-mails need to be addressed so as not to interfere with e-mail user convenience. The most common method for detection of anti-spam works by separating or grouping the contents of the email and then studying a number of spam emails that have been detected. The spam email detection method used by the Naïve Bayes Classifier (NB) and Single-Layer Perceptron Classifier (SLP) respectively have resulted in accuracy performance of 88,75% and 81,25. The NB method is 7,50% higher than the SLP method for spam detection.

Keyword: Naïve Bayes Classifier, Spam Email, Single-Layer Classifier.

KATA PENGANTAR

Segala puji bagi Allah SWT yang telah memberikan rahmat dan hidayah-Nya sehingga dapat menyelesaikan Laporan Tugas Akhir ini dengan baik. Tugas Akhir ini disusun sebagai persyaratan kelulusan tingkat sarjana pada Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Selama pembuatan Laporan Tugas Akhir ini, disadari penulis tak dapat luput dari hambatan dan kesulitan, namun berkat bimbingan dan pengarahan serta bantuan dari berbagai pihak, maka penulis dapat menyelesaikan laporan ini. Untuk itu pada kesempatan ini penulis ingin menyampaikan ucapan terima kasih kepada:

1. Orang tuaku, Edy Herman dan Elvera Movida, saudaraku, Aan, Adi dan Yudith dan seluruh keluarga besar tercinta yang selalu memberikan dukungan moril, materi dan do'a tanpa henti.
2. Pamanku Rully Rusdy Cosim, Tante Linda dan Bude Erna yang telah bersabar dan merawatku selalu memberikan dukungan selama merantau.
3. Bapak Jaidan Jauhari, M. T., selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Bapak Rifkie Primartha, M. T., selaku Ketua Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
5. Bapak Syamsuryadi, S.Si., M.Kom., Ph.D., selaku pembimbing satu Tugas Akhir dan Pembimbing Akademik dan Bapak Danny Matthew Saputra,S.T., M.Sc., selaku pembimbing dua Tugas Akhir, merekalah yang telah banyak memberikan bimbingan, masukan, arahan, dan pengetahuan selama proses penyelesaian Tugas Akhir ini.

6. Ibuk Rizki Kurniati, M. T. dan Bapak Rifkie Primartha, M. T., selaku dosen penguji yang telah memberikan koreksi dan masukan untuk Tugas Akhir ini.
7. Segenap staff pengajar di Fakultas Ilmu Komputer Universitas Sriwijaya yang telah mengajar, membimbing, dan memberikan ilmu kepada penulis.
8. Mbak Wiwin Juliani S. SI., selaku staff administrasi Teknik Informatika Bilingual yang telah membantu dalam hal urusan akademik dan administrasi selama perkuliahan penulis.
9. Muhammad Fajriandi dan Dyah Lindung Pengasih yang telah membantu penulis untuk mengembangkan pengalaman dan ide, serta memberikan inspirasi akademik dan moril beserta hal yang lainnya.
10. Sahabat-sahabat seperjuangan IF BILINGUAL 2011 yang selalu saling mendukung dalam suka dan duka;
11. Serta pihak-pihak lainnya yang terlibat selama pelaksanaan Tugas Akhir ini yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa Tugas Akhir ini jauh dari kesempurnaan. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun untuk kesempurnaan Tugas Akhir ini. Semoga Tugas Akhir ini dapat memberikan manfaat bagi pihak yang membacanya.

Palembang, 31 Juli 2018

Muhammad Zaki Tamimy
NIM 09111402011

DAFTAR ISI

BAB IV PENGEMBANGAN PERANGKAT LUNAK

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Pendahuluan

Berperan sebagai pembuka, bab ini akan menjabarkan pondasi-pondasi yang digunakan pada penelitian. Penjelasan mengenai pembandingan pendeteksian spam akan dimulai dari latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan penelitian hingga ke sistematika penulisan.

1.2 Latar Belakang Masalah

Spam adalah pesan sampah atau pesan tidak diinginkan yang dikirimkan kepada anda dengan konten dari hal-hal yang tidak penting atau tidak terkait seperti iklan produk, penawaran jasa, atau bahkan berisi konten tidak senonoh. Permasalahan yang disebabkan oleh spam telah meningkat sepanjang tahun. Sebuah statistik tahun 2011 menemukan bahwa 40% dari semua *email spam* atau tepatnya 15.4 triliun spam email tersebar setiap harinya dan berdampak terhadap pengguna internet sehingga menyebabkan kerugian hingga 355 juta dolar per tahun (Awad dan Elsoufi, 2011). *Email* menyediakan cara yang terbaik dalam mengirimkan jutaan iklan bebas biaya oleh pengirim. Namun, fakta ini secara luas dimanfaatkan oleh beberapa individu dan organisasi. Akibatnya, kotak surat elektronik jutaan orang dipenuhi dengan semua email yang tidak diinginkan yang juga diketahui sebagai "*spam*" atau "*junk mail*".

Cara yang paling sederhana untuk mengatasi *spam* adalah dengan memblokir email yang yang datang dari alamat IP tertentu atau dengan memfilter judul subjek tertentu. Metode yang paling umum yang disebut sebagai pendeteksian *anti-spam* bekerja dengan memisahkan atau mengelompokkan isi *email* kemudian mempelajari sejumlah *email spam* yang telah terdeteksi. Proses ini dapat mencegah spam masuk kedalam kotak pesan dan memperingatkan penerima tentang *spam* tersebut dikemudian harinya. Aktor yang menyebarkan email yang mengandung spam disebut sebagai *spammer*. Saat ini *spammer* menggunakan metode yang unik untuk dapat melewati *filter* dengan mengirimkan *spam* dengan alamat acak atau memasukkan karakter acak pada awal dan akhir subjek *email* unuk melewati mesin pendeteksi *anti-spam*. Perilaku ini menjadi alasan mengapa masalah spam tampaknya masih bertahan dan dengan teknik tercanggih masa kini pun tidak memberikan cukup bukti dalam mengatasi masalah ini.

Saat ini, banyak riset yang telah dilakukan untuk mendeteksi spam dengan menggunakan *Bayesian classifier* dan *Single-Layer Perceptron* (SLP). Marsono et al., (2009) telah mendemonstrasikan bahwa konten klasifikasi *Naïve Bayes* (NB) dapat diadaptasi untuk prosesing layer-3, tanpa memerlukan perakitan ulang. James Clatke telah berhasil menetapkan Jaringan Syaraf Tiruan (JST) yang akan digunakan dalam proses email menuju kotak surat otomatis dan filter *spam.*

Metode NB dan SLP dijadikan subjek untuk perbandingan karena kedua metode tersebut memiliki prinsip cara kerja yang berbeda. Dapat dilihat dari *Naive Bayes* yang menggunakan probabilitas dan statistik dalam pendeteksian sedangkan SLP menggunakan bobot yang disesuaikan secara bertahap per-sampel. Menurut

Awad dan Elsoufi (2011) *machine learning* telah dipelajari secara luas dan terdapat banyak metode yang dapat digunakan dalam penyaringan email. Salah-satu metode yang disebut adalah Naive Bayes dan Single-Layer Perceptron. Dalam penelitiannya NB mendapatkan hasil akurasi yang lebih tinggi sedangkan metode SLP dengan sifatnya yang adaptif memiliki akurasi lebih rendah, tetapi dalam penelitian yang dilakukan oleh Konstantin Tretyakof (2004) metode SLP-lah yang mendapatkan hasil lebih baik daripada NB, dengan SLP mendapatkan akurasi tertinggi diantara metode lainnya. Penggunaan NB dan SLP dipertimbangkan oleh peneliti karena kedua metode tersebut termasuk kedalam metode yang sering dipakai dalam sistem penyaringan spam. Oleh karena itu, maka diperlukan suatu perangkat lunak untuk pendeteksian spam dengan membandingkan metode *Naive Bayes* dan *Single-Layer Perceptron.*

1.3 Rumusan Masalah

Permasalahan yang terjadi pada pembandingan metode pendeteksian spam email sebagai berikut:

1. Bagaimana mengurangi jumlah pesan sampah (*spam*) yang masuk kedalam *inbox* dengan mendeteksinya dengan metode pengklasifikasi Bayesian dan pengklasifikasi SLP.
2. Bagaimana menentukan metode terbaik dalam mendeteksi spam yang menghasilkan jumlah penjaringan spam tertinggi dan jumlah terjaringnya pesan asli (*ham*) terendah.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Mengembangkan perangkat lunak komputer untuk pendeteksian menggunakan pengklasifikasi *Bayesian* dan pengklasifikasi SLP untuk membaca fitur yang telah disediakan.
2. Memperoleh tingkat performa dalam mendeteksi *spam email* menggunakan metode pengklasifikasi *Bayesian* dan pengklasifikasi SLP.
3. Menunjukkan perbandingan hasil dari dua metode yang berbeda cara kerjanya.

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Dapat mendeteksi *spam* diantara sekumpulan *email* yang berisikan dari berbagai macam konten.
2. Dapat mengetahui metode manakah yang memiliki performa terbaik dalam mendeteksi *spam*
3. Hasil penelitian dapat digunakan sebagai acuan untuk penelitian-penelitian pendeteksi spam email selanjutnya.

1.6 Batasan Masalah

Batasan masalah dalam penelitian tugas akhir ini adalah:

1. Pesan *email* akan diklasifikasikan kedalam *spam* dan *ham*.

2. Pendeteksian akan dilakukan dalam kata-kata didalam *email*. Tidak meliputi *attachment* dan gambar didalamnya
3. Urutan pembelajarannya akan menggunakan sekumpulan *dataset offline* tangan kedua dari website pada alamxat berikut http://csmining.org/index.php/spam-email-datasets-.html.
4. Jumlah *dataset* yang digunakan adalah sebanyak 200 email dengan perbandingan 1:1 antara spam dan ham.
5. Riset yang dimaksud akan dikembangkan menggunakan bahasa pemprograman Java.

1.7 Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini adalah sebagai berikut :

BAB I. PENDAHULUAN

Bab ke 1 berisi penjelasan mengenai latar belakang, rumusan masalah, tujuan, manfaat, batasan masalah, metodologi penelitian, metode pengembangan perangkat lunak dan sistematika penulisan.

BAB II. KAJIAN LITERATUR

Bab ke 2 menguraikan dasar-dasar teori yang digunakan pada penelitian, seperti *Electronic Mail, Pre-Processing, Bayesian Classifier*, Jaringan Syaraf Tiruan, pengembangan perangkat lunak, desain model, bahasa pemrograman Java.

BAB III. METODOLOGI PENELITIAN

Bab ke 3 berisi analisis serta perancangan terhadap penggunaan algoritma Bayesian dan SLP dalam melakukan pelatihan dan pengenalan pada sampel *email* yang dikembangkan untuk pembanding dalam perangkat lunak pendeteksian *spam*.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Bab ke 4 membahas mengenai lingkungan implementasi algoritma Bayesian dan SLP dalam melakukan pelatihan dan pengenalan pada sampel *email* serta implementasi program dan pengujian.

BAB V. HASIL DAN ANALISIS PENELITIAN

Bab ke 5 berisi kesimpulan dari semua uraian-uraian pada bab - bab sebelumnya dan saran untuk penelitian selanjutnya.

BAB VI. KESIMPULAN DAN SARAN

Bab ke 6 berisi kesimpulan dari semua uraian-uraian pada bab-bab sebelumnya dan juga berisi saran-saran yang diharapkan berguna dalam pembandingan pendeteksi *spam* ini.

1.8 Kesimpulan

Penelitian yang bertajuk Pembandingan antara *Naive Bayes* dan *Single-Layer Perceptron* untuk pendeteksian *spam email* berlatar belakang keingin tahuan atas tingkat kualitas metode dalam mendeteksi *spam*. Dibatasi *dataset offline*, penelitian diharapkan menunjukkan performa dari metode yang digunakan sehingga menghasilkan bahan kajian untuk metode terbaik.

DAFTAR PUSTAKA

Almeida, T. A., Almeida, J., & Yamakami, A. (2011). Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers.*Journal of Internet Services and Applications*, *1*(3), 183-200.

Almeida, T., Hidalgo, J. M. G., & Silva, T. P. (2013). Towards sms spam filtering: Results under a new dataset. *International Journal of Information Security Science*, *2*(1), 1-18.

Arram, A. W. (2013). *Spam detection using hybrid of artificial neural network and genetic algorithm* (Doctoral dissertation, Universiti Teknologi Malaysia, Faculty of Computing).

Awad, W. A., & ELseuofi, S. M. (2011). Machine Learning methods for E-mail Classification. *International Journal of Computer Applications*, *16*(1).

Ceska, Z., & Fox, C. (2009). The influence of text pre-processing on plagiarism detection. In *Proceedings of the International Conference RANLP-2009* (pp. 55-59).

Christina, V., Karpagavalli, S., & Suganya, G. (2010). A study on email spam filtering techniques. *International Journal of Computer Applications*, *12*(1), 0975-8887.

Eberhardt, J. J. (2015). Bayesian spam detection. *Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal*, *2*(1), 2.

Graham, P. (2003). A plan for spam, 2002. *Available from World Wide Web: http://www. paulgraham. com/spam. html*.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Haykin, S. Neural networks, 1999. *A Comprehensive Foundation*.

Hecht-Nielsen, R. (1998, October). A Theory of the Cerebral Cortex. In *ICONIP*(pp. 1459-1464).

Hershkop, S. (2006). *Behavior-based email analysis with application to spam detection* (Doctoral dissertation, Columbia University).

Knoernschild, K. (2002). *Java design: objects, UML, and process*. Addison-Wesley Professional.

Krenker, A., Bester, J., & Kos, A. (2011). Introduction to the artificial neural networks. *Artificial neural networks: methodological advances and biomedical applications. InTech, Rijeka. ISBN*, 978-953.

Kruchten, P. (2000). From Waterfall to Iterative Lifecycle-a tough transition for project managers.

Kusumadewi, Sri. 2004. Membangun Jaringan Syaraf Tiruan Menggunakan Matlab dan Excel Link. Yogyakarta: Graha Ilmu.

Marsono, M. N., El-Kharashi, M. W., & Gebali, F. (2008). Binary LNS-based naïve Bayes inference engine for spam control: noise analysis and FPGA implementation. *Computers & Digital Techniques, IET*, *2*(1), 56-62.

Marsono, M. N., El-Kharashi, M. W., & Gebali, F. (2009). Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification. *Computer Networks*, *53*(6), 835-848.

Padhy, N., Mishra, D., & Panigrahi, R. (2012). The survey of data mining applications and feature scope. *arXiv preprint arXiv:1211.5723*.

Sutojo, T., Mulyanto, E., & Suhartono, V. (2011). Kecerdasan Buatan. *Edisi Pertama. Andi Offset, Yogyakarta*.

Tretyakov, Konstantin. "Machine learning techniques in spam filtering." *Data Mining Problem-oriented Seminar, MTAT*. Vol. 3. No. 177. 2004.

Triawati, C., Bijaksana, M. A., Indrawati, N., & Saputro, W. A. (2009). Pemodelan Berbasis Konsep untuk Kategorisasi Artikel Berita Berbahasa Indonesia. In *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*. Dokumen Berbahasa Indonesia.