PROPOSAL TUGAS AKHIR

PENERAPAN NAÏVE BAYES DAN NAMED ENTITY RECOGNITION PADA INTRUSION DETECTION BERBASIS JURNAL

Disusun Oleh:

KUNI NUR AINI

NIM M0514029

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

SURAKARTA

2017

UNIVERSITAS SEBELAS MARET
PROGRAM STUDI INFORMATIKA

PROPOSAL TUGAS AKHIR

Nama : **KUNI NUR AINI**

NIM : **M0514029**

PERSETUJUAN PEMBIMBING

Proposal Tugas Akhir ini telah disetujui oleh

Kandidat Pembimbing I,

Drs. Bambang Harjito, M.App.Sc.,Ph.D.

NIP. 196211301991031002

1 JUDUL

PENERAPAN NAÏVE BAYES DAN NAMED ENTITY RECOGNITION PADA INTRUSION DETECTION BERBASIS JURNAL

2 PENDAHULUAN

2.1 Latar Belakang

Informasi saat ini sudah menjadi sebuah komoditi yang sangat penting. Jatuhnya informasi ke tangan yang tidak berhak dapat menimbulkan kerugian bagi pemilik informasi. Dalam jaringan komputer, informasi dapat disediakan dengan cepat. Sebagai dampaknya, terhubungnya komputer ke suatu jaringan komputer membuka potensi adanya lubang keamanan *(security hole).* Untuk mencapai tujuan, paket informasi yang dikirimkan melalui jaringan komputer harus melalui beberapa sistem (router, gateway, hosts, atau perangkat-perangkat komunikasi lainnya) yang berada di luar kendali. Setiap titik yang dilalui memiliki potensi untuk dibobol, disadap, dipalsukan. Hal ini mengundang penjahat untuk melakukan suatu tindakan criminal (McClure, Scambray, Kurtz, & Kurtz, 2009).

Serangan terhadap keamanan sistem informasi melalui jaringan komputer semakin meningkat baik dari segi kuantitas maupun kualitas (McClure et al., 2009). Berbagai macam serangan pada *layer network* dapat terjadi. Hal ini menjadi salah satu permasalahan yang menarik untuk dijadikan sebagai sebuah *study*.

Intrusion Detection System (IDS) adalah sistem yang mendeteksi serangan yang terdapat pada jaringan komputer. Sistem tersebut mendeteksi aktivitas yang menyimpang dari pola aktivitas normal (Azis & Adiwijaya). Pada peneitian ini sistem akan digunakan untuk mendeteksi jenis serangan dalam dokumen.

Banyaknya dokumen memungkinkan sistem harus dapat mengkategorikan jenis serangan yang dibahas dalam dokumen tersebut. Untuk itu, maka perlu dilakukankan pengelompokan dokumen agar sistem dapat mendeteksi dengan benar mengenai pembahasan jenis serangan dalam

suatu dokumen. Pengelompokan dokumen bisa dilakukan dengan berbagai teknik, salah satunya adalah *Text Mining*. *Text Mining* adalah salah satu bidang khusus dari Data Mining yang merupakan suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis, salah satuya adalah melakukan kategorisasi/klasifikasi (Triawati, 2009).

Kategorisasi teks memiliki berbagai cara pendekatan salah satunya berbasis numeris, misalnya pendekatan *probabilistic, support vector machine, dan artificial neural network.* Pendekatan berbasis probabilistic *Naïve Bayes Classifier (NBC)* memiliki beberapa kelebihan antara lain, sederhana, cepat dan berakurasi tinggi. Metode *NBC* untuk klasifikasi atau kategorisasi teks menggunakan atribut kata yang muncul dalam suatu dokumen sebagai dasar klasifikasinya.

Penelitian (Li, Li, Xia, & Zhang, 2011) menujukkan bahwa metode *Naïve Bayes* bekerja dengan baik dalam klasifikasi dokumen teks untuk pelacakan topik pada dokumen berita. Metode *Naïve Bayes* juga dibahas oleh (Hamzah, 2012) untuk klasifikasi teks berita dan abstract akademis, penelitian tersebut menunjukkan bahwa meskipun asumsi independensi antar kata dalam dokumen tidak sepenuhnya dapat dipenuhi, tetapi kinerja *NBC* dalam klasifikasi relative sangat bagus dengan hasil akurasi maksimal pada dokumen berita sebesar 91% sedangkan pada dokumen akademik 82%. Penelitian (Wulandini & Nugroho, 2009) membandingkan metode klasifikasi teks *NBC* dengan metode *Support Vector machine (SVM), C4.5* dan *K-Nearest Neighbour (K-NN).* Hasil penelitian menunjukkan akurasi masing-masing metode secara urut dari yang terbaik adalah *SVM* akurasi 92%, *NBC* akurasi 90%, *C4.5* akurasi 77.5%% dan yang terendah *K-NN* akurasi 50%. Penelitian (Wirawan & Eksistyanto, 2015) menerapkan metode *Naive Bayes* pada *Intrusion Detection System* dengan diskritisasi variable memberikan hasil bahwa penambahan diskritisasi variable menjadikan probabilitas dari algoritma naive bayes yang digunakan dapat lebih dihandalkan untuk menentukan kelas dari suatu data. Penelitian (Mukherjee

& Sharma, 2012) menerapkan metode *Naïve Bayes* pada *Intrusion Detection System* dengan pengurangan fitur masukan sehingga memberikan hasil yang lebih efisien secara komputasi.

Untuk mendapatkan hasil akurasi yang lebih baik, maka sebelum data diklasifikasikan menggunakan *Naïve Bayes*, perlu dilakukan ekstrasi informasi terlebih dahulu pada data tersebut. Ekstraksi informasi pada jurnal penelitian dari header dan referensi jurnal sangat dibutuhkan untuk berbagai aplikasi, misalnya pencarian berbasis bidang, analisis penulis, dan analisis kutipan (Soderland, 1999). Ekstraksi informasi merupakan suatu sistem untuk mencari data spesifik dalam *natural language text*. Kelompok fitur yang akan digunakan pada penelitian ini adalah *named-entity*. Fitur *named entity* adalah fitur non-lokal yang diekstraksi dengan library *Stanford Named Entity Recognition (NER).* Penelitian (Sazali, Rahman, & Bakar, 2016) menerapkan metode *Named Entity Recognition* dalam ekstrasi informasi kata benda pada artikel surat kabar Melayu dan hasil yang didapat dari proses ekstrasi informasi adalah informasi terstruktur yang mirip dengan catatan database. Penelitian (Mahalakshmi, Antony, & Roshini, 2016) menerapkan metode *Named Entity Recognition* dan metode klasifikasi *Naïve Bayes* pada klasifikasi teks bahasa Tamil dan menjukkan hasil akurasi yang cukup baik yaitu 79%.

Intrusion Detection yang akan dilakukan pada penelitian ini menggunakan data berbentuk dokumen jurnal. Metode *Named Entity Recognition* digunakan untuk mengenali entitas tahun. Sedangkan *Naïve Bayes* digunakan untuk proses klasifikasi dan identifikasi jenis serangan. Keunggulan yang dimiliki oleh *Naïve Bayes* yaitu berkeja dengan cepat dan mempunyai akurasi tinggi dan *Named Entity Recognition* bekerja dengan baik dalam mengenali suatu entitas tertentu. Sehingga sangat cocok apabila *Naïve Bayes* dan *Named Entity Recognition* dijadikan sebagai metode untuk melakukan deteksi jenis serangan pada dokumen.

2.2 Rumusan Masalah

Berdasarkan uraian latar belakang, maka rumusan masalah untuk penelitian ini adalah sebagai berikut:

Bagaimana menerapkan metode *Naïve Bayes* dan *Named Entity Recognition* pada *Intrusion Detection* berbasis jurnal*?*

2.3 Batasan Masalah

Dalam penulisan skripsi ini, batasan masalah yang digunakan adalah:

1. Sistem dibangun untuk mengidentifikasi jenis serangan pada dokumen penelitian.
2. Sistem yang dibangun pada tugas akhir ini menggunakan bahasa pemrograman PHP dan database *Mysql*.
3. Data yang digunakan adalah dokumen jurnal mengenai serangan pada *Layer Network.*
4. Data yang diperoleh berasal dari www.sciencedirect.com, ieeexplore.ieee.org, dan penelitan diluar kedua website tersebut.
5. Keyword yang digunakan dalam pengambilan data adalah: attack network, attack layer network, issue attack network, dos, man-in-the-middle, spoofing, malicious code, phishing, probing, dictionary attack, brute-force attack.
6. Data yang diambil adalah data tahun 2014-2017.
7. Jenis serangan yang dapat diidentifikasi dibagi menjadi 7 kategori yaitu: DOS, Probing, Spoofing, Malicious code, Phishing, Man in the middle, dan General.

2.4 Tujuan Penelitian

Bagaimana menerapkan metode Naïve Bayes pada Intrusion Detection System?

Tujuan dari penelitian ini adalah sebagai berikut.

Menerapkan metode *Naïve Bayes* dan *Named Entity Recognition* pada *Intrusion Detection* berbasis jurnal.

2.5 Manfaat Penelitian

Manfaat penulisan skripsi ini diharapkan dapat memberikan manfaat teoritis maupun praktis. Manfaat teoritis yang diperoleh adalah dapat menjadi tambahan acuan dalam pengembangan *Intrusion Detection* berbasis jurnal menggunakan metode *Naïve Bayes* dan *Named Entity Recognition*. Sedangkan manfaat praktis yang diperoleh adalah dapat mengidentifikasi jenis serangan yang dibahas dalam suatu dokumen.

3 PENELITIAN TERKAIT

Berikut adalah beberapa penelitian yang berkaitan dengan penelitian yang diajukan.

3.1 Information Extraction: Evaluating Named Entity Recognition from Classical Malay Documents. 2016 (Siti Syakira Sazali, Nurazzah Abdul Rahman - Faculty of Computer and Mathematical Sciences Universiti Teknologi MARA Shah Alam dan Zainab Abu Bakar Faculty of Computers & Information Technology Al-Madinah International University Shah Alam)

Penelitian ini membahas mengenai proses ekstrasi kata benda pada artikel surat kabar Melayu. Ekstrasi kata benda dilakukan dengan aturan morfologi (Verb, Adjective and Noun Affixes), yaitu dengan proses menganalisis teks berdasarkan teori dan teknologi. Proses tersebut disebut dengan Natural Language Processing (NLP). Salah satu teknologi NLP yang diterapkan pada penelitian ini adalah *Information Extraction*, metode yang dipiilih dalam pada penelitian ini adalah *Named Entity Recognition*. Hasil yang didapat dari proses ekstrasi informasi adalah informasi terstruktur yang mirip dengan catatan database. NER mengacu pada proses pencarian bagian dari teks yang mewakili nama yang tepat dan kemudian mengklasifikasikan nama tersebut ke dalam kategori yang sesuai. Adapun hasil dari penelitian ini yaitu metode NER dengan aturan morfologi (Verb, Adjective and Noun Affixes) adalah cara terbaik untuk mengidentifikasi kata benda pada artikel surat kabar Melayu.

3.2 Domain Based Named Entity Recognition using Naive Bayes Classification. 2016 (G.S. Mahalakshmi, Betina Antony J, Akshaya Kumar, and Bagawathi Roshini S - Department of Computer Science & Engineering., College of Engineering Guindy, Anna University)

Penelitian ini membahas mengenai penggunaan metode *Named Entity Recognition* untuk pengenalan entitas seperti: Lokasi, tuhan, nama dewi, waktu dan sejarah pada teks bahasa Tamil yang akan diklasifikan menggunakan metode *Naïve Bayes.* Pendekatan yang digunakan dalam penelitian ini adalah tokenisasi dan parsing teks. Kerangka kerja pemrosesan statistic menggunakan kamus yang dibuat dari *data training* yang termasuk label standar dari entitas nama. Hasil yang diperoleh padda penelitian ini bahwa menentukan entitas nama dengan teks yang relevan memberikan hasil akurasi sebesar 79%.

3.3 Topic Tracking Based on Naïve Bayes. 2011 (Shuping Li, ChunyanXia, and Wei Zhang - Department of Computer Science and Technology, Mudanjiang Normal University dan Shengdong Li Dept. of Computer Engineering, Langfang Yanjing Vocational and Technical College)

Penelitian ini membahas mengenai klasifikasi dokumen teks untuk pelacakan topic menggunakan metode *Naïve Bayes* dan dimensi fitur *Vector Space Model (VSM). Naïve Bayes* diterapkan sebagai metode untuk melakukan pelacakan topic, sedangkan *VSM* diterapkan sebagai model pre-process topik atau mewakili topik. *VSM* diterapkan dengan menggunakan algoritma *TF-IDF*. Dari 14150 teks berita berbahasa China ang disediakan oleh Dr. Tan Songbo di CAS Institute of Computing terdapat 12 topik berita pada korpus. Evaluasi rata-rata yang diperoleh menggunakan *Recall* yaitu 0.6713 dengan dimensi fitur 500, sedangkan ketika dilakukan penambahan dimensi fitur menjadi 2500 didapatkan eevauasi rata-rata sebesar 0,7191 atau meningkat 7,12%. Sama halnya ketika diukur menggunakan *Precision* hasilnya meningkat dari 0.7399 menjadi 0.7934 atau 7,23%. Dengan

demikian maka *Naïve Bayes* bekerja dengan baik dalam melakukan pelacakan topic dalam dokuen teks.

3.4 Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita Dan *Abstract* Akademis. 2012 (Amir Hamzah - Jurusaan Teknik Informatika, Fakultas Teknologi Industri, Institut Sains dan Teknologi AKPRIND)

Penelitian ini membahas mengenai klasifikasi teks berita dan *abstract* akademis menggunakan metode *Naïve Bayes Classifier (NBC)*. Metode *NBC* mempunyai beberapa kelebihan kesederhanaan dalam komputasinya. Namun metode ini memiliki kelemahan dalam asumsi yang sulit dipenuhi, yaitu independensi feature kata. Penelitian menggunakan data 1000 dokumen berita dan 450 dokumen abstrak akademik. Hasil penelitian menunjukkan pada dokumen berita akurasi maksimal dicapai 91% sedangkan pada dokumen akademik 82%. Sehingga dapat dikatakan bahwa Algoritma NBC memiliki kinerja yang cukup tinggi untuk klasifikasi dokumen teks, baik dokumen berita maupun dokumen akademik.

3.5 Penerapan Naive Bayes pada Intrusion Detection System dengan Diskritisasi Variabel. 2015 (Nyoman Trisna Wirawan dan Ivan Eksistyanto - Institut Teknologi Sepuluh Nopember Surabaya)

Penelitian ini membahas mengenai sistem deteksi gangguan pada jaringan komputer menggunakan metode *Naïve Bayes* dengan *Diskritisasi Variabel. Diskritiasa Variabel* diperlukan untuk mengubah atribut kontinu ke dalam bentuk diskrit. Tujuan penelitian ini yaitu menerapkan *Naïve Bayes Classifier* dengan menggunakan pemilihan atribut berdasarkan pada korelasi serta preprocessing data dengan diskritisasi dengan menggunakan metode mean/standar deviasi untuk atribut kontinu dengan menggunakan 3-interval dan 5-interval. Hasil percobaan menunjukan bahwa penerapan naive bayes pada klasifikasi data yang telah melewati proses diskritisasi mampu memberikan akurasi hingga 89% dengan running time rata-rata adalah 31 detik. Dari hasil tersebut dapa dikatakan bahwa proses klasifikasi dengan

menggunakan proses diskritasasi menjadikan probabilitas dari algoritma naive bayes lebih dihandalkan untuk menentukan kelas dari suatu data, namun proses diskritisasi ini juga menghilangkan beberapa informasi penting yang ada dalam dataset karena teknik ini tidak mempertimbangkan kelas dari suatu data sebelum melewati proses diskritisasi.

3.6 Intrusion Detection using Naive Bayes Classifier with Feature Reduction. 2012 (Saurabh Mukherjee and Neelam Sharma - Department of Computer Science, Banasthali University)

Penelitian ini membahas mengenai sistem deteksi gangguan pada jaringan komputer menggunakan metode *Naïve Bayes* dengan *Feature Reduction.* Tujuan penelitian ini adalah mengidentifikasi pentingnya pengurangan fitur input pada penerapan *IDS,* sehingga akan memberikan hasil yang lebih efisien secara komputasi. Fitur yang digunakan pada penelitian ini adalah Fitur Vitality berbasis metode *Reduction.* Hasil dari penelitian menunjukkan bahwa attribute pengurangan yang sudah dipilih memberikan kinerja yang lebih baik dalam merancang *IDS* yang efektif dan efisien. Selain itu juga digunakan fitur seleksi yaitu *CFS, IG*, dan *GR* untuk dibandingkan dengan *Feature Reduction*. Hasil percobaan menunjukkan bahwa fitur seleksi *(CFS)* dapat lebih banyak meningkatkan hasil akurasi klasifikasi *Naïve Bayes* dibanding dengan IG dan GR. Namun metode metode Fitur Vitality berbasis metode *Reduction* menunjukkan peningkatan yang jauh lebih baik dibanding CFS, tetapi membutuhkan waktu lebih lama.

Perbandingan hasil penelitian terkait yang telah dijabarkan dirangkum dalam Tabel 3.1.

Tabel 1 Penelitian Terkait

No.	Judul Penelitian	Tujuan	Metode	Kelebihan	Kekurangan
1.	Information Extraction: Evaluating Named Entity Recognition from Classical Malay Documents (Sazali et al., 2016)	Mengetahui bagaimana menerapkan metode *Named Entity Recognition* dalam ekstrasi informasi kata benda pada artikel surat kabar Melayu.	*Named Entity Recognition*	- Metode *NER* memberikan informasi terstruktur yang mirip dengan catatan database - Metode *NER* dengan aturan morfologi (Verb, Adjective and Noun Affixes) menjadi cara terbaik untuk mengidentifikasi kata benda pada artikel surat kabar Melayu	- Implementasi metode *NER* berada pada kemampuan definisi pola yang biasanya dilakukan oleh ahli bahasa. - *NER* memiliki ketergantungan yang besar dengan bahasa yang digunakan.
2.	Domain Based Named Entity Recognition using Naive Bayes Classification (Mahalakshmi et al., 2016)	Mengetahui bagaimana menerapkan metode *Named Entity Recognition* dan metode klasifikasi *Naïve Bayes* pada klasifikasi teks bahasa Tamil	*Named Entity Recognition, Naïve Bayes*	- Penerapan *NER* pada penelitian ini dapat dilakukan untuk penenalan entitas seperti: Lokasi, tuhan, nama dewi, waktu dan	- Identifikasi sejarah candi pada teks menggunakan *NER* menimbulkan tantangan, karena banyak kata dalam teks

				sejarah - *Named Entity Recognition* dan *Naïve Bayes* bekerja dengan akurasi yang cukup baik yaitu 79%	akan mengandung nama entitas yang lainnya.
3.	Topic Tracking Based on Naïve Bayes (Li et al., 2011)	Mengetahui bagaimana menerapkan metode *Naïve Bayes* untuk melakukan pelacakan topik pada dokumen teks	Naïve Bayes dengan dimensi fitur *Vector Space Model*	- Dimensi fitur *Vector Space Model* mempengaruhi kinerja pelacakan topic pada dokumen teks - Metode *Naïve Bayes* bekerja dengan baik dalam klasifikasi dokumen teks untuk pelacakan topic.	- Untuk meningkatkan akurasi kerja *Vector Space Model* harus dilakukan penambahan fitur deminsi yang cukup banyak.
4.	Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis (Hamzah, 2012)	Mengetahui hasil klasifikasi dan hasil akurasi penggunaan metode Naïve Bayes Classifier pada pengelompokan teks berita dan abstract akademis.	Naïve Bayes Classifier	- Pada penelitian ini *Naïve Bayes* bekerja cepat dan mempunyai akurasi yang tinggi - *Naïve Bayes* mempunyai	- *Naïve Bayes* mempunyai asumsi independensi antar kata dalam dokumen tidak sepenuhnya dapat dipenuhi.

				kesederhanaan dalam komputasinya	
5.	Penerapan Naive Bayes Pada Intrusion Detection System Dengan Diskritisasi Variabel (Wirawan & Eksistyanto, 2015)	Mendeteksi gangguan pada sistem menggunakan metode Naïve Bayes dan diskritisasi varibael	Naïve Bayes, Corellation Feature Selection, Diskritisasi variabel	- Proses klasifikasi memberikan hasil yang lebih efisien secara komputasi dengan menggunakan Diskritisasi Variabel - Proses binning (diskritisasi) menjadikan probabilitas dari algoritma naive bayes yang digunakan dapat lebih dihandalkan untuk menentukan kelas dari suatu data.	- *Naïve Bayes* menghasilkan nilai probabilitas yang sangat kecil jika terdapat sangat banyak nilai yang berbeda dalam suatu atribut - Proses diskritisasi menghilangkan beberapa informasi penting yang ada dalam dataset karena teknik ini tidak mempertimbangkan kelas dari suatu data sebelum melewati proses diskritisasi
6.	Intrusion Detection using Naive Bayes Classifier with Feature Reduction	Mengidentifikasi pentingnya pengurangan fitur masukan dalam penerapan IDS, sehingga	Naive Bayes Classifier, Feature Reduction	- Pemilihan metode *Naïve Bayes* karena metode *Bayes*	- Proses klasifikasi dengan dilengkapi metode *Feature*

	(Mukherjee & Sharma, 2012)	diperoleh hasil yang lebih efisien secara komputasi.		beroperasi dengan asumsi independensi yang kuat. - Memberikan hasil akurasi yang lebih baik dengan *Feature Reduction* - *Naïve Bayes* menjadi metode yang lebih dihandalkan dengan *Feature Reduction*	*Reduction* membutuhkan waktu yang lebih lama

4 DASAR TEORI

4.1 Keamanan Jaringan Komputer

Keamanan telah menjadi permasalahan setiap orang. Terkadang bukan mesin atau sistem yang menjadi penyebab masalah, melainkan penyebabnya adalah manusia sendiri. Di masa berkembangan sekarang ini, peralatan teknologi tinggi yang juga dapat menghalangi orang-orang yang ingin merugikan kita, sama seperti kita mengunci pintu kita dari penyusup. Permasalahan keamanan harus dipertimbangkan baik-baik sebagai pokok yang paling mendasar dalam setiap topik diskusi tentang keamanan (Simmonds, Sandilands, & Van Ekert, 2004).

Aktivitas hacking saat ini semakin bervariasi, baik dari segi tekniknya maupun dampak kerusakan yang ditimbulkannya. Para hacker sendiri terus berinovasi dalam melakukan serangan terhadap suatu sistem komputer. Mereka tanpa bosan terus mencari kelemahan dan celah yang bisa dimanfaatkan dari sebuah sistem komputer. Seiring dengan berkembangnya teknologi yang semakin pesat maka semakin banyak pula variasi serangan dari para hacker ini. Kemajuan teknologi saat ini menghadirkan banyak tools-tools bagi para hacker untuk melancarkan serangannya. Aktivitas hacking menjadi sangat mudah sekalipun para ahli keamanan komputer terus memperbaiki celah keamanan yang ditemukan namun tetap saja membendung serangan dari hacker adalah kegiatan yang tidak mudah. Kenyataannya ada aktivitas umum yang dilakukan oleh hacker dalam menyerang suatu sistem keamanan jaringan komputer. Kegiatan ini mulai dari persiapan dalam melakukan penyerangan hingga teknik yang digunakan sesuai dengan kelemahan yang ditemukan dalam suatu sistem komputer. (Agung, M.F., 2011).

Jenis serangan pada jaringan komputer terdiri dari empat kelas utama (Antoon):

1. *Reconnaissance* (Pengintaian)

   Pengintaian adalah tindakan tidak sah dan pemetaan sistem, layanan, atau kerentanan. Pengintai juga bisa dikatakan sebuah fase persiapan sebelum (attacker) melakukan penyerangan, dimana kegiatan intinya adalah mengumpulkan informasi sebanyak mungkin mengenai sasaran. Tindakan yan termasuk pengintaian yaitu:

   - *Packet sniffers*

     Packet sniffers adalah sebuah program yan menangkap atau mengcapture data dari paket yang lewat di jaringan. (*username, password,* dan informasi penting lainnya)

   - *Port scans*

     Port Scanning adalah aktivitas yang dilakukan untuk memeriksa status *port TCP* dan *UDP* pada sebuah mesin.

   - *Ping sweeps*

     Ping sweeps adalah sebuah metode yang dapat menetapkan berbagai alamat IP yang memetakan ke host hidup.

   - *Internet information queries*

2. *Access* (Mengakses) :

   Sistem akses adalah kemampuan penyusup yang tidak berwenang untuk mendapatkan akses ke perangkat, walaupun penyusup tersebut tidak mempunyai akun dan kata sandi. Seseorang yang tidak mempunyai kewenangan untuk mengakses, biasanya ia menjalankan hack, script, atau tool untuk memanfaatkan kerentanan sistem atau aplikasi yang dapat diserang. Tindakan yang termasuk penyusupan yaitu:

   - *Password attacks*

     Password attack dapat diimplementasikan dengan menggunakan beberapa metode termasuk serangan *brute force, dictionary attack,* dan *Trojan horse.*

   - *Trust exploitation*

     Sekumpulan teknik untuk memanipulasi orang sehingga orang tersebut membocorkan informasi rahasia. Meskipun hal ini mirip

dengan permainan kepercayaan atau penipuan sederhana, istilah ini mengacu kepada penipuan untuk mendapatkan informasi atau akses sistem komputer.

- *Man-in-the-middle attacks*

  Man-in-the-middle attacks adalah serangan yang mengharuskan hacker memiliki akses ke paket jaringan yang dating melintasi jaringan. *Man-in-the-middle* dapat diimpelentasikan menggunakan *packet sniffers* jaringan, *routing*, dan *transport protocols.* Kemungkinan penggunaan serangan tersebut adalah pencurian informasi, pembajakan pada sesi yang sedang berlangsung untuk mendapatkan akses ke sumber daya jaringan pribadi, menganalisis lalu lintas untuk mendapatkan informasi tentang jaringan dan penggunanya. penolakan layanan, korupsi data yang dikirim, dan pengenala informasi baru ke dalam sesi jaringan

- *Social engineering*

  Social engineering bisa dilakukan dengan mengelabui anggota sebuah organisasi untuk memberikan informasi yang berharga, seperti lokasi file, dan server, dan password, tetapi lebih mudah dari yang dibayangkan.

- *Phishing*

  Phishing adalah sejenis serangan rekayasa sosial yang melibatkan penggunaan e-mail atau jenis pesan lainnya dalam upaya untuk mengelabui orang lain agar memberikan informasi sensitif, seperti nomor kartu kredit atau kata sandi. Penyerang menyamar sebagai pihak yang terpercaya yang memiliki legitimasi untuk informasi sensitive.

3. Denial of Service (DOS)

Serangan pada *DOS* adalah tindakan menonaktifkan atau merusak jaringan, sistem, atau layanan dengan maksud menolak layanan pengguna. Serangan *DOS* melibatkan kerusakan sistem atau memperlambat sistem sampai sistem tersebut tidak dapat digunakan lagi.

Serangan DOS yang sederhana yaitu meghapus atau merusak informasi, menjalankan *hack* atau *script.*

4. *Malicious Software (Worm, viruses, and Trojan horse)*

   Malicious Software adalah perangkat lunak berbahaya yang masuk ke dalam *host* dengan tujuan merusak sistem, mereplikasi dirinya sendiri, atau menolak layanan atau akses ke jaringan, sistem ataupun layanan.

   Trojan horses adalah replikasi dari virus yang dapat digunakan untuk meminta pengguna masukkan informasi penting pada tampilan layar yang biasa ia percayai. Misalnya penyerang masuk ke layar kotak *windows* dan menjalankan program yang terlihat seperti layar kotak *windows* yang sebenarnya, sehingga pengguna akan mengetikkan *username* dan *password* miliknya.

   Virus adalah program komputer yang dapat menggandakan atau menyalin dirinya sendiri dan menyebar dengan cara menyisipkan salinan dirinya ke dalam program atau dokumen lain.

   Worm adalah jenis virus yang tidak menginfeksi program lainnya. Ia membuat copy dirinya sendiri dan menginfeksi komputer lainnya (biasanya menggunakan hubungan jaringan) tetapi tidak mengkaitkan dirinya dengan program lainnya; akan tetapi sebuah worm dapat mengubah atau merusak file dan program.

4.2 Intrusion Detection System (IDS)

Intrusion Detection adalah proses mengidentifikasi dan menanggapi aktivitas berbahaya yang ditargetkan pada komputasi dan sumber daya jaringan. Sedangkan *Intrusion Detection System* yaitu suatu cara untuk mencoba menemukan konfigurasi yang ada, mengindikasikan, atau dapat mempengaruhi aktivitas berbahaya. *Intrusion Detection System (IDS)* terdiri dari perangkat keras dan perangkat lunak yang bekerja sama untuk menemukan kejadian tak terduga yang mungkin mengindikasikan akan adanya serangan yang terjadi, sedang terjadi, dan telah terjadi.

Tujuan dari *Intrusion Detection System* adalah untuk memberikan indikasi adanya potensi atau serangan nyata. Serangan atau intrusi adalah peristiwa sementara, sedangkan kerentanan merupakan salah satu hal yang membawa potensi serangan atau gangguan. Perbedaan antara Serangan dan kerentanan adalah bahwa serangan terjadi pada waktu tertentu, sementara kerentanan ada secara independen dari waktu pengamatan (Amoroso & Kwapniewski, 1998).

4.3 Text Mining

Text mining, yang juga disebut sebagai Teks Data Mining (TDM) atau Knowledge Discovery in Text (KDT), secara umum mengacu pada proses ekstraksi informasi dari dokumen-dokumen teks tak terstruktur (unstructured). Text mining dapat didefinisikan sebagai penemuan informasi baru dan tidak diketahui sebelumnya oleh komputer, yang secara otomatis mengekstrak informasi dari sumber-sumber teks tak terstruktur yang berbeda. Kunci dari proses ini adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Nugroho, 2011).

Tujuan utama text mining adalah mendukung proses knowledge discovery pada koleksi dokumen yang besar. Pada prinsipnya, text mining adalah bidang ilmu multidisipliner, melibatkan information retrieval (IR), text analysis, information extraction (IE), clustering, categorization, visualization, database technology, natural language processing (NLP), machinelearning, dan data mining. Dapat pula dikatakan bahwa text mining merupakan salah satu bentuk aplikasi kecerdasan buatan (artificial intelligence / AI) (Nugroho, 2011).

Text mining mencoba memecahkan masalah information overload dengan menggunakan teknik-teknik dari bidang ilmu yang terkait. Text mining dapat dipandang sebagai suatu perluasan dari data mining atau knowledge-discovery in database (KDD), yang mencoba untuk menemukan pola-pola menarik dari basis data berskala besar. Namun text mining

memiliki potensi komersil yang lebih tinggi dibandingkan dengan data mining, karena kebanyakan format alami dari penyimpanan informasi adalah berupa teks. Text mining menggunakan informasi teks tak terstruktur dan mengujinya dalam upaya mengungkap struktur dan arti yang tersembunyi di dalam teks (Nugroho, 2011).

Ada empat tahap proses pokok dalam text mining, yaitu pemrosesan awal terhadap teks (text preprocessing), transformasi teks (text transformation), pemilihan fitur (feature selection), dan penemuan pola (pattern discovery) (Nugroho, 2011).

a. Text Preprocessing

   Tahap ini melakukan analisis semantik (kebenaran arti) dan sintaktik (kebenaran susunan) terhadap teks. Tujuan dari pemrosesan awal adalah untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan lebih lanjut. Operasi yang dapat dilakukan pada tahap ini meliputi part-of-speech (PoS) tagging, menghasilkan parse tree untuk tiap-tiap kalimat, dan pembersihan teks.

b. Text Transformation

   Transformasi teks atau pembentukan atribut mengacu pada proses untuk mendapatkan representasi dokumen yang diharapkan. Pendekatan representasi dokumen yang lazim digunakan oleh model "bag of words" dan model ruang vector (vector space model). Transformasi teks sekaligus juga melakukan pengubahan kata-kata ke bentuk dasarnya dan pengurangan dimensi kata di dalam dokumen. Tindakan ini diwujudkan dengan menerapkan stemming dan menghapus stop words.

c. Feature Selection

   Pemilihan fitur (kata) merupakan tahap lanjut dari pengurangan dimensi pada proses transformasi teks. Walaupun tahap sebelumnya sudah melakukan penghapusan kata-kata yang tidak deskriptif (stopwords), namun tidak semua kata-kata di dalam dokumen memiliki

arti penting. Oleh karena itu, untuk mengurangi dimensi, pemilihan hanya dilakukan terhadap kata-kata yang relevan yang benar-benar merepresentasikan isi dari suatu dokumen. Ide dasar dari pemilihan fitur adalah menghapus kata-kata yang kemunculannya di suatu dokumen terlalu sedikit atau terlalu banyak.

d. Pattern Discovery

Pattern discovery merupakan tahap penting untuk menemukan pola atau pengetahuan (knowledge) dari keseluruhan teks. Tindakan yang lazim dilakukan pada tahap ini adalah operasi text mining, dan biasanya menggunakan teknik-teknik data mining. Dalam penemuan pola ini, proses text mining dikombinasikan dengan proses-proses data mining. Masukan awal dari proses text mining adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai hasil interpretasi atau evaluasi. Apabila hasil keluaran dari penemuan pola belum sesuai untuk aplikasi, dilanjutkan evaluasi dengan melakukan iterasi ke satu atau beberapa tahap sebelumnya. Sebaliknya, hasil interpretasi merupakan tahap akhir dari proses text mining dan akan disajikan ke pengguna dalam bentuk visual (Nugroho, 2011).

4.4 Ekstraksi Dokumen

Cara yang digunakan dalam mempelajari suatu data teks, adalah dengan terlebih dahulu menentukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Sebelum menentukan fitur-fitur yang mewakili, diperlukan tahap preprocessing yang dilakukan secara umum dalam teks mining pada dokumen, yaitu case folding, tokenizing, filtering, stemming, tagging dan analyzing.

a. Case folding

Case folding adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf “a” sampai dengan “z” yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.

b. Tokenizing

Tahap tokenizing / parsing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.

c. Filtering

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma stoplist (membuang kata yang kurang penting) atau wordlist CASE FOLDING TOKENIZING FILTERING STEMMING (menyimpan kata penting). Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah “yang”, “dan”, “di”, “dari”, dan seterusnya.

d. Stemming

Tahap stemming adalah tahap mencari root kata dari tiap kata hasil filtering. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa nggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki rumus bentuk baku yang permanen (Nugroho, 2011).

4.5 Algoritma TF-IDF

Metode TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada *information retrieval*. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat (Robertson, 2004). Metode ini akan menghitung nilai *Term Frequency (TF)* dan *Inverse Document Frequency (IDF)* pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus:

$$W_{dt} = tf_{dt} * IDF_t$$

Dimana :

d : dokumen ke-d

t : kata ke-t dari kata kunci

W : bobot dokumen ke-d terhadap kata ke-t

tf : banyaknya kata yang dicari pada sebuah dokumen

IDF : Inversed Document Frequency

Nilai *IDF* didapatkan dari

IDF : *log2 (D/df)*

dimana

D : total dokumen

df : banyak dokumen yang mengandung kata yang dicari

Setelah bobot *(W)* masing-masing dokumen diketahui, maka dilakukan proses pengurutan dimana semakin besar nilai *W*, semakin besar tingkat similaritas dokumen tersebut terhadap kata kunci, demikian sebaliknya.

4.6 Named Entity Recognition (NER)

Named Entity Recognition (NER) adalah salah satu bagian penting dari Natural Pengolahan Bahasa (NLP). NER digunakan untuk menemukan dan mengklasifikasikan ungkapan mengenai arti khusus dalam teks yang ditulis dalam bahasa alami. NER umumnya digunakan untuk mendeteksi nama orang, nama tempat dan organisasi dari sebuah dokumen, dan sering menjadi informasi penting dalam teks. NER dapat digunakan untuk berbagai tugas penting, yaitu dapat menjadi alat untuk penelusuran dan penyaringan teks. NER juga bisa digunakan sebagai alat preprocessing untuk tugas NLP lainnya. Dalam NLP, NER biasa digunakan untuk Machine Translation, Question Answering, Text Summarization, Language Modeling atau Sentiment Analisis (Konkol, Brychcín, & Konopík, 2015).

4.7 Naïve Bayes

Metode NBC menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan vocabulary, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang

sedapat mungkin menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas prior bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi.

Lebih konkritnya jika diasumsikan dimiliki koleksi dokumen $D = d_i \mid i=1,2,\ldots|D|\}=\{d_1,d_2,\ldots,d_{|D|}\}$ dan koleksi kategori $V=\{v_j \mid j=1,2,\ldots|V|\}=\{v_1,v_2,\ldots,v_{|V|}\}$. Klasifikasi NBC dilakukan dengan cara mencari probabilitas $P(V=v_j \mid D=d_i)$, yaitu probabilitas category v_j jika diketahui dokumen d_i . Dokumen d_i dipandang sebagai tuple dari kata-kata dalam dokumen, yaitu $<a_1,a_2,\ldots a_n>$, yang frekuensi kemunculannya diasumsikan sebagai variable random dengan distribusi probabilitas Bernoulli (McCallum & Nigam, 1998). Dalam klasifikasi teks, tujuan kita adalah untuk menemukan kelas terbaik untuk dokumen. Kelas terbaik dalam klasifikasi NB adalah yang paling besar kemungkinannya atau kelas *maximum a posteriori* (MAP):

$$V_{map} = \arg\max_{vj \in V} P(v_j \mid a_{1,} a_{2,} \ldots, a_n) \quad (1)$$

Teorema Bayes menyatakan tentang probabilitas bersyarat menyatakan :

$$P(B|A) = \frac{P(A \mid B)P(B)}{P(A)} \quad (2)$$

Dengan menerapkan teorema Bayes persamaan (1) dapat ditulis :

$$V_{map} = \arg\max_{vj \in V} \frac{P(a_{1,}a_{2,} \ldots, a_{n,}|v_j)\, P(v_{j,})}{P(a_{1,}a_{2,} \ldots, a_n)} \quad (3)$$

Karena nilai $P(a_{1,}a_{2,} \ldots, a_n)$ untuk semua v_j besarnya sama maka nilainya dapat diabaikan, sehingga persamaan (3) menjadi :

$$V_{map} = \arg\max_{vj \in V} P(a_{1,} a_{2,} \ldots, a_n \mid v_j) P(v_j) \quad (4)$$

Dengan mengasumsikan bahwa setiap kata dalam <a_1, a_2,…,a_n> adalah independent, maka $P(a_{1,}a_{2,}\dots,a_n\,|v_j\,)$ dalam persamaan (4) dapat ditulis sebagai :

$$P(a_{1,}a_{2,}\dots,a_n\,|v_j\,) = \prod_i P(a_i \mid v_j) \quad (5)$$

Sehingga persamaan (4) dapat ditulis :

$$V_{map} = \underset{vj \in V}{\arg\max}\, P\,(v_j) \prod_i P(a_i \mid v_j) \quad (6)$$

Nilai $P(v_j)$ ditentukan pada saat pelatihan, yang nilainya didekati dengan :

$$P(v_j) = \frac{|doc_j|}{|\text{Contoh}|} \quad (7)$$

Dimana $|doc_j|$ adalah banyaknya dokumen yang memiliki kategori j dalam pelatihan, sedangkan $|\text{Contoh}|$ banyaknya dokumen dalam contoh yang digunakan untuk pelatihan. Untuk nilai $P(w_k|v_j)$ yaitu probabilitas kata w_k dalam kategori j ditentukan dengan :

$$P(w_k|v_j) = \frac{n_k + 1}{n + |\text{vocabulary}|} \quad (8)$$

Dimana n_k adalah frekuensi munculnya kata w_k dalam dokumen yang ber kategori v_j, sedangkan nilai n adalah banyaknya seluruh kata dalam dokumen berkategori v_j, dan vocabulary adalah banyaknya kata dalam contoh pelatihan.

4.8 Evaluasi

Supervised Machine Learning memiliki beberapa cara untuk mengevaluasi kinerja algoritma pembelajaran dan pengklasifikasi yang mereka hasilkan. Pengukuran kualitas klasifikasi yang dibangun dari *"confusion matrix"* yang mencatat contoh yang dikenali secara benar dan salah oleh untuk masing-masing kelas (Sokolova, Japkowicz, & Szpakowicz, 2006). Tabel dibawah ini menyajikan "confusion matrix" untuk

klasifikasi biner, di mana tp adalah true positive, fp - false positive, fn - false negative, dan tn - true negative.

Tabel 2 Confusion Matrix

Class \ Recognized	Positive	Negative
Positive	Tp	fn
Negative	Fp	tn

Pengukuran evaluasi kinerja yang sering diterima adalah *accuracy*. *Accuracy* tidak membedakan antara jumlah label yang benar dari kelas yang berbeda. *Accuracy* dapat dihitung dengan persamaan (8) (Sokolova et al., 2006)

$$accuracy = \frac{tp+tn}{tp+fp+fn+tn} \qquad (8)$$

Pengukuran kinerja lain yang dipilih untuk kelas positif adalah *precision, recall*, dan *F – measure*. *Precision* dihitung dengan persamaan berikut

$$precision = \frac{tp}{tp+fp} \qquad (9)$$

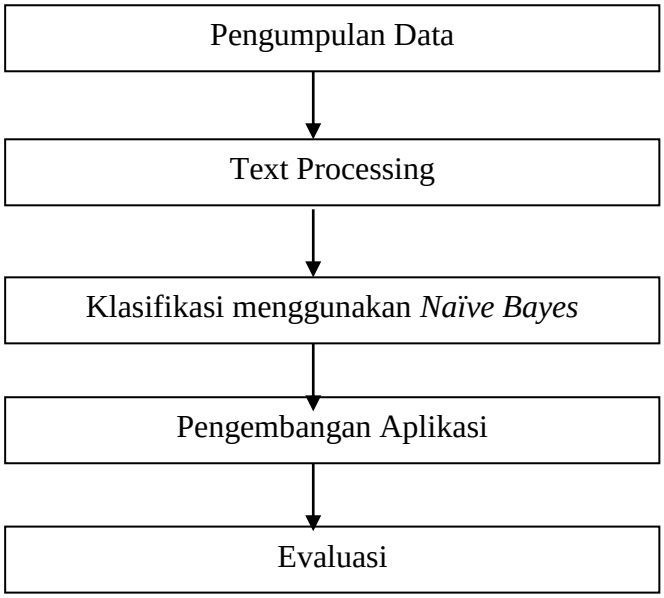
Precision adalah sebuah fungsi dari contoh positive yang diklasifikasikan dengan benar (*true positive*) dan contoh bernilai negative salah diklasifikasikan sebagai positive (*false positive*). (Sokolova et al., 2006).

$$recall = \frac{tp}{tp+fn} \qquad (10)$$

Recall adalah sebuah fungsi dari contoh positive yang diklasifikasikan dengan benar (*true positive*) dan contoh positive yang salah diklasifikasikan sebagai negative (*false negative*). (Sokolova et al., 2006).

5 METODOLOGI

Penelitian ini dilakukan dengan beberapa tahapan yang ditunjukkan pada gambar di bawah ini.



Gambar 1 Metodologi Penelitian

5.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini didapatkan dari laman www.sciencedirect.com, ieeexplore.ieee.org, dan penelitan diluar kedua website tersebut. Data berisi dokumen jurnal penelitan tentang serangan pada *layer network* dari tahun 2014 -2017. Data yang digunakan tediri dari 7 kategori, yaitu 6 kategori serangan yang mungkin terjadi pada *layer network,* dan 1 kategori jurnal serangan secara umum. Jumlah data yang diambil sekitar 800 buah dokumen jurnal. Setelah data didapat, maka akan dilakukan proses *preprocessing.*

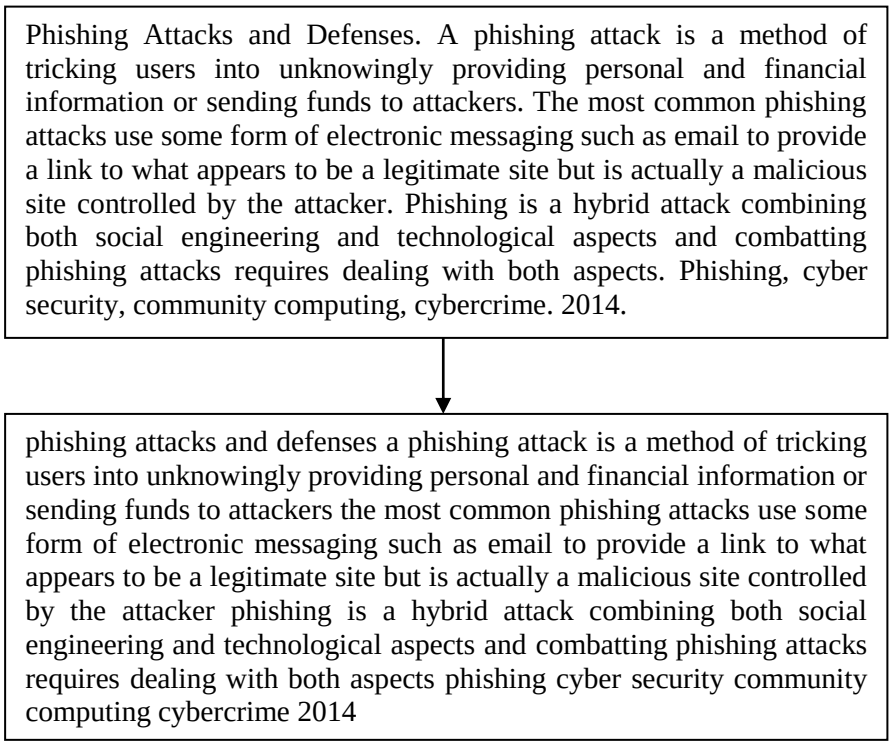
5.2 Text Processing

Proses *preprocessing* pada teks dokumen mencangkup *Case Folding, Tokenization, Filtering/Stopword Removel, Stemmer*. Setelah dilakukan

preprocessing kemudian menghitung *tf-idf* untuk pembobotan setiap kata. Setelah didapat hasil *preprocessing* kemudian menggunakan metode *Named Entity Recognition* untuk mengenali entitas tahun sehingga masing-masing dokumen dapat diketahui informasi tahun .

5.2.1 Case Folding

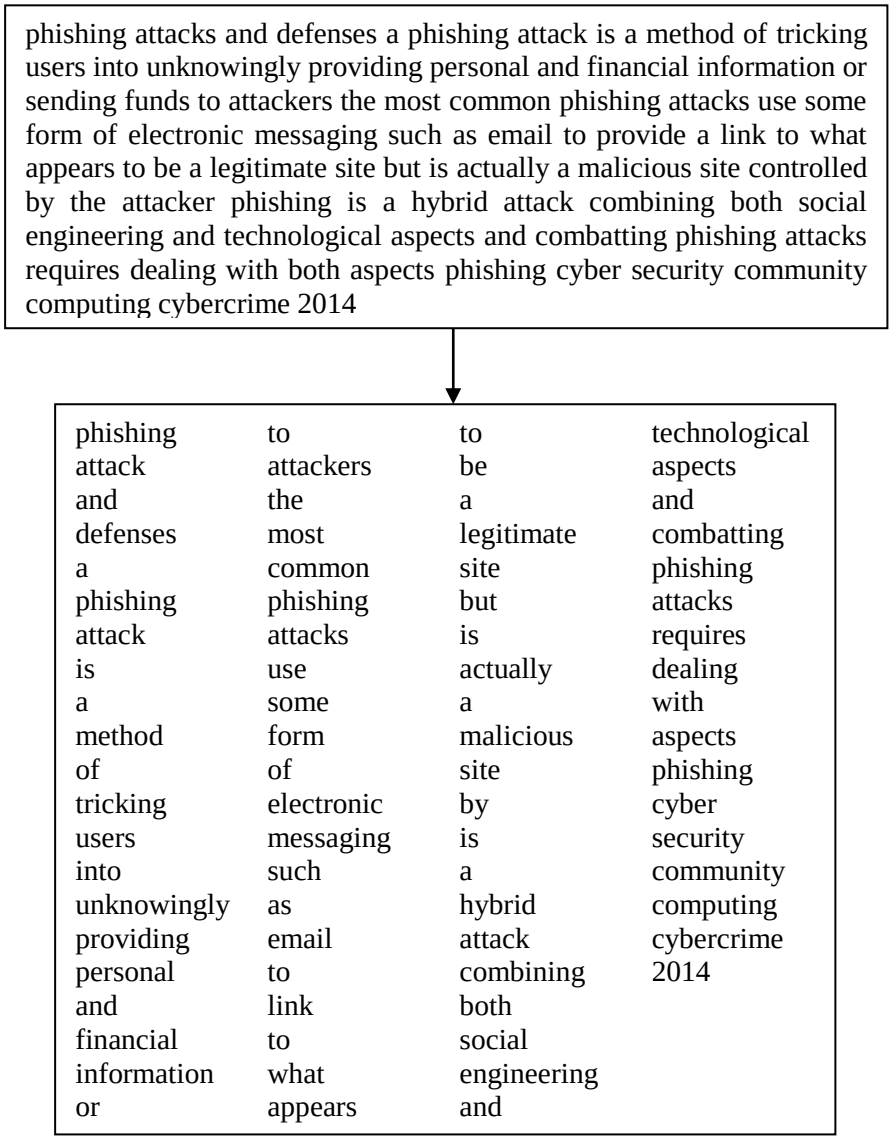
Case Folding dilakukan dengan mengubah semua huruf dalam dokumen jurnal (judul, abstrak, dan keyword) menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter. Berikut contoh dari case folding.



Gambar 2 Contoh Penerapan Case Folding

5.2.2 Tokenization

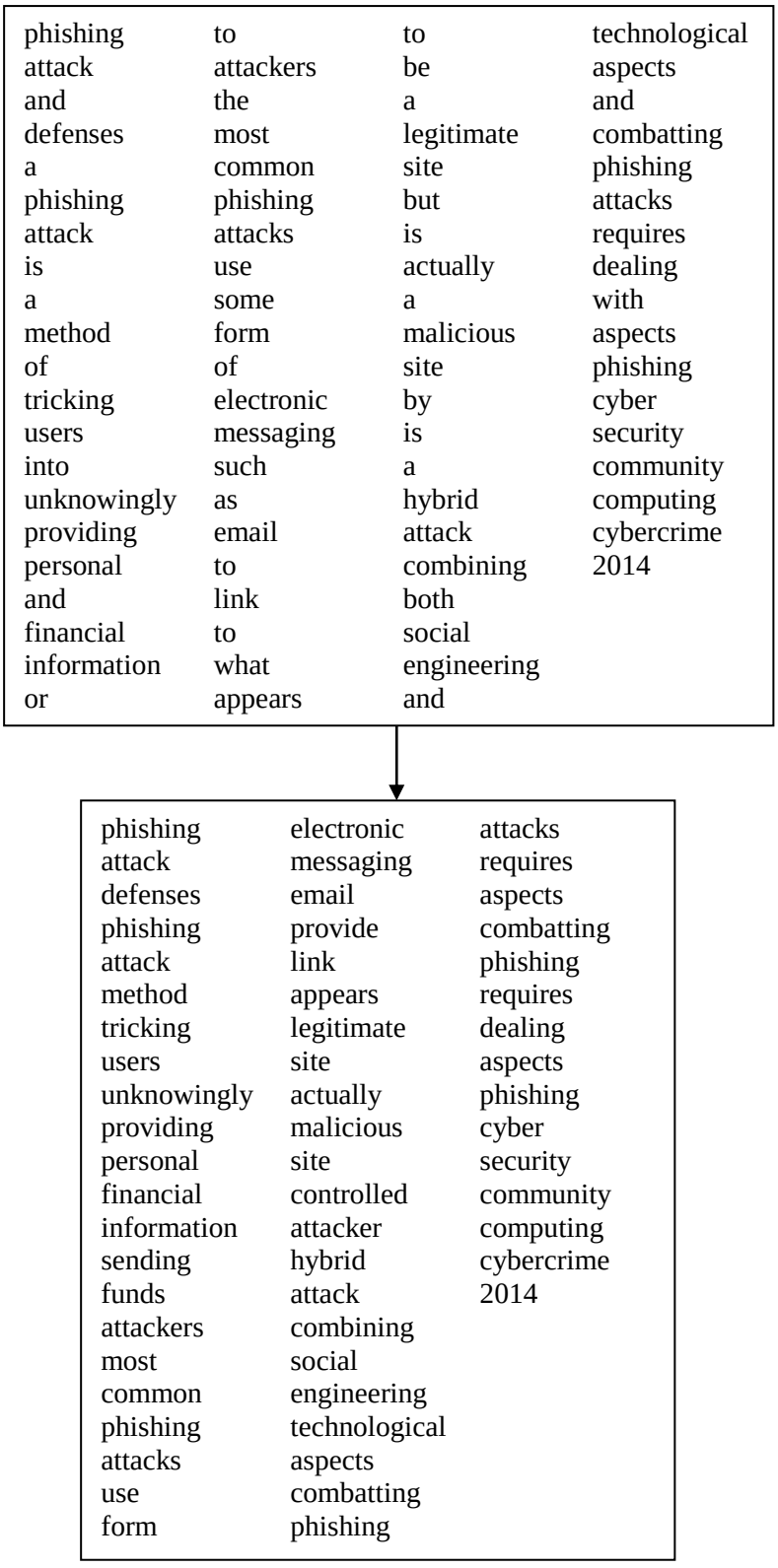
Tahap tokenization dilakukan pemotongan string input berdasarkan tiap kata yang menyusunnya. Berikut merupakan contoh dari penerapan tokenization.



Gambar 3 Contoh Penerapan Tokenization

5.2.3 Filtering/Stopword Removel

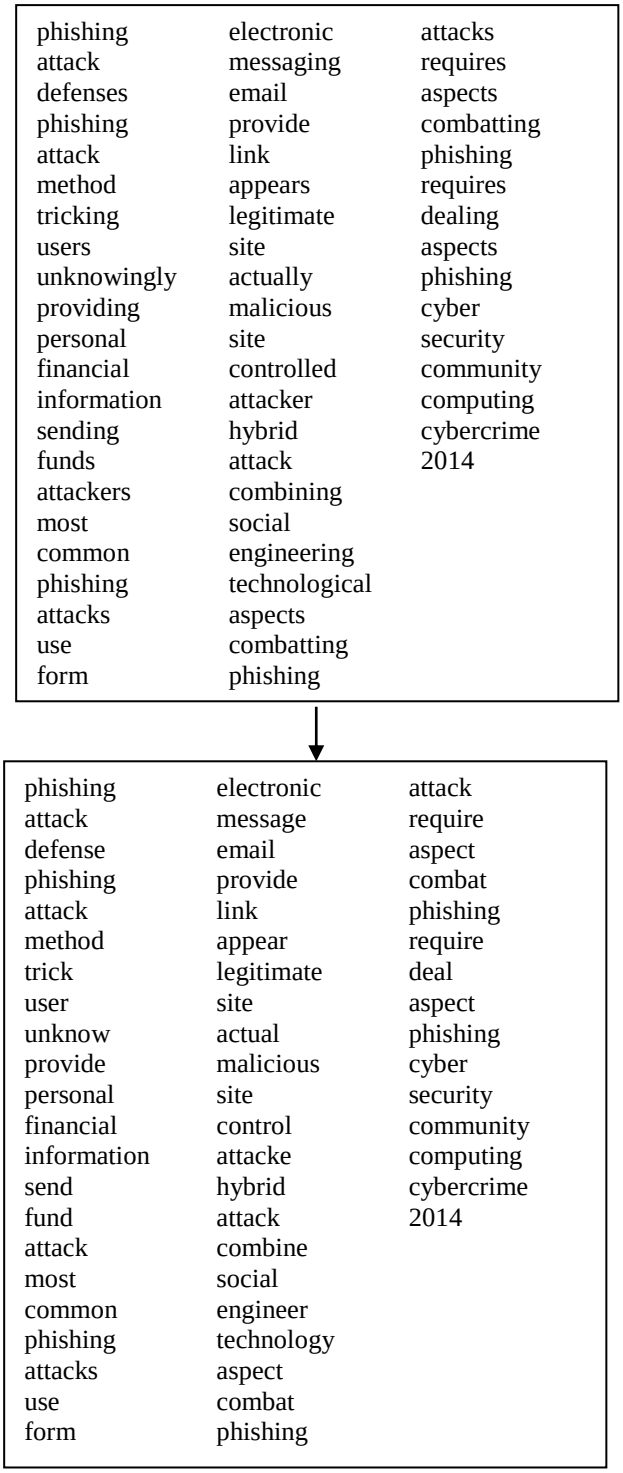
Tahap *filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bs menggunakan algoritma stoplist (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting).



Gambar 4 Contoh Penerapan Filtering/Stopword Removel

5.2.4 Stemming

Tahap *stemming* dilakukan dilakukan dengan tahap mencari *root* kata dari tiap kata. Berikut adalah contoh penerapan stemming.



Gambar 5 Contoh Penerapan Stemming

5.2.5 Penerapan Algoritma *TF-IDF*

Dalam melakukan pembobotan dengan *TF-IDF*, kata yang akan digunakan adalah kata-kata penting dalam dokumen tersebut, sehingga sebelum dilakukan pembobotan maka setiap dokumen harus di lakukan preprocessing terlebih dahulu. Metode ini akan menghitung bobot setiap kata pada dokumen d dengan rumus:

$W_{dt} = tf_{dt} * IDF_t$

Dimana :

d : dokumen ke-d

t : kata ke-t dari kata kunci

W : bobot dokumen ke-d terhadap kata ke-t

tf : banyaknya kata yang dicari pada sebuah dokumen

IDF : Inversed Document Frequency

Nilai *IDF* didapatkan dari

IDF : *log2 (D/df)*

dimana

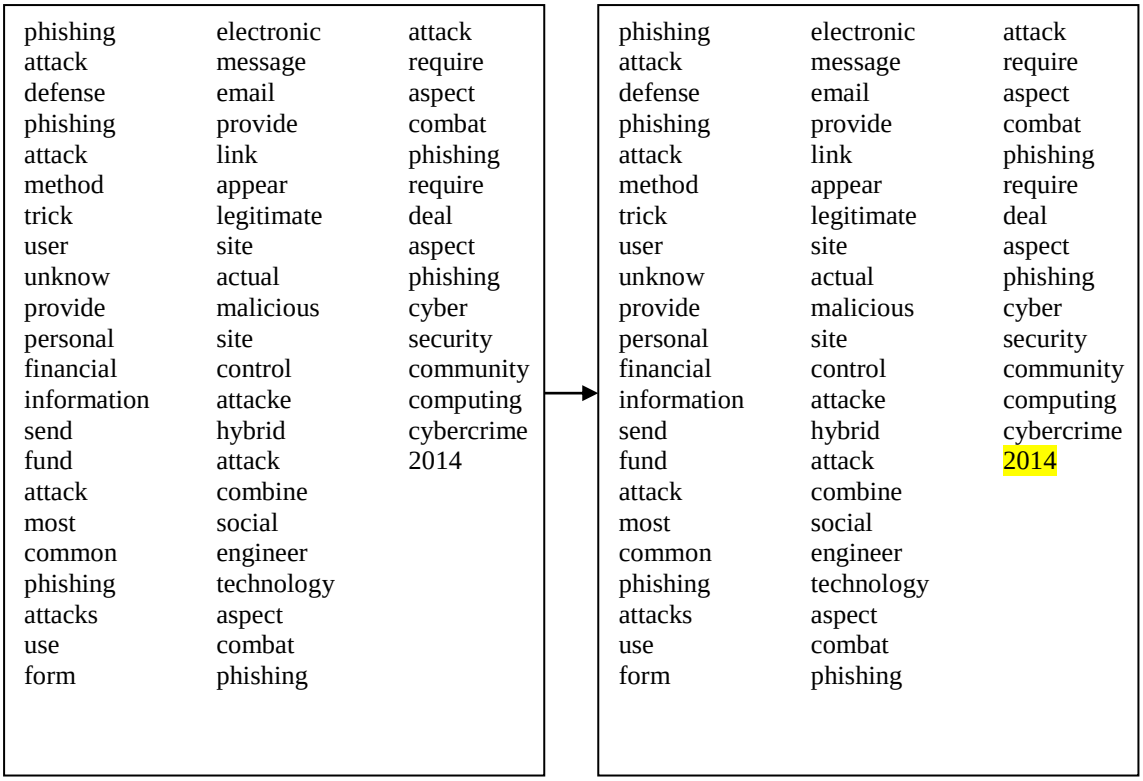
D : total dokumen

df : banyak dokumen yang mengandung kata yang dicari

Setelah bobot *(W)* masing-masing dokumen diketahui, maka dilakukan proses pengurutan dimana semakin besar nilai *W*, semakin besar tingkat similaritas dokumen tersebut terhadap kata kunci, demikian sebaliknya.

5.2.6 Penerapan Metode Named Entity Recognition

Pada tahap ini dilakukan penerapan *Named Entity Recognition* untuk pengenalan entitas tahun. Sehingga setiap dokumen dapat diketahui tahunnya.



Berdasarkan data tersebut maka diperoleh informasi bahwa jurnal tersebut adalah jurnal tahun 2014.

5.3 Klasifikasi Menggunakan Naïve Bayes

Penerapan Metode Naïve Bayes dilakukan dengan perhitungan untuk mengklasifikasikan dokumen serangan pada *Layer Network* ke dalam 7 kelas yaitu kelas DOS, spoofing, probing, phishing, malicious code, man-in-the-middle, dan general, dengan langkah sebagai berikut:

$$V_{map} = \arg\max_{vj \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Dengan Vmap adalah kelas terbaik dalam suatu dokumen, nilai $P(v_j)$ ditentukan pada saat pelatihan, yang nilainya didekati dengan :

$$\mathrm{P}(v_j) = \frac{|doc_j|}{|\mathrm{Contoh}|}$$

dimana :

$P(v_j)$ adalah probabilitas prior dari dokumen yang ada pada kategori v_j

$|doc_j|$ adalah jumlah dokumen yang memiliki kategori j dalam pelatihan

|Contoh| adalah jumlah dokumen dalam contoh yang digunakan untuk pelatihan.

Untuk nilai $P(w_k|v_j)$ yaitu probabilitas kata w_k dalam kategori j ditentukan dengan :

$$P(w_k|v_j) = \frac{n_k + 1}{n + |\text{vocabulary}|}$$

dimana :

$P(w_k|v_j)$ adalah probabilitas bersyarat dari kata w_k yang terjadi dalam sebuah dokumen v_j.

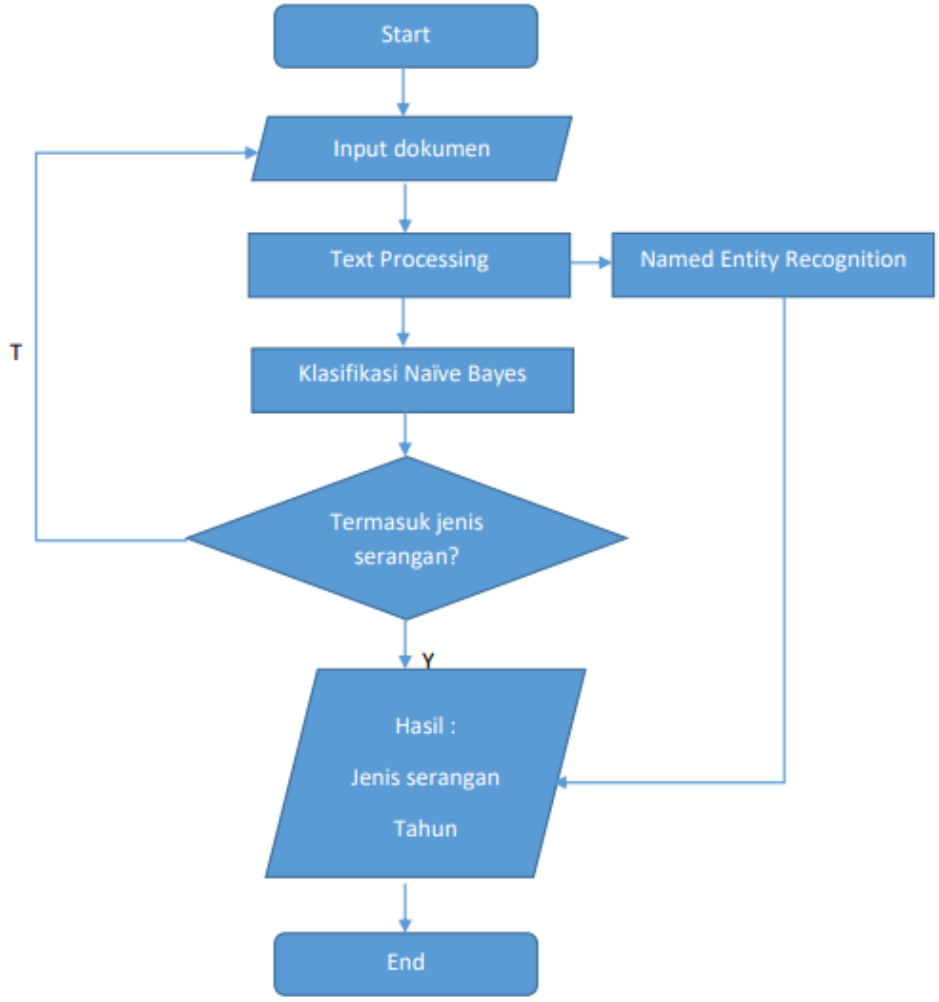
n_k adalah frekuensi munculnya kata w_k dalam dokumen yang berkategori v_j.

n adalah banyaknya seluruh kata dalam dokumen berkategori v_j.

vocabulary adalah banyaknya kata dalam contoh pelatihan.

5.4 Pengembangan Aplikasi

Aplikasi dikembangkan menggunakan bahasa pemrograman PHP dengan database MySQL. Pada tahap ini akan dilakukan pengembangan program sesuai dengan tujuan disertai dengan desain *interface* yang mendukung jalannya program untuk proses deteksi jenis serangan yang dibahas dalam setiap dokumen. Alur program digambarkan menggunakan flowchart dibawah ini:



Gambar 6 Flowchart Program

Keterangan :

a. Input dokumen dengan format PDF.
b. Text Processing terdiri dari : *case folding, tokenisasi, filtering, stemmer,* penerapan *algoritma tf-idf.*
c. Named Entity Recognition : Mengenali entitas tahun pada dokumen yang sudah dilakukan *preprocessing.*
d. Klasifikasi *Naïve Bayes* untuk mengidentifikasi jenis serangan pada dokumen.
e. Hasil informasi dari dokumen yang telah melalui proses berupa jenis serangan yang dibahas dan tahun.

5.5 Evaluasi

Evaluasi kinerja klasifikasi dokumen akan dihitung menggunakan pengukuran *accuracy*, *precision*, *recall*, *F1-score*, dan kurva ROC dengan perhitungan menggunakan persamaan (8), (9), dan (10).

6 JADWAL PELAKSANAAN

Berikut adalah jadwal penelitian yang akan dilakukan.

No.	Kegiatan	Bulan															
		September				Oktober				November				Desember			
		Minggu ke-															
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Pengumpulan Data																
2	Text Preprocessing																
3	Ekstrasi Informasi : Named Entity Recognition																
4	Perhitungan Klasifikasi dengan Metode Naïve Bayes																
5	Pengembangan Aplikasi																
6	Pengujian																

DAFTAR PUSTAKA

Amoroso, E., & Kwapniewski, R. 1998. ***A selection criteria for intrusion detection systems***. Paper presented at the Computer Security Applications Conference, 1998. Proceedings. 14th Annual.

Antoon, R. ***Network Security 1 and 2 Companion Guide***: Pearson Education India.

Azis, M. S., & Adiwijaya, B. M. Deteksi Anomaly pada Intrusion Detection System (IDS) menggunakan Backpropagation Termodifikasi.

Dewi, E. K., & Azhari, S. 2013. ***ANALISIS KEAMANAN SISTEM PERANGKAT LUNAK***. Paper presented at the Seminar Nasional Aplikasi Teknologi Informasi (SNATI).

Godse, V. 2010. Building an ecosystem for cyber security and data protection in india. ***Ethics and Policy of Biometrics***: 138-145.

Hamzah, A. 2012. ***Klasifikasi teks dengan naïve bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis***. Paper presented at the Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III.

Inyaem, U., Meesad, P., & Haruechaiyasak, C. 2009. ***Named-entity techniques for terrorism event extraction and classification***. Paper presented at the Natural Language Processing, 2009. SNLP'09. Eighth International Symposium on.

Konkol, M., Brychcín, T., & Konopík, M. 2015. Latent semantics in named entity recognition. ***Expert Systems with Applications***, 42(7): 3470-3479.

Li, S., Li, S., Xia, C., & Zhang, W. 2011. ***Topic tracking based on Naive bayes***. Paper presented at the Computer Science and Network Technology (ICCSNT), 2011 International Conference on.

Lumbanraja, F. R. 2013. Sistem Pencarian Data Teks dengan Menggunakan Metode Klasifikasi Rocchio (Studi Kasus: Dokumen Teks Skripsi). ***Prosiding SEMIRATA 2013***, 1(1).

Maarif, A. A. 2015. Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah. ***Teknik Informatika Universitas Dian Nuswantoro, Semarang***.

Mahalakshmi, G., Antony, J., & Roshini, S. 2016. Domain Based Named Entity Recognition Using Naive Bayes Classification.

Manisha, D. 2014. Mukesh Kumar, Network Layer Attacks and Their Countermeasures in Manet: A Review. ***IOSR Journal of Computer Engineering***, 16(2): 113-116.

McCallum, A., & Nigam, K. 1998. ***A comparison of event models for naive bayes text classification***. Paper presented at the AAAI-98 workshop on learning for text categorization.

McClure, S., Scambray, J., Kurtz, G., & Kurtz. 2009. Hacking exposed: network security secrets and solutions.

Mukherjee, S., & Sharma, N. 2012. Intrusion detection using naive Bayes classifier with feature reduction. ***Procedia Technology***, 4: 119-128.

Nugroho, E. 2011. Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp. ***Jurusan Ilmu Komputer, Universitas Muhammadiyah Malang***.

Pfleeger, C. P., & Pfleeger, S. L. 2012. ***Analyzing computer security: a threat/vulnerability/countermeasure approach***: Prentice Hall Professional.

Rish, I. 2001. ***An empirical study of the naive Bayes classifier***. Paper presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence.

Robertson, S. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. ***Journal of documentation***, 60(5): 503-520.

Sazali, S. S., Rahman, N. A., & Bakar, Z. A. 2016. ***Information extraction: Evaluating named entity recognition from classical Malay documents***. Paper presented at the Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on.

Simmonds, A., Sandilands, P., & Van Ekert, L. 2004. ***An ontology for network security attacks***. Paper presented at the Asian Applied Computing Conference.

Soderland, S. 1999. Learning information extraction rules for semi-structured and free text. ***Machine learning***, 34(1): 233-272.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. 2006. ***Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation***. Paper presented at the Australian conference on artificial intelligence.

Triawati, C. 2009. Text mining. ***Retrieved May***, 19: 2015.

Wirawan, I. N. T., & Eksistyanto, I. 2015. Penerapan Naive Bayes Pada Intrusion Detection System Dengan Diskritisasi Variabel. ***JUTI: Jurnal Ilmiah Teknologi Informasi***, 13(2): 182-189.

Wulandini, F., & Nugroho, A. S. 2009. ***Text Classification Using Support Vector Machine for Web mining Based Spation Temporal Analysis of the Spread of Tropical Diseases***. Paper presented at the International Conference on Rural Information and Communication Technology.