Jurnal Integrasi, vol. 6, no. 1, 2014, 11-20
ISSN: 2085-3858 (print version)

Article History
Received 10 February 2014
Accepted 11 March 2014

Aplikasi Pendeteksi Opini Spam

Hilda Widyastuti[1)], Ari Wibowo[2)]
Batam State Polytechnic
Informatic Engineering Study Program
Parkway Street, Batam Centre, Batam 29461, Indonesia
Email : hilda@polibatam.ac.id [1)]

Batam State Polytechnic
Network Multimedia Study Program
Parkway Street, Batam Centre, Batam 29461, Indonesia
Email : wibowo@polibatam.ac.id [2)]

Abstrak

Opinion mining bertujuan untuk menentukan apakah suatu opini termasuk kategori opini positif, opini negatif, atau netral. Penerapan *opinion mining* pada sekumpulan opini pembaca menghasilkan rangkuman persentase jumlah opini positif, opini negatif, atau netral pembaca. Dalam konteks berita politik online, rangkuman ini bermanfaat bagi para pembaca untuk lebih memahami situasi politik yang berubah sangat cepat dan bermanfaat bagi pengelola berita online untuk mengetahui respon dari masyarakat terhadap berita yang disajikan. Opini pembaca ini juga bisa digunakan untuk mempengaruhi opini publik, mengakibatkan munculnya opini spam yang menyesatkan. Opini spam perlu dideteksi pada *opinion mining*, yang selanjutnya akan opini spam akan diabaikan dalam *opinion mining*.

Penelitian ini bertujuan mendeteksi opini spam. Penelitian ini menggunakan metode pengecekan duplikasi komentar dan pendeteksian perilaku yang menyimpang dengan *support unexpectedness*. Pengecekan duplikasi komentar dilakukan dengan membandingkan setiap komentar yang ada di semua artikel dengan menggunakan *cosine similarity*. Penelitian ini berhasil menemukan opini spam dengan kedua metode tersebut.

Kata kunci : deteksi spam, *opinion mining,* pengecekan duplikasi, *support unexpectedness*

Abstract

Opinion mining aims to determine an opinion category, positive opinion, negative opinion, or neutral. Opinion mining implementation on public opinions, produces a percentage summary of positive opinions, negative opinions, or neutral from reader. In online political news context, this summary has advantage for readers to understand political situation which changed very fast and has advantage for online news manager to know respond from society towards presented news. Actually reader's opinion also can be used to influenced public opinion, motivate appearance mislead spam opinion. Spam opinion must be detected in opinion mining, further spam opinion will be ignored.

This research purpose is opinion spam detection. This research used two methods, comments duplication checking and aberrant behavior detection. Duplication checking is done by comparing each existing comments on all articles by using the cosine similarity. This research can find some spam opinions by those methods.

Keywords: spam detection, opinion mining, duplicate checking, support unexpectedness

1 Pendahuluan

Keberadaan media sosial seperti Facebook dan Twitter yang sangat digemari oleh masyarakat, menghasilkan banyak data berupa status, komentar terhadap status yang diposting, foto, video, message, dan lain-lain. Media sosial telah digunakan sebagai salah satu sarana interaksi antar individu. Selain itu, keberadaan media sosial juga dipakai oleh perusahaan untuk mempromosikan produknya. Perusahaan-perusahaan membuat *fan page* di Facebook dan Twitter, dan pemakai Facebook dan Twitter bisa memberikan umpan balik kepada *fan page* tersebut dalam bentuk status atau komentar.

Keberadaan media sosial menyebabkan terkumpulnya data yang sangat besar. Tantangannya adalah bagaimana mendayagunakan data tersebut menjadi pengetahuan yang mempunyai nilai tambah. Salah satu cara mendayagunakan data tersebut menggunakan *data mining*. Salah satu bagian dari *data mining* adalah *text mining*, yaitu *data mining* yang mengolah data yang berbentuk teks yang tidak terstruktur. Untuk memudahkan orang menggunakan opini publik, telah dikembangkan *opinion mining*, yang dapat menentukan apakah suatu opini itu termasuk kategori opini positif, opini negatif, atau opini yang bersifat netral. Selain di media sosial, masyarakat juga sering memberikan opini pada berita-berita di media online seperti www.detik.com, www.kompas.com, dan web site jual beli seperti www.amazon.com. Berdasarkan pengamatan penulis, masyarakat Indonesia senang memberikan komentar pada berita-berita politik.

Ada beberapa tujuan pembaca media online memberikan komentar pada berita-berita politik yaitu untuk mengekspresikan aspirasi mereka tentang kondisi politik yang dinamis yang terjadi saat ini, mempengaruhi pembaca lainnya untuk mengikuti opininya, dan menangkis isi pemberitaan pada media online yang dirasakannya kurang tepat. Pembaca surat kabar juga senang membaca komentar dari pembaca lainnya, supaya mengerti situasi politik yang berubah sangat cepat. Dalam sudut pandang sumber berita dan pengelola berita, mereka mempunyai kebutuhan mengumpulkan umpan balik dari masyarakat terhadap berita-berita yang mereka muat.

Keberadaan opini pembaca yang dapat mempengaruhi pembaca lainnya, mendorong munculnya opini spam yang menyesatkan para pembaca karena opini yang diberikan merupakan opini palsu. Pada *opinion mining* perlu diketahui opini yang merupakan opini spam dan opini bukan spam. Setelah diketahui, opini spam akan diabaikan dalam *opinion mining*.

Menurut [7] ada tiga jenis opini spam yaitu (1) opini yang sebenarnya tidak dapat dipercaya, tetapi sangat meyakinkan, (2) opini yang mengarah pada merk atau produsen atau penjual produk yang berbeda dengan objek yang dibicarakan, (3) komentar yang bukan opini, karena berisi iklan atau opini yang tidak relevan misalnya berisi pertanyaan, jawaban, atau teks-teks acak. Opini jenis kedua dan ketiga lebih jelas, sehingga mudah ditebak dan ditentukan sendiri oleh admin, sedangkan penentuan opini spam jenis pertama sulit dilakukan secara manual.

Rumusan masalah penelitian ini adalah bagaimana cara merancang dan membangun aplikasi yang dapat menentukan apakah suatu opini berbahasa Indonesia termasuk opini spam atau tidak. Batasan masalah penelitian ini adalah menggunakan opini berbahasa Indonesia

2 Landasan Teori

2.1 *Data mining*

Data mining muncul berdasarkan fakta bahwa pertumbuhan data yang sangat pesat, tetapi pengetahuan yang ada di dalamnya tidak banyak diketahui. Kemampuan teknologi informasi dalam mengumpulkan dan menyimpan berbagai tipe data jauh meninggalkan kemampuan untuk menganalisis, meringkas dan mengekstraksi pengetahuan dari data. Menurut [6] *data mining* adalah proses mengekstraksi informasi atau pola-pola yang menarik (tidak remeh-temeh, implisit, belum diketahui sebelumnya, dan berpotensi bermanfaat) dari data yang berukuran besar. Data mining merupakan salah satu langkah dalam menemukan pengetahuan (*knowledge discovery*). Berdasarkan [6], langkah-langkah dalam menemukan pengetahuan meliputi :

a. Membersihkan data, membuang *noise*, dan memperbaiki data yang tidak konsisten.
b. Integrasi data, yaitu menggabungkan data dari berbagai sumber data.
c. Seleksi data, yaitu memilih data yang relevan dengan kepentingan *data mining*.
d. Transformasi data, yaitu mengubah atau menggabungkan data ke bentuk-bentuk yang cocok untuk keperluan data mining, misalnya dengan operasi summary, agregasi, normalisasi.
e. Data mining, yaitu metode berintelegensia diterapkan dalam rangka mengekstrak pola-pola data. Ada beberapa metode data mining yang dapat dipilih antara lain asosiasi, klasifikasi, prediksi, estimasi, dan clustering.
f. Evaluasi pola, yaitu mengidentifikasi pola yang menarik yang merepresentasikan pengetahuan, berdasarkan ukuran-ukuran tertentu.
g. Menampilkan pengetahuan, yaitu menggunakan visualisasi dan teknik representasi pengetahuan

untuk menampilkan pengetahuan hasil proses mining kepada pengguna.

2.2 *Text mining*

Text mining adalah proses mendapatkan pengetahuan yang intensif di mana pengguna berinteraksi dengan kumpulan dokumen dengan menggunakan perangkat analisis tertentu. Jika dianalogikan dengan *data mining*, *text mining* mengekstrak informasi yang berguna dari berbagai sumber data melalui identifikasi dan eksplorasi pola-pola yang menarik. Sumber datanya adalah kumpulan dokumen, dan pola-pola yang menarik yang ditemukan tidak diformalisasi ke dalam *record* di database tetapi dalam data tekstual yang tidak terstruktur[10].

Contoh pemanfaatan *text mining* menurut [9] :

a. Aplikasi media online
   Text mining digunakan oleh perusahaan media besar seperti perusahaan Tribune untuk menghilangkan informasi ambigu, memberikan pembaca pengalaman pencarian yang lebih baik sehingga meningkatkan loyalitas pada situs dan meningkatkan pendapatan.
b. Analisis sentimen
   Analisis sentimen melibatkan analisis dari para pereview film untuk memperkirakan berapa baik review untuk sebuah film. Analisis semacam ini memerlukan kumpulan data berlabel misalnya WordNet.
c. Aplikasi akademik
   Masalah *text mining* penting bagi penerbit yang memiliki database besar untuk mendapatkan informasi yang memerlukan pengindeksan untuk pencarian. Hal ini terutama berlaku dalam ilmu sains, di mana informasi yang sangat spesifik sering terkandung dalam teks tertulis.
d. Filter email spam
   Text mining digunakan dalam beberapa filter email spam sebagai cara untuk menentukan karakteristik pesan yang mungkin berupa iklan atau materi yang tidak diinginkan lainnya.

2.3 *Opinion mining*

Menurut [11] opini adalah pendapat, pikiran, pendirian. *Opinion mining* yaitu komputasi tentang opini publik, penghargaan, sikap terhadap suatu objek, permasalahan, kejadian, suatu topik bahasan, dan atribut-atributnya. *Opinion mining* dapat menentukan apakah suatu opini itu termasuk kategori opini positif, opini negatif, atau opini yang bersifat netral. Penerapan *opinion mining* pada sekumpulan opini pembaca terhadap artikel tentang produk tertentu, menghasilkan rangkuman berapa persentase opini positif, opini negatif, atau opini netral pembaca.

2.4 Opini Spam

Berdasarkan [12] spam adalah email yang tidak diinginkan, email yang dikirim ke banyak orang dan sebagian besar berisi iklan. Meskipun definisinya fokus ke email, ternyata ada aktifitas lainnya yang merupakan spam, misalnya opini spam. Opini spam adalah opini palsu, opini abal-abal. Nama lain opini spam adalah *fake opinion*, *bogus opinion*, atau *fake review*. Pembuatan opini spam adalah kegiatan ilegal yang mencoba untuk menyesatkan pembaca atau *opinion mining* otomatis atau sistem analisis sentimen, dengan memberikan pendapat positif tidak layak untuk beberapa entitas yang menjadi sasaran untuk mempromosikan entitas tersebur dan atau dengan memberikan opini negatif palsu untuk beberapa entitas lain untuk merusak reputasi entitas yang diulas[3]

Ada dua jenis pembuat opini spam yaitu penulis spam tunggal dan penulis spam yang bekerja berkelompok. Penulis spam tunggal, tidak bekerja dengan orang lain, saat menulis review yang bersifat spam. Penulis spam terdaftar sebagai penulis tunggal atau mempunyai beberapa *user id*. Penulis spam berkelompok yaitu sekelompok spammer yang bekerja sama untuk mempromosikan entitas tertentu dan atau membahayakan reputasi entitas tertentu. Kelompok ini bisa sangat berbahaya, karena mereka dapat mengontrol sentimen produk dan menjerumuskan konsumen potensial.

2.5 Cara Mendeteksi Spam

Metode untuk mendeteksi spam menurut [3] meliputi:

a. Pendeteksian spam dengan pembelajaran supervisi
   Pendeteksian spam dianggap sebagai masalah klasifikasi dengan dua kelas, yaitu kelas spam dan kelas bukan spam. Di proses pembentukan model dibutuhkan data training yang mengandung kelas. Salah satu cara membuat data training adalah melabeli data training secara manual. Hal ini bisa dilakukan untuk opini spam tipe kedua dan ketiga. [4]
b. Pendeteksian spam dengan deteksi duplikasi
   Pemberian label secara manual untuk opini spam tipe pertama sangat sulit, karena penulis spam menulis opini spam seperti penulis opini lainnya. Pada kenyataannya sering sekali ditemukan opini yang duplikat atau mendekati duplikat. Jenis-jenis duplikasi opini yang termasuk jenis spam adalah :
   - Duplikasi dari pembuat opini dengan id yang berbeda pada produk yang sama
   - Duplikasi dari pembuat opini dengan id yang sama pada produk yang berbeda

- Duplikasi dari pembuat opini dengan id yang berbeda pada produk yang berbeda

Sedangkan duplikasi dari pembuat opini dengan id yang sama pada produk yang sama tidak digolongkan ke dalam spam. Misalnya tanpa sengaja mengklik mouse dua kali, sehingga data diabaikan salah satu [7].

c. Pendeteksian spam berdasarkan perilaku yang menyimpang
Ternyata banyak juga opini yang tidak duplikat yang merupakan spam. Untuk mengatasi hal itu dilakukan identifikasi pola kebiasaan pembuat opini yang tidak biasa dengan menemukan *unexpected rule.* Jenisnya meliputi :
   - *Confidence unexpectedness*
     Pengukuran ini berguna untuk menemukan pembuat opini yang memberikan rating tinggi pada merek tertentu, tetapi sebagian besar pembuat opini lain memberikan rating negatif pada merek tersebut.
   - *Support unexpectedness*
     Pengukuran ini berguna untuk menemukan pembuat opini yang yang menulis berbagai review pada suatu produk, sementara pembuat opini lainnya hanya menulis satu review.
   - *Attribute distribution unexpectedness*
     Pengukuran ini berguna untuk menemukan fakta bahwa opini positif yang paling banyak untuk suatu merek produk berasal dari satu orang pembuat opini saja, meskipun ada banyak pembuat opini lainnya yang juga mereview produk tersebut
- *Attribute unexpectedness*
  Pengukuran ini berguna untuk menemukan pembuat opini yang hanya menulis opini positif kepada merek tertentu dan hanya menulis opini negatif terhadap merek lainnya.

2.6 Pengukuran Tingkat Persamaan Dokumen

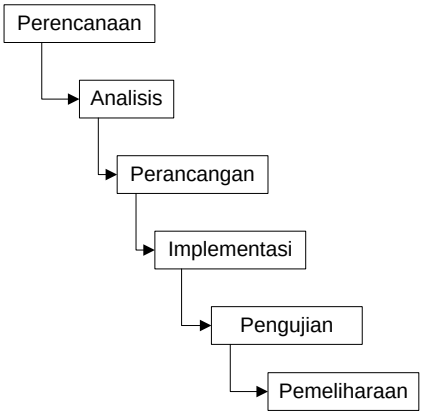
Menurut [5] dokumen direpresentasikan dengan representasi *non sequence* dan *sequence*. Pada representasi *non sequence* dokumen dianggap sebagai kumpulan kata-kata (*bag of words*) yang memperhatikan frekuensi kemunculan kata-kata saja sedangkan urutannya tidak diperhatikan. Sedangkan pada representasi *sequence* dokumen dianggap sebagai kumpulan *sequence of word* atau n-gram. Pada tahapan *preprocessing* dokumen teks, representasi *non sequence* dan *sequence* dijadikan sebagai vektor dengan kata-kata atau n-gram sebagai komponen vektornya.

Tingkat persamaan dua dokumen berhubungan dengan keterhubungan antara vektor yang diukur dengan nilai *cosine* antara dua vektor yang disebut *cosine similarity* [1].

Rumus *cosine similarity* = $\frac{A \cdot B}{||A||x||B||}$ (1)

A dan B menunjukkan vektor yang berisi jumlah kemunculan kata-kata yang berdimensi m, di mana A={a_1, a_2, a_3, a_4.., a_m} dan B= {a_1, a_2, a_3, a_4.., a_m}

2.7 Pengembangan Perangkat Lunak



Gambar 1 Tahapan Pengembangan Perangkat Lunak

Tahapan pengembangan perangkat lunak menurut [2] meliputi enam tahap. Tahap pertama, yaitu tahap perencanaan menyangkut studi tentang kebutuhan pengguna, studi kelayakan baik secara teknis maupun secara teknologi, serta penjadwalan pengembangan perangkat lunak. Pada tahap ini digunakan *use case diagram* untuk menangkap kebutuhan dan harapan pengguna. Tahap analisis, yaitu tahap untuk mengenali semua permasalahan yang muncul pada pengguna dengan mendekomposisi dan merealisasikan *use case diagram*, mengenali komponen-komponen sistem atau perangkat lunak, obje-objek, hubungan antar objek.

Tahap desain adalah tahap mencari solusi permasalahan yang didapat dari tahap analisis. Tahap perancangan ada dua yaitu tahap perancangan yang lebih menekankan pada platform implementasi dan tahap perancangan untuk menghaluskan kelas-kelas yang didapat pada tahap analisis dan menambahkan serta memodifikasi kelas-kelas akan mengefisienkan perangkat lunak yang dikembangkan.

Tahap keempat adalah tahap implementasi, yaitu mengimplementasikan perancangan sistem ke situasi nyata. Selanjutnya tahap pengujian yang digunakan untuk menentukan apakah perangkat lunak sudah sesuai dengan kebutuhan pengguna atau belum. Di samping itu tujuan lain pengujian adalah untuk menghilangkan atau meminimalisasi cacat program.Tahap terakhir adalah tahap

pemeliharaan, yaitu tahap pengoparasian sistem, dan jika diperlukan melakukan perbaikan-perbaikan kecil.

3 Pengembangan Perangkat Lunak

3.1 Tahap Perencanaan

Penelitian ini menggunakan sebelas berita politik dari www.detik.com. Setiap berita mempunyai komentar-komentar yang berasal dari pembaca. Data-data komentar tersebut akan disalin oleh admin ke dalam file teks, kemudian diinputkan ke sofware Spam Checker. Software akan melakukan beberapa proses, yaitu memecah-pecah isi file teks (*parsing*), menginputkannya ke basis data, menentukan apakah komentar-komentar tersebut termasuk spam atau tidak, dan menampilkan daftar komentar yang termasuk jenis spam ke layar.

Kebutuhan fungsional pengguna meliputi F1 s.d. F4 di bawah ini :

F1 : *Preprocessing*
F2 : Identifikasi spam dengan duplikasi
F3 : Identifikasi spam dengan pengelompokan jumlah komentar
F4 : Menampilkan daftar spam

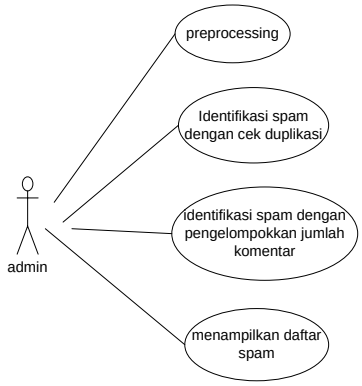
Kebutuhan non fungsionalnya adalah aplikasi mampu menangani data yang menggunakan bahasa Indonesia.

Perangkat lunak yang dibutuhkan selama pengembangan aplikasi adalah :

a. Sistem Operasi : Windows 7
b. DMBS : MySQL
c. Perangkat pengembangan : Netbeans

3.2 Tahap Analisis

Penelitian ini menggunakan *use case diagram* di tahap analisis. Ada satu aktor yang terlibat yaitu admin. Admin mempunyai hak akses terhadap empat proses, yaitu *preprocessing*, identifikasi spam dengan pengecekan duplikasi, identifikasi spam dengan pengelompokan jumlah komentar, dan menampilkan daftar spam. Use case diagram ada di gambar 2, sedangkan skenario keempat proses ada di tabel 1, tabel 2, tabel 3, dan tabel 4.



Gambar 1 Use Case Diagram

TABEL 1
SKENARIO *PREPROCESSING*

Use Case	*preprocessing*
Deskripsi	Melakukan pengolahan data dari file text yang tidak terstruktur ke bentuk yang terstruktur
Kondisi Awal	File text terisi
Kondisi Akhir	Tabel t_komentar dan t_artikel terisi
Skenario	1. Admin memasukkan data ke t_artikel 2. Admin menyalin data komentar dari www.detik.com ke file teks 3. Aplikasi memparsing data berdasarkan kriteria tertentu dan menyimpannya ke tabel t_komentar

TABEL 2
SKENARIO IDENTIFIKASI SPAM DENGAN CEK DUPLIKASI

Use Case	identifikasi spam dengan cek duplikasi
Deskripsi	Membandingkan tingkat persamaan antara dua komentar yang ada pada t_komentar
Kondisi Awal	Tabel t_komentar terisi
Kondisi Akhir	Tabel t_hasil terisi
Skenario	1. Membandingkan tingkat persamaan antara dua komentar yang ada pada t_komentar 2. Jika tingkat persamaan >= 80%, informasi kedua komentar akan disimpan ke t_hasil 3. Menampilkan isi t_hasil 4. Admin akan menentukan apakah komentar tersebut spam atau tidak 5. Jika spam, dilakukan update isi kolom spam_class dengan ‘spam’ di tabel t_komentar

TABEL 3
SKENARIO IDENTIFIKASI SPAM DENGAN PENGELOMPOKAN JUMLAH KOMENTAR

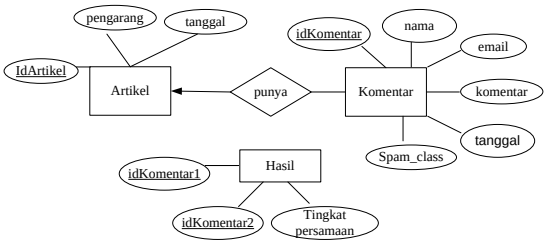
Use Case	identifikasi spam dengan pengelompokan jumlah komentar

Deskripsi	Menghitung jumlah komentar yang ditulis oleh pembaca pada sebuah artikel
Kondisi Awal	Tabel t_komentar terisi
Kondisi Akhir	Tabel t_komentar terupdate
Skenario	1. Menghitung jumlah komentar yang ditulis oleh pembaca pada sebuah artikel 2. Menampilkan jumlah komentar yang ditulis oleh pembaca pada sebuah artikel 3. Admin akan menentukan apakah komentar tersebut spam atau tidak 4. Jika spam, dilakukan update kolom spam_class dengan 'spam' di tabel t_komentar

TABEL 4
SKENARIO MENAMPILKAN DAFTAR SPAM

Use Case	menampilkan daftar spam
Deskripsi	menampilkan semua komentar yang berjenis spam
Kondisi Awal	Tabel t_komentar terisi
Kondisi Akhir	Menampilkan informasi daftar spam ke layar
Skenario	Menampilkan data di tabel t_komentar yang mempunyai isi spam_class adalah 'spam' ke layar

Pada tahap analisis ini juga dibuat diagram *Entity Relationship* (ER). Di diagram ER di gambar 3, terdapat tiga entitas yang terlibat, yaitu artikel, komentar, dan hasil. Setiap artikel mempunyai banyak komentar. Setiap komentar milik 1 artikel saja. Sedangkan entitas hasil menyimpan hasil perbandingan antara dua komentar yang memiliki nilai tingkat persamaan >= 80% dengan metode *cosine similarity.*



Gambar 2 Diagram ER

3.3 Tahap Perancangan

Perancangan tabel berdasarkan diagram ER ada di tabel 5, tabel 6, dan tabel 7.

TABEL 5
T_ARTIKEL

Field	Type data	Keterangan
IdArtikel	Integer	Primary key
Pengarang	Varchar(50)	
Tanggal	Varchar(30)	

TABEL 6
T_KOMENTAR

Field	Type data	Keterangan
idKomentar	integer	Primary key
idArtikel	Integer	Foreign key dari t_artikel
Nama	Varchar(30)	
email	Varchar(100)	
Tanggal	Varchar(30)	
Komentar	Varchar(2000)	
Spam_class	Varchar(10)	

TABEL 7
T_HASIL

Field	Type data	Keterangan
idKomentar1	integer	Primary key Foreign key dari t_komentar
idArtikel1	Integer	Primary key Foreign key dari t_komentar
Nama1	Varchar(30)	
Tanggal1	Varchar(30)	
idKomentar2	Integer	Primary key Foreign key dari t_komentar
idArtikel2	Integer	Primary key Foreign key dari t_komentar
Nama2	Varchar(30)	
Tanggal2	Varchar(30)	
TingkatPersa-maan	double	

Algoritma *preprocessing*

Deskripsi : melakukan pengolahan data dari file text yang tidak terstruktur ke bentuk yang terstruktur
Kondisi awal : File text terisi
Kondisi akhir : Tabel t_komentar dan t_artikel terisi
Deklarasi :
file, nama_file_komentar, a: String
nama, email, tanggal, komentar : String
id_artikel, id_komentar : integer
hasilParsing : array[0..3] of String
Algoritma :
input (id_artikel)

```
//input nama file yang berisi semua komentar dari sebuah artikel
file←input (nama_file_komentar)
open (file)
id_komentar ←1
while (a=bacaBaris() <> null )
begin
    //memparsing a untuk mendapatkan info nama, email, tanggal, komentar
    parsing (a, hasilParsing)

    //simpan hasil parsing ke t_komentar
    nama ←hasilParsing[0]
    email ←hasilParsing[1]
    tanggal ←hasilParsing[2]
    komentar←hasilParsing[3]

    //jalankan query :
    insert into t_komentar values (id_komentar, id_artikel, nama, email, tanggal, komentar, null)

    id_komentar ←id_komentar + 1
end while
close(file)
```

Algoritma identifikasi spam dengan cek duplikasi

Deskripsi : identifikasi spam dengan pengecekan duplikasi
Kondisi awal : tabel t_komentar terisi
Kondisi akhir : tabel t_hasil terisi
Deklarasi :
st, st1, st2, st3,st4 : PreparedStatement
rs, rs1, rs3 : ResultSet
id_komentar, id_artikel : String
idkomentar1, idartikel1, nama1:string
tanggal1, komentar1: string
idkomentar2, idartikel2, nama2:string
tanggal2, komentar2: string
A, B : string
sim_score : double

Algoritma :

```
//jalankan query:
st ←select * from t_komentar
rs←st.executeQuery()
while (rs.next())
begin
    idkomentar1←rs.getInt("idKomentar")
    idartikel1←rs.getInt("idArtikel")
    nama1←rs.getString("nama")
    tanggal1←rs.getString("tanggal")
    komentar1←rs.getString("komentar")

    //jalankan query
    st1←select * from t_komentar where idstatus>?
    st1.setInt(1, IdStatus1)
    rs1 ←st1.executeQuery()
    while (rs1.next())
    begin
        idkomentar2←rs1.getInt("idKomentar")
        idArtikel2←rs1.getInt("idArtikel")
        nama2←rs1.getString("nama")
        tanggal2← rs1.getString("tanggal")
        komentar2←rs1.getString("komentar")

        //bandingkan komentar1 dengan komentar2
        A←komentar1
        B←komentar2
        //memanggil fungsi cosine similarity
        sim_score=Cos_Similarity(A,B)
        if (sim_score>=0.80) then
        begin
            //masukkan ke t_hasil
            st2←insert into t_hasil values (idkomentar1, idartikel1, nama1, tanggal1, idkomentar2, idartikel2, nama2, tanggal2, sim_score)
        end if
    end while
end while

//Menampilkan isi t_hasil
st3 ←select * from t_hasil
rs3←st.executeQuery()
while (rs.next())
begin
    output(idkomentar1)
    output(idartikel1)
    output(nama1)
    output(tanggal1)
    output(idkomentar2)
    output(idartikel2)
    output(nama2)
    output(tanggal2)
    output(sim_score)
end while


//Admin menentukan apakah komentar tersebut spam
//atau tidak. Jika spam, dilakukan update isi kolom
//spam_class dengan 'spam' di tabel t_komentar
input(id_komentar)
input(id_artikel)
St4←update t_komentar set spam_class="spam" where idKomentar=id_komentar and idArtikel=id_artikel
```

Algoritma identifikasi spam dengan pengelompokan jumlah komentar

Deskripsi : identifikasi spam dengan pengelompokan jumlah komentar
Kondisi awal : Tabel t_komentar terisi
Kondisi akhir : Tabel t_komentar terupdate
Deklarasi :
st, st1: PreparedStatement
rs: ResultSet

Algoritma :

```
//menghitung jumlah komentar yang ditulis oleh
//pembaca pada sebuah artikel
For i←1 to 11 do
  st ←select nama, count(idKomentar) from
      t_komentar where idArtikel=i group by nama
   rs←st.executeQuery()
   while (rs.next())
       //menampilkan jumlah komentar yang ditulis
oleh
       // pembaca pada sebuah artikel
       output(i)
       output(rs.getString("nama"))
       output(rs.getInt("count(idKomentar)"))
    end while
end for

//Admin akan menentukan apakah komentar
tersebut //spam atau tidak. Jika spam, dilakukan
update kolom //spam_class dengan 'spam' di tabel
t_komentar
input(id_komentar)
input(id_artikel)
st1←update t_komentar set spam_class="spam"
where idKomentar=id_komentar and
idArtikel=id_artikel
```

Algoritma menampilkan daftar spam

Deskripsi : menampilkan daftar spam
Kondisi awal : Tabel t_komentar terisi
Kondisi akhir : Menampilkan informasi daftar spam ke layar
Deklarasi :
st: PreparedStatement
rs: ResultSet
Algoritma :

```
//jalankan query:
st←select * from t_komentar where
  spam_class="spam"
rs←st.executeQuery()
while (rs.next())
begin
      output(idkomentar)
      output(idartikel)
      output(name)
      output(komentar)
      output(spam_class)
end while
```

3.4 Tahap Implementasi

Perangkat lunak *Spam Checker* dibuat sesuai dengan hasil analisis dan perancangan. Perangkat lunak tersebut dikembangkan dengan menggunakan bahasa Java dengan menggunakan IDE Netbeans. Implementasi *cosine similariry* menggunakan kode yang dibuat oleh [8]. Data disimpan dalam perangkat lunak *Database Management System* (DBMS) MySQL. Nama database adalah komentarpolitik, yang berisi tiga tabel yaitu t_artikel, t_komentar, dan t_hasil.

3.5 Tahap Pengujian

Pengujian perangkat lunak dilakukan dengan metode *black box*. Hasil pengujian ada di tabel 8.

TABEL 8
HASIL PENGUJIAN PERANGKAT LUNAK

Id	Deskripsi	Hasil
F1	*Preprocessing*	OK
F2	Identifikasi spam dengan pengecekan duplikasi	OK
F3	Identifikasi spam dengan pengelompokan jumlah komentar	OK
F4	Menampilkan daftar spam	OK

4 Hasil dan Pembahasan

4.1 Use Case Data Preprocessing

Use case data preprocessing menghasilkan data komentar sebanyak 993 buah berasal dari sebelas artikel berita politik. Tahapan *preprocessing data* tidak melibatkan proses pembuangan *stop word list* dan tidak melakukan proses *stemming* atau pengubahan bentuk kata ke dalam kata dasar, karena peneliti ingin memproses komentar dalam bentuk yang asli atau apa adanya. Dampaknya penelitian ini bisa memproses data komentar dalam banyak bahasa yang menggunakan huruf latin, tidak hanya dibatasi bahasa Indonesia saja.

4.2 *Use Case* Identifikasi Spam dengan Duplikasi

TABEL 9
HASIL EKSPERIMEN YANG MEMPUNYAI TINGKAT PERSAMAAN DUA KOMENTAR >=80%

	Komentar 1			Komentar 2		
Id Artikel	Id komentar	Nama	Waktu	id komentar	Nama	Waktu
1	2	Stella Gracia	1 day ago	3	Stella Gracia	1 day ago
1	81	Ronaldo Rabbani	14:02:19	155	Ronaldo Rabbani	13:02:14
1	109	agus setiawan	13:28:58	205	Rahmad Budiani	12:13:34
2	232	Indrayana Harja	16:06:45	233	Indrayana Harja	15:31:02
2	296	Heri Purwanto	09:42:23	318	bejogembul	09:26:51
6	655	Latif Djukborneo	11:52:11	656	Latif Djukborneo	11:43:01
6	655	Latif Djukborneo	11:52:11	657	Latif Djukborneo	11:42:32
6	655	Latif Djukborneo	11:52:11	658	Latif Djukborneo	11:41:48

Id Artikel	Komentar 1			Komentar 2		
	Id komentar	Nama	Waktu	id komentar	Nama	Waktu
6	655	Latif Djukborneo	11:52:11	659	Latif Djukborneo	11:40:44
6	655	Latif Djukborneo	11:52:11	660	Latif Djukborneo	11:40:13
6	656	Latif Djukborneo	11:43:01	657	Latif Djukborneo	11:42:32
6	656	Latif Djukborneo	11:43:01	658	Latif Djukborneo	11:41:48
6	656	Latif Djukborneo	11:43:01	659	Latif Djukborneo	11:40:44
6	656	Latif Djukborneo	11:43:01	660	Latif Djukborneo	11:40:13
6	657	Latif Djukborneo	11:42:32	658	Latif Djukborneo	11:41:48
6	657	Latif Djukborneo	11:42:32	659	Latif Djukborneo	11:40:44
6	657	Latif Djukborneo	11:42:32	660	Latif Djukborneo	11:40:13
6	658	Latif Djukborneo	11:41:48	659	Latif Djukborneo	11:40:44
6	658	Latif Djukborneo	11:41:48	660	Latif Djukborneo	11:40:13
6	659	Latif Djukborneo	11:40:44	660	Latif Djukborneo	11:40:13
9	866	M_ikwan	an hour ago	912	M_ikwan	2 hours ago
9	903	Jkwopfer	2 hours ago	906	Jkwopfer	2 hours ago
9	903	Jkwopfer	2 hours ago	907	Jkwopfer	2 hours ago
9	903	Jkwopfer	2 hours ago	918	Jkwopfer	2 hours ago
9	906	Jkwopfer	2 hours ago	907	Jkwopfer	2 hours ago
9	906	Jkwopfer	2 hours ago	918	Jkwopfer	2 hours ago

Hasil eksperimen yang mempunyai tingkat persamaan dua komentar lebih besar atau sama dengan 80% ada di tabel 9. Tabel tersebut ditampilkan ke layar, sehingga admin bisa melakukan justifikasi apakah komentar tersebut termasuk spam atau tidak. Hasil justifikasi admin adalah :

a. Ronaldo Rabbani adalah spammer karena mengirim komentar yang sama dalam selang waktu 1 jam dan id komentar miliknya jaraknya cukup jauh (81 dan 155) pada artikel 1.
b. Agus Setiawan adalah spammer, karena mengirim komentar yang sama persis (nilai *cosine similirity*=100%) dengan Rahmad Budiani. Ini sesuai dengan teori tentang duplikasi opini yang termasuk jenis spam yang berbunyi "duplikasi dari pembuat opini dengan id yang berbeda pada produk yang sama".
c. Rahmad Budiani adalah spammer. Alasannya seperti penjelasan di point b.
d. Indrayana Harja adaah spammer karena mengirim komentar yang sama dalam selang waktu cukup lama dan komentar yang dikirim tidak sama persis, sehingga tidak bisa dianggap sebagai kesalahan input data(misalnya : salah klik mouse).
e. Heri Purwanto dan Bejogembul adalah spammer. Alasan seperti penjelasan di point b.
f. Latif Djukborneo adalah spammer, karena mengirim enam komentar dalam selang waktu cukup lama.
g. M_ikwan adalah spammer karena mengirim komentar dalam selang waktu cukup lama dan id komentar miliknya jaraknya cukup jauh (866 dan 912).
h. Jkwopfer adalah spammer karena mengirim tiga komentar yang tidak sama persis, sehingga tidak bisa dianggap sebagai kesalahan input data (misal : salah klik mouse).

Dalam penelitian ini, peneliti tidak menemukan jenis "duplikasi dari pembuat opini dengan id yang sama pada produk yang berbeda" dan "duplikasi dari pembuat opini dengan id yang berbeda pada produk yang berbeda". Sebenarnya metode penelitian yang digunakan memungkinkan untuk menemukan dua jenis duplikasi opini lainnya tersebut, hanya saja data *sample* yang dipakai ternyata tidak mengandung kedua jenis duplikasi tersebut.

4.3 *Use Case* Identifikasi Spam dengan Pengelompokan Jumlah Komentar

Proses ini dilakukan berdasarkan teori tentang pendeteksian spam berdasarkan perilaku yang menyimpang yaitu *support unexpectedness.* Pengukuran *support unexpectedness* menemukan pembuat opini yang menulis berbagai review pada suatu produk, sementara pembuat opini lainnya hanya menulis satu review. SQL query dan hasil query ada di tabel 10. Nama-nama yang tertera di tabel 10 adalah spammer karena mereka mengirim jumlah komentar lebih dari dua. Rata-rata penulis komentar hanya menulis satu komentar saja.

TABEL 10
HASIL PENGELOMPOKAN JUMLAH KOMENTAR BERDASARKAN NAMA PEMBERI KOMENTAR

Id Artikel	Nama	Jumlah kemunculan

1	Apa22222	3
	Eimhard	3
	Meizon	3
2	Bayubisma	3
	Dodi.irawan	11
	Indrayana Harja	4
	mamad123	3
	M_ikwan	3
	Politikkejam	3
6	Latif Djukborneo	6
9	GundulPacul5	3
	Jkwopfer	4
	M_ikwan	3
	santaiajah	4
	Tony Admono	4
	Usil	3

5 Kesimpulan dan Saran

Kesimpulan penelitian ini adalah :

1. Penelitian ini menemukan spammer dengan metode pengecekan duplikasi komentar. Pengecekan duplikasi membandingkan setiap komentar yang ada di semua artikel dengan menggunakan *cosine similarity*.
2. Penelitian ini menemukan duplikasi dari pembuat opini dengan id yang berbeda pada produk yang sama.
3. Penelitian ini menemukan spammer dengan pendeteksian perilaku yang menyimpang(*support unexpectedness)*.
4. Parangkat lunak yang dikembangkan bisa mendeteksi komentar spam yang ditulis dalam bahasa apapun, tidak terbatas bahasa Indonesia saja.

Saran pengembangan penelitian ini adalah :

1. Membuat fitur untuk menangkap komentar dari Twitter atau Facebook, sehingga admin tidak perlu melakukan penyalinan data komentar dari media online ke file teks.
2. Menerapkan fitur pengelolaan sentimen, untuk menandai apakah sutu komentar termasuk komentar positif, negatif, atau netral. Penerapan fitur ini memungkinkan implementasi identifikasi spam dengan metode *confidence unexpectedness*, *attribute distribution unexpectedness*, dan *attribute unexpectedness.*

Daftar Pustaka

[1] A. Huang, “Similarity Measures for Text Document Clustering”, *New Zealand Computer Science Research Student Conference*, pp. 49-56, 2008

[2] A. Nugroho, *Rekayasa Perangkat Lunak Berorientasi Objek dengan Metode USDP (Unified Software Development Process)*, Penerbit Andi Yogyakarta, 2010

[3] B. Liu, L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis”, *Mining Text Data*, pp.415-463, Springer, 2012

[4] B. Liu, “Sentiment Analysis and Opinion Mining”, Mining Text Data, Morgan & Claypool, 2012.

[5] H. Widyastuti, “Studi Representasi N-Gram pada Algoritma HMRF-KMeans untuk Document Clustering”, *Tesis*, Institut Teknologi Bandung, 2008

[6] J. Han, M. Kamber, *Data Mining: Concepts and Techniques 2nd edition*, Morgan Kaufmann: San Fransisco, 2006.

[7] N. Jindal, B. Liu, “Opinion spam and Analysis”, *Proceeding of Conference on Web Search and Web Data Mining*, 2008.

[8] N. Kumar, “Cosine_Similarity”, https://sites.google.com/site/nirajatweb/home/technical_and_coding_stuff/cosine_similarity, 2014, diakses pada 29 November 2014.

[9] N. W. S. Saraswati, ”Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis”, *Tesis*, Program Pascasarjana Universitas Udayana, Denpasar, Indonesia, 2011.

[10] R. Feldman, J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007.

[11] Kamus besar bahasa Indonesia, http://kbbi.web.id, diakses pada 10 Maret 2014.

[12] An Encyclopedia Britannica Company, http://www.merriam-webster.com/, diakses pada 10 Maret 2014.