**TESIS**

**OPINION MINING PADA TWITTER UNTUK BAHASA INDONESIA DENGAN METODE SUPPORT VECTOR MACHINE DAN METODE BERBASIS LEXICON**

****

**JAN KRISTANTO WIBISONO**

**10/306141/PPA/03230**

**PROGRAM STUDI S2 ILMU KOMPUTER**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS GADJAH MADA**

**YOGYAKARTA**

**2013**

# DAFTAR ISI

[DAFTAR ISI v](#_Toc367271047)

[BAB I 1](#_Toc367271048)

[PENDAHULUAN 1](#_Toc367271049)

[BAB II 3](#_Toc367271050)

[TINJAUAN PUSTAKA 3](#_Toc367271051)

[BAB III 6](#_Toc367271052)

[METODOLOGI PENELITIAN 6](#_Toc367271053)

[3.1 Pengumpulan Data 6](#_Toc367271054)

[3.2 Analisis Sistem 6](#_Toc367271055)

[3.3 Garis Besar Sistem 8](#_Toc367271056)

[BAB IV 10](#_Toc367271057)

[HASIL PENELITIAN DAN PEMBAHASAN 10](#_Toc367271058)

[4.1 Perbandingan proses pelabelan data pelatihan 10](#_Toc367271059)

[4.2 Akurasi Pemberian label 12](#_Toc367271060)

[BAB V 13](#_Toc367271061)

[KESIMPULAN DAN SARAN 13](#_Toc367271062)

[5.1 Kesimpulan 13](#_Toc367271063)

[5.2 Saran 13](#_Toc367271064)

[DAFTAR PUSTAKA 15](#_Toc367271065)

# 

# PENDAHULUAN

Opini adalah pendapat pribadi yang tidak obyektif dan tidak melalui proses verifikasi [1]. Opini orang lain tentang suatu hal menjadi penting dalam pengambilan keputusan. Misalkan seseorang akan membeli sebuah mobil biasanya dia akan bertanya kepada orang lain bagaimana pendapat mereka tentang mobil tersebut. Atau sering juga ketika kita membeli produk tertentu, kita dipersilakan untuk mengisi kuisoner yang berkaitan dengan produk tersebut. Sekarang ini tidak perlu kita langsung bertanya tentang pendapat orang lain secara langsung. Seiring dengan perkembangan internet dan web, telah banyak orang yang berbagi informasi tentang pendapat dan pengalaman bahkan kritik mereka terhadap suatu produk tertentu.

Social media kini telah berkembang dengan pesat, dimulai dari awal Friendster sampai era sekarang ini Facebook, Twitter, Google+ dan Foursquare. Menurut socialbakers.com pengguna Facebook di Indonesia telah mecapai 42.5 juta, yang menjadikan Indonesia menjadi peringkat ke empat pengguna Facebook setelah Amerika, India dan Brazil. Pengguna Twitter di Indonesia juga tidak kalah besar jumlahnya yang mencapai 19.5 juta [2] yang menempati peringkat kelima setelah Amerika, Brazil, Japan dan U.K. Perkembangan yang begitu pesat ini membuat orang-orang berkomunikasi dengan mudah. Mereka saling berbagi informasi, baik berbagi tentang aktifitas sehari-hari mereka ataupun memberikan pemberitahuan pada komunitas mereka, bahkan keluhan mereka terhadap suatu hal.

Begitu populernya social media, pengguna social media sekarang ini bukan hanya kalangan pribadi atau perseorangan saja. Tapi kalangan corporate juga memanfaatkan social media sebagai media komunikasinya dengan masyarakat. Seperti pengenalan produk baru, penawaran produk, bahkan para legislative juga mempromosikan diri mereka melalui social media. Selain yang dianggap lebih murah dari segi biaya, juga lebih tepat pada sasaran.

Berdasarkan pada paparan diatas terdapat potensi besar pada social media untuk dilakukan pengekstrakan opini atau yang disebut opinion mining, untuk mengetahui pendapat orang lain tentang suatu hal, yang dapat dimanfaatkan selama proses pengambilan keputusan. Beberapa penelitian yang sudah ada sekarang ini kebanyakan menggunakan machine learning baik dengan naïve bayes [3],[4], [5],[6], [7] , Support Vector Machine [4],[5],[8], [6] dan Maximum Entropy [6] ataupun yang berbasiskan lexicon [9]. Pada penelitian ini akan dilakukan opinion mining dengan mengkombinasikan antara pendekatan berbasis machine learning dan berbasiskan lexicon. Metode yang akan digunakan untuk mewakili pendekatan machine learning adalah Support Vector Machine (SVM), yang menurut paper Zhang et al. [8] sering memberikan akurasi yang terbaik dibanding dengan metode learning yang lainnya. Dengan melakukan kombinasi antara metode Support Vector Machine (SVM) dan berbasiskan lexicon diharapkan dapat meningkatkan akurasi pada proses pengklasifikasian opini.

# 

# TINJAUAN PUSTAKA

Bab ini memuat uraian mengenai penelitian-penelitian yang telah dilakukan sebelumnya yang berkaitan dengan *opinion mining* atau *sentiment analysis* dalam Bahasa Inggris dan Bahasa Indonesia.

Pang et al. [4] menggunakan *machine learning* untuk mengklasifikasi *movie reviews*. Penelitian ini melakukan klasifikasi sentimen terhadap review film dan menentukan apakah *review* tersebut memiliki sentimen positif atau negarif. Fitur yang berbeda dari review film tersebut diekstrak dan digunakan algoritma *machine learning* Naïve Bayes dan Support Vector Machine (SVM) untuk menghasilkan klasifikasi model. Mereka memperoleh akurasi antara 78.7% ketika menggunakan Naïve Bayes pada penggunakan *unigram*. Akurasi yang diperoleh ketika menggunakan SVM dengan *unigram* adalah 72.8%.

Franky dan Manurung [5] mencoba mengulangi eksperimen klasifikasi sentimen *movie reviews* yang dilakukan oleh Pang et al. [4]. Perbedaan utama penelitian ini, adalalah penelitian ini mengambil fokus *movie reviews* dalam bahasa indonesia. Berkaitan dengan ketidaktersediaannya *traning corpora* untuk bahasa Indonesia, maka diaplikasikan *machine translation tools* untuk mentranslasikan corpus Bahasa Inggris yang dibuat Pang et al. [4] yang asli ke Bahasa Indonesia dan hasil translasinya digunakan, mulai dari *commercial tool* hingga translasi sederhana kata demi kata dan metoda klasifikasi teks dicoba. Hasil rata-rata yang diperoleh untuk metode Naïve Bayes adalah 74.62% dan 75.62% untuk metode SVM.

*Opinion mining* dalam bahasa china pernah dilakukan oleh Zhang et al. [8]. Penelitian ini, juga terinspirasi dari penelitian Pang et al. [4]. Jika pada penelitian Pang et al. [4] menggunakan data *movie reviews* dalam bahasa inggris, penelitian ini menggunakan *online reviews* produk komputer dalam bahasa cina. Mereka mengembangkan *opinion mining* berdasarkan metode *learning*. Menggunakan *dataset* opini yang ada pada AmazonCN. Metode pembelajaran yang digunakan adalah Support Vector Machine dengan menggunakan String kernel.

Pada tahun 2009 ketika *social media* mulai diminati pengguna internet. Go et al. [6] melakukan *opinion mining* pada *social media* twitter. Penelitian ini memulai ide pemanfaatan *emoticons* dalam mempermudah pelabelan data *training* dari tweet berbahasa inggris. Ide ini melengkapi penelitian sebelumnya yang dilakukan Pang et al. [4] yang masih menggunakan cara manual untuk mendapatkan data *training*. Penelitian ini mengklasifikasikan sentimen tweet atas 2 kelas yaitu kelas sentiment positif dan negatif. Akurasi yang diperoleh adalah 81,3% dengan menggunakan Naïve Bayes dan 80,5 menggunakan Maximum Entropy serta 82,2% dengan menggunakan SVM.

Pak dan Paroubek [7] menggunakan *emoticons* untuk membangun data *training* berbahasa inggris dari twitter dengan sentimen positif, negatif dan netral. Ide membangun data *training* dengan menggunakan *emoticons* sebenarnya sudah cetuskan oleh Go et al. [6]. Penelitian ini menambahkan, untuk data *training* kelas netral mengambil data dari *tweet* akun media berbahasa inggris. Metode yang digunakan adalah Naïve Bayes dengan n-grams. Performasi terbaik dihasilkan ketika menggunakan bigram.

Komarsilam dan Winarko [9] melakukan opinion mining pada review film berbahasa Indonesia dengan metode *unsupervised learning* yang berbasis pada lexicon berbahasa Indonesia dan menggunakan rule untuk menentukan frasa opini dan entitas yang berkaitan dengan hal opini tersebut. Hasil yang diperoleh berupa polaritas sentimen dari suatu teks apakah positif, negatif ataupun netral dan entitas yang menyertai opini yang telah ditentukan, dimana hasil analisa digambarkan dalam sebuah diagram yang menunjukan jumlah persentasi sentimen dan ditampilkan juga label yang merupakan detail dari hasil analisa sentiment. Dari hasil pengujian terhadap sistemnya didapatkan nilai presisi sebesar 0.616 dan nilai *recall* sebesar 0.643.

Alliandu dan Winarko [3] melakukan opinion mining pada twitter berbahasa Indonesia. Metode yang digunakan adalah *supervised learning* yaitu naïve bayes. Metode pengumpulan data *training* mengambil ide dari penelitian Pak dan Paroubek [7] yang menggunakan metode *emoticons* dan untuk kelas netral mengambil data dari akun media. Fitur pembobotan yang digunakan adalah *term frequency* dengan *laplace smoothing* dan TF-IDR. Dari hasil penelitian diperoleh akurasi 77.45% untuk *term frequency* dengan *laplace smoothing* dan akurasi 75.86% untuk TF-IDF pada test set yang dianotasikan menggunakan *emoticon*.

# 

# METODOLOGI PENELITIAN

## Pengumpulan Data

Data yang dibutuhkan pada penelitian ini terdiri atas dua jenis data, yaitu data yang digunakan untuk pelatihan dan data pengujian pemberian label. Data yang digunakan untuk pelatihan berupa dokumen tweet berbahasa indonesia yang diberi label dengan metode lexicon based. Dokumen ini diambil dari penelitian Alliandu dan Winarko [3] yang jumlah 81500 tweet. Namun setelah diproses menggunakan lexicon based data yang digunakan untuk pelatihan hanya 25000.

Berbeda dengan data yang digunakan untuk pelatihan , data yang digunakan pada proses pelabelan menggunakan tweet yang belum memiliki label. Dokumen tweet tersebut didapatkan dengan memanfaatkan API twitter yang dapat mencari tweet yang ada pada twitter dengan parameter tertentu.

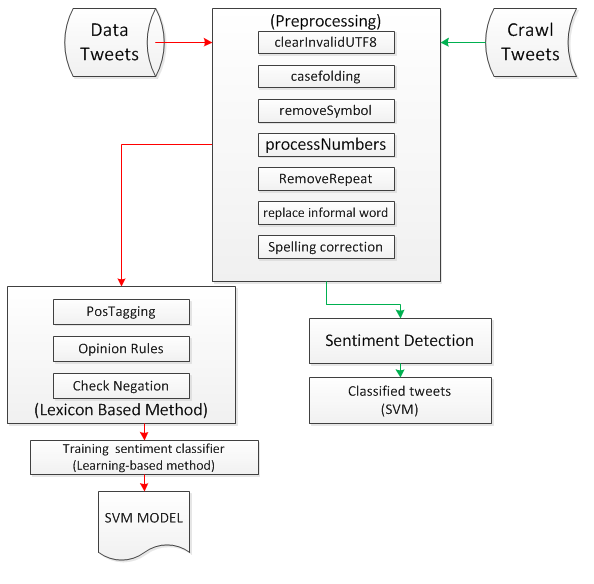
## Analisis Sistem

Sistem *Opinion Mining* ini memiliki dua buah proses utama, yaitu pelatihan dan penentuan kelas sentimen pada data baru. Pelatihan digunakan untuk mendapatkan model yang akan digunakan untuk menentuan kelas sentimen pada data baru. Proses-proses yang terjadi dalam sistem ini diperlihatkan pada gambar 1.

Pada proses pelatihan dengan menggunakan masukan berupa corpus yang digunakan pada penelitian Komarsilam dan Winarko [9] yang telah dilakukan pelabelan dengen metode *lexicon based*. Langkah awal dari pelabelan dengan menggunakan *lexicon based* adalah melakukan *preprocessing* terhadap data. Dalam *preprocessing* terdapat beberapa tahap, secara berurutan adalah membersihkan yang bukan termasuk karater UTF-8, merubah semua karakter menjadi huruf kecil, menghapus simbol-simbol, memproses angka-angka, menghapus perulangan karakter, melakukan penggantian terhadap kata tidak baku, melakukan *spelling correction*. Setelah selesai tahap *preprocessing*, selanjutnya akan masuk dalam tahan penentuan label. Dalam tahap ini juga terdapat beberapa proses diantaranya *POS taging*, menganalisis tweet dengan rules yang sudah ada, mengecek negasi. Sampai disini telah dihasilkan corpus yang labelnya ditentukan dengan menggunakan *lexicon based* dan siap digunakan untuk proses pelatihan.

Proses berikutnya adalah proses inti pelatihan, dalam penelitian ini terdapat 25000 data pelatihan yang sudah ditentukan labelnya dengan metode *lexicon based,* terdiri dari data 12500 data berlabel positif dan 12500 data berlabel negatif. Proses pelatihan ini nantinya akan menghasilkan SVM model. Yaitu sebuah model yang nantinya digunakan untuk menentukan sentimen dengan metode SVM.

Proses utama yang kedua adalah penentuan label *tweet* baru dengan metode SVM. Langkah pertama pada proses utama ini adalah sistem akan mencari *tweet* yang berhubungan dengan kata kunci yang diinginkan. Selanjutnya akan masuk dalam tahap *preprocessing* yang telah dijelaskan sebelumnya. Kemudian dilakukan pengecekan indikasi keberadaan frase opini pada *tweet* yang telah dipreprocessing. Selanjutnya untuk *tweet* yang terindikasi memiliki frase opini, sentimennya akan ditentukan dengan metode SVM.



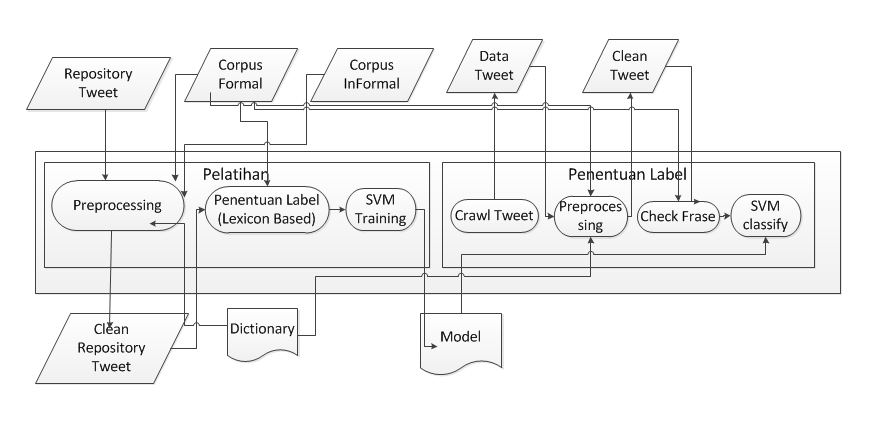
Gambar . Proses Opinion Mining

## Garis Besar Sistem

Arsitektur sistem Opinion Mining yang dibuat ditunjukan pada gambar 2 Pada sistem Opinion Mining yang dibuat terdapat dua proses utama. Yaitu proses pelatihan dan proses penentuan pelabelan atau testing. Proses pelatihan menggunakan metode lexicon based untuk menentukan label, yang kemudian hasil pelabelan tersebut akan menjadi data training pada proses testing yang menggunakan algoritma Support Vector Machine.

Alur pada proses pelatihan adalah melakukan proprocessing pada data yang sudah ada. Dimana dalam preprocessing melibatkan corpus kata formal dan kata informal, yang kemudian akan menghasilkan data tweet yang sudah di preprocessing atau dengan istilah data bersih. Setelah selesai dilakukan preprocessing akan dilakukan penentuan label pada untuk data tweet yang sudah dipreprocessing dengan menggunakan metode lexicon based. Sampai disini sistem sudah memiliki data tweet yang sudah memiliki label. Selanjutnya data tweet yang sudah memiliki label akan diubah menjadi vektor sebagai data training SVM menggunakan metode TF-IDF. Kemudian dilakukan proses pelatihan dengan metode SVM. Pelatihan ini akan menghasilkan SVM model yang nanti akan digunakan pada proses penentuan label atau testing.

Alur proses besar berikutnya adalah proses penentuan label atau testing. Pertama sistem akan mencari tweet menggunakan tweetAPI yang sesuai dengan kata kunci yang dimasukan oleh pengguna. Proses ini akan menghasilkan data tweet dan akan disimpan dalam basisdata. Kemudian dilakukan preprocessing pada data tweet tersebut. Prerpocessing yang dilakukan sama seperti yang dilakukan pada alur proses pelatihan. Setelah dilakukan preprocessing akan dilakukan pengecekan frase, apakah mengandung frase opini atau tidak. Jika mengandung frase opini maka akan dimasukan dalam proses klasifikasi.



Gambar . Arsitektur Sistem

# 

# HASIL PENELITIAN DAN PEMBAHASAN

hasil penelitian yang terdiri dari tiga bagian. Pertama, perbandingan antara proses pelabelan data pelatihan yang dilakukan Alliandu dan Winarko[3] yaitu menggunakan metode emoticon dan pelabelan data pelatihan yang dilakukan dengan metode *lexicon* *based*. Kedua, percobaan pemberian label dengan menggunakan metode gabungan antara *lexicon* *based* dan *support* *vector* *machine*. Ketiga, perbandingan hasil pelabelan metode gabungan dan pelabelan dengan cara manual

## Perbandingan proses pelabelan data pelatihan

Proses pelabelan data yang dilakukan dalam penelitian ini akan menggunakan metode *lexicon* *based* sedangkan data pelatihannya mengambil dari pelatihan yang dilakukan oleh Alliandu dan Winarko [3]. Pada pelatihan yang dilakukan Alliandu dan Winarko [1] metode yang digunakan adalah dengan metode emoticon. Metode emoticon adalah dengan cara mendeteksi emoticon-emoticon tertentu, yang setiap emoticon memiliki korelasi dengan label.

Percobaan dilakukan menggunakan 300 data yang diambil secara random yang sebelumnya sudah diberi label pada penelitian Alliandu dan Winarko[3]. Pada percobaan ini yang dibandingan adalah tingkat akurasi antara metode emoticon yang digunakan pada penelitian Alliandu dan Winarko [3] dan pelabelan manual dibandingkan dengan metode *lexicon* *based* yang dilakukan pada penelitian ini dan pelabelan manual. Pada pengujian yang dilakukan terhadap metode emoticon dan pelabelan manual memiliki akurasi 0.49 yang secara detail ditampilkan pada Tabel 6.

Tabel . Perbandingan Metode Emoticon dan Manual

|  | **Manual** | | | |
| --- | --- | --- | --- | --- |
| **Emoticon** |  | **Positif** | **Netral** | **Negatif** |
| **Positif** | 40 | 46 | 14 |
| **Netral** | 12 | 57 | 31 |
| **Negatif** | 11 | 40 | 49 |
|  | **63** | **143** | **94** |

Pada pengujian yang dilakukan terhadap metode *lexicon* *based* dan pelabelan manual memiliki akurasi 0.68 yang secara detail ditampilkan pada Tabel 7.

Tabel . Perbandingan Metode Lexicon dan Manual

|  | **Manual** | | | |
| --- | --- | --- | --- | --- |
| **Lexicon** |  | **Positif** | **Netral** | **Negatif** |
| **Positif** | 36 | 13 | 8 |
| **Netral** | 23 | 119 | 37 |
| **Negatif** | 4 | 11 | 49 |
|  | **63** | **143** | **94** |

Berdasarkan pada kedua percobaan yang sudah dilakukan, pelabelan yang dilakukan dengan metode lexicon based ternyata memiliki akurasi yang lebih baik yaitu 0.68 dibandingkan pelabelan yang dilakukan dengan metode emoticon yaitu hanya 0.49.

## Akurasi Pemberian label

Penghitungan akurasi dari penelitian ini adalah dengan membandingkan hasil pelabelan yang dilakukan dengan metode kombinasi antara *lexicon* *based* dan SVM dibandingkan dengan pelabelan yang dilakukan secara manual. Data yang digunakan untuk perbandingan adalah data dari 3 kali percobaan dengan masukan kata kunci yang berbeda-beda. Berdasarkan 3 percobaan yang dilakukan didapatkan 499 data *tweet.* Gambar 8 menunjukan hasil perbandingan yang sudah dilakukan. Pada gambar 6.9 ditunjukan dari 499 *tweet* , jumlah *tweet* yang diberi label oleh sistem dengan benar berjumlah 388 *tweet*. Berdasarkan data tersebut, akurasi dari penelitian ini adalah 0.78

Tabel . Akurasi Sistem

|  | **Manual** | | | |
| --- | --- | --- | --- | --- |
| **Emoticon** |  | **Positif** | **Netral** | **Negatif** |
| **Positif** | 40 | 46 | 14 |
| **Netral** | 12 | 57 | 31 |
| **Negatif** | 11 | 40 | 49 |
|  | **63** | **143** | **94** |

# 

# KESIMPULAN DAN SARAN

## Kesimpulan

Berdasarkan hasil pengujian yang telah diperoleh, dapat ditarik kesimpulan:

1. Pada proses pengumpulan data training metode lexicon based memberikan akurasi yang lebih baik dari pada metode emoticon. Dalam penelitian ini metode lexicon based mencapai akurasi 68% sedangkan metode emoticon hanya mencapai akurasi 49%. Pengujian ini berdasarkan 300 tweet random berbahasa indonesia yang terdapat emoticon pada semua tweet.

2. Pada pengujian akhir metode gabungan antara lexicon based dan support vector machine menghasilkan akurasi 78%. Pengujian ini berdasarkan 499 tweet random berbahasa indonesia yang diambil langsung dari twitter.

## Saran

Pada penelitian ini metode postagging yang digunakan adalah menggunakan dictionary. Ternyata hasil postagging menggunakan dictionary kurang bagus. Padahal postagging digunakan dalam lexicon based yang merupakan metode yang digunakan untuk pengumpulan data training. Untuk itu diperlukan metode postagging yang memiliki akurasi lebih tinggi. Sehingga hasil dari akurasi pengumpulan data dapat lebih baik lagi.

Pada metode lexicon based pencocokan frase dengan data sentimen yang ada pada dictionary bersifat tunggal. Artinya satu kata hanya memiliki satu sentimen. Padahal sangat dimungkinkan satu kata dapat memiliki lebih dari satu sentimen. Untuk itu dibutuhkan dictionary kata sentimen yang masing-masing katanya dapat memiliki lebih dari satu sentimen.

Pada penelitan ini menggunakan data pelatihan sebanyak 25000. Ternyata data pelatihan belumlah cukup. Penambahan data training dapat meningkatkan akurasi dari klasifikasi yang dilakukan dengan SVM

# DAFTAR PUSTAKA

[1]Quirk, R., Greenbaum, S., Leech, G. dan Svartvik, J., 1985. A Comprehensive Grammar of the English Language, Longman, London..

[2]Semiocast, 2012. Brazil becomes 2nd country on Twitter, Japan 3rd Netherlands most active country. http://semiocast.com/en/publications/2012\_01\_31\_Brazil\_becomes\_2nd\_country\_on\_Twitter\_superseds\_Japan, diakses 30 Mei 2012.

[3]Alliandu, P. dan Winarko, E., 2012. Analisis Sentimen Tweet Berbahasa Indonesia di Twitter. Tesis. : Program Pasca Sarjana Universitas Gadjah Mada.Yogyakarta.

[4]Pang, B., Lee, L. dan Vaithyanathan, S., 2002. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. Stroudsburg.

[5]Franky dan Manurung, R., 2008. Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews. In Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications.

[6]Go, A., Bhayani,R. dan Huang, L., 2009. Twitter Sentiment Classification using Distant Supervision, CS224N Project Report, Standford

[7]Pak, A. dan Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. European Language Resources Association, Valletta.

[8]Zhang, C., Zuo, W., Peng, T. dan He, F., 2008. Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel. In Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology. Washington, DC.

[9]Komarsilam, E. dan Winarko, E., 2012. Penambangan Opini Pada Situs Review Film Berbahasa Indonesia, Tesis, Program Pasca Sarjana FMIPA UGM, Yogyakarta.