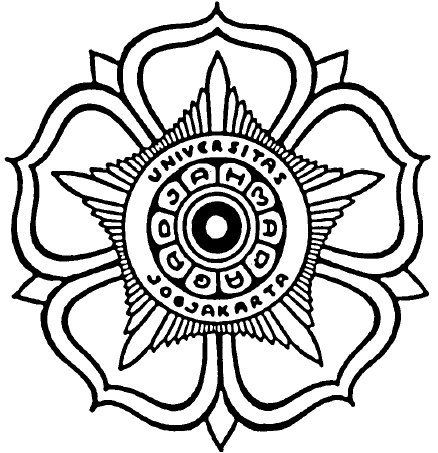
**USULAN PENELITIAN S2**

**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**

**JURUSAN ILMU KOMPUTER DAN ELEKTRONIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS GADJAH MADA**

**YOGYAKARTA**

**2012**

**HALAMAN PERSETUJUAN**

**USULAN PENELITIAN S2**

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

**DIUSULKAN OLEH**

**JAN KRISTANTO WIBISONO**

**10/306141/PPA/03230**

**Telah disetujui**

**pada tanggal Mei 2012**

**Pembimbing**

**Drs. Edi Winarko, M.Sc., PhD**

**NIP. 196302231987031002**

# Latar Belakang

Opini adalah pendapat pribadi yang tidak obyektif dan tidak melalui proses verifikasi (Quirk et al., 1985). Opini orang lain tentang suatu 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.

Sekarang ini *social media* telah berkembang dengan pesat, dimulai dari awal Friendster sampai era sekarang ini Facebook, Twitter, Google+ dan Foursquare. Menurut <http://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 ([http://semiocast.com](http://semiocast.com/)) 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 (aliandu., 2012 , Pang 2002, Franky., 2008, Go et al., 2009, Pak dan Paurobek., 2009) , Support Vector Machine (Pang., 2002 , Franky., 2008 , Zhang., 2008, Go et al., 2009) dan Maximum Entropy(Go et al., 2009) ataupun yang berbasiskan lexicon (komansilan., 2012). 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 (2008) 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.

# Perumusan Masalah

Berdasarkan latar belakang masalah yang telah dipaparkan diatas dapat dirumuskan hal-hal sebagai berikut :

1. Bagaimana melakukan *preprocessing* terhadap kalimat tidak baku.
2. Bagaimana melakukan *opinion mining* pada *social media* untuk bahasa Indonesia.
3. Bagaimana mengkombinasikan Support Vector Machine (SVM) dengan Lexicon Based untuk menentukan suatu kalimat beropini positif, negative atau netral?

# Batasan Masalah

Adapun yang menjadi batasan masalah pada penelitian ini adalah :

1. Bahasa yang digunakan dalam *opinion mining* dalam penelitian ini adalah bahasa Indonesia.
2. Sumber dari *social media* yang digunakan adalah dari Twitter.
3. Untuk menentukan apakah suatu kalimat merupakan kalimat positif, negatif, maupun netral adalah dengan mengkombinasikan antara metode *lexicon-based* dan *learning-based*

# Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah menciptakan mesin yang dapat melakukan *opinion mining* dalam bahasa Indonesia pada *social media* dengan mengkombinasikan metode *lexicon-based* dan *learning-based*.

# Manfaat Penelitian

Penelitian ini diharapkan dapat menghasil mesin yang bisa mengekstrak opini orang lain yang dapat menjadi pertimbangan dalam proses pengambilan keputusan.

# Keaslian Penelitian

Penelitian tentang *opinion mining* dalam bahasa Indonesia pernah dilakukan oleh Franky dan Ruli Manurung (2008). Pada penelitian tersebut sumber opini yang digunakan adalah *movie reviews* bahasa inggris yang diterjemahkan dalam bahasa Indonesia menggunakan mesin penerjemah. Pendekatan yang digunakan adalah *machine learning* dengan tiga teknik klasifikasi. Yaitu, naïve bayes, maximum entropy dan support vector machine. Pada penelitian ini akurasi analisis hasil terjemahan bahasa Indonesia 78.82% sedangkan untuk analisis dalam bahasa inggris 80.09%.

# Tinjauan Pustaka

Pang dkk. (2002) 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 (2008) mencoba mengulangi eksperimen klasifikasi sentimen *movie reviews* oleh Pang et al(2002) untuk 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(2002) 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. Hasil terbaik diperoleh sama dengan yang diperoleh ketika menggunakan eksperimen dalam bahasa inggris

*Opinion mining* dalam bahasa china pernah dilakukan oleh Zhang (2008). Dalam penelitian ini, mereka mempresentasikan hasil kerjanya dengan penekanan *opinion mining* pada *online reviews*. Mereka mengembangkan *opinion mining* tersebut berdasarkan metode *learning*. Menggunakan *dataset* opini yang ada pada AmazonCN. Metode pembelajaran yang digunakan adalah Support Vector Machine dengan menggunakan String kernel.

Pak dan Paurobek (2010) menggunakan *emoticons* untuk membangun *corpus* berbahasa inggris dari twitter dengan sentimen positif, negatif dan netral. Untuk kelas netral pak dan paurobek mengambil data *training* dari tweet akun media berbahasa inggris. Metode yang digunakan adalah Naïve Bayes dengan n-grams. Performasi terbaik dihasilkan ketika menggunakan bigram.

Rasyid dan Purwarianti (2011) menggunakan *machine learning* untuk melakukan klasifikasi opini pada *social media*. Yaitu dengan metode SVM dan Maximum entropy akan mengklasifikasikan menjadi kalimat pertanyaan, sentimen positif, sentiment negatif atau netral. Berdasarkan penelitian yang dilakukan klasifikasi terbaik adalah menggunakan metode SVM dengan akurasi 86,66%.

Go et al. (2009) melakukan *opinion mining* pada *social media* twitter. Memanfaatkan *emoticons* dalam mempermudah pelabelan data pada analisis sentimen dari tweet berbahasa inggris. 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.

Harb et al. (2008) melakukan opinion mining pada media blog. Pendekatan yang dilakukan hanya menghitung jumlah kata sifat positif dan negatif yang terdapat dalam artikel. Kemudian menghitung jumlah perbedaan antara sentimen positif dan negatif. Untuk menambah akurasi sistem, mereka juga melakukan analisa pada kata kerja dan juga kata-kata negasi.

Komansilan dan Winarko (2012) 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.

Aliandu dan Winarko (2012) melakukan opinion mining pada twitter berbahasa Indonesia. Metode yang digunakan adalah *supervised learning* yaitu naïve bayes. 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*.

Perbandingan antara penelitian yang telah dijelaskan diatas dengan sistem yang akan dibangun akan dirangkum dalam tabel 1.

Tabel 1. Perbandingan penelitian lampau dengan sistem yang akan dibangun

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Judul | Penulis & Tahun | Persamaan | Perbedaan |
| 1. | *Thumbs up? Sentiment Classification using Machine Learning Techniques* | Pang et al .,2002 | Metode *learning* yang digunakan Support Vector Machine (SVM) | Pada sistem yang akan dibangun tidak hanya menggunakan metode learning SVM tetapi juga menggunakan *lexicon based*. |
| 2. | *Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews* | Franky dan Ruli .,2008 | Bahasa yang digunakan bahasa Indonesia. Menggunakan metode *learning* SVM | Bahasa Indonesia yang digunakan pada penelitian ini bukan hasil translate dari bahasa inggris dan diambil dari *social media*. |
| 3. | *Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel* | Zhang et al .,2008 | Metode *learning* menggunakan SVM | Bahasa yang digunakan dalam penelitian ini bahasa indoenesia. Metode yang digunakan tidak hanya SVM saja tapi gabungan dengan *lexicon based*. |
| 4. | *Twitter as a Corpus for Sentiment Analysis and Opinion Mining* | Pak dan Paroubek .,2010 | Menganalisis dari sumber twitter | Bahasa yang digunakan dalam penetian ini bahasa indonesia.  Metode yang digunakan gabungan antara *lexicon based* dan *machine learning* |
| 5. | *Sentiment Classification for Indonesian Message in Social Media* | Rasyid dan Purwarianti .,2011 | Sumber yang digunakan dari *social media* berbahasa indonesia | Penelitian ini metode yang digunakan gabungan antara *lexicon based* dan *machine learning*. |
| 6. | *Twitter Sentiment Classification using Distant Supervision* | Go, et al ., 2009 | Menganalisis dari sumber twitter | Penelitian ini dengan tidak hanya dengan *machine learning* saja tetapi dengan *lexicon based* juga |
| 7. | *Web Opinion Mining: How to extract opinions from blogs?* | Harb, et al., 2008 | Metode yang digunakan lexicon based dan pada media online. | Selain *lexicon based* juga menggunakan *machine learning*. |
| 8. | Penambangan Opini Pada situs Review Film Berbahasa Indonesia | Komansilan dan Winarko., 2012 | Analisis pada bahasa Indonesia. | Selain menggunakan *lexicon based* penelitian ini juga menggunakan *machine learning*.  Pada penelitian ini sumber data yang digunakan dari twiiter. |
| 9. | Analisis Sentimen Tweet Berbahasa Indonesia di Twitter | Aliandu dan Winarko., 2012 | Sumber data yang digunakan dari twiiter.  Analisis pada bahasa Indonesia. | Metode yang digunakan pada penelitian ini gabungan SVM dan lexicon. |

# Metodologi Penelitian

Metode penelitian yang akan dilakukan terdiri atas beberapa tahap sebagai berikut :

1. Studi Kepustakaan

Melakukan berbagai macam pengumpulan bahan referensi, seperti jurnal penelitian, tesis, buku-buku teori dan sumber lain termasuk informasi yang diperoleh dari internet. Mempelajari litelatur dan teori pendukung penelitian yaitu mengenai klasifikasi khususnya SVM (Support Vector Machine), Twitter API.

1. Analisis Sistem

Melakukan analisa terhadap requirement yang dibutuhkan dalam pengembangan sistem.

1. Perancangan

Pada tahap ini akan dilakukan perancangan metode berdasar lexicon, dan *machine learning*, yaitu Support Vector Machine (SVM) yang akan diimplementasikan untuk *Opinion mining*. Data-data yang diambil dalam penetian ini diambil dari twitter dengan memanfaatkan Twitter API yang sudah disediakan, selanjutnya data tersebut disimpan dalam basis data. Langkah pertama dalam proses besar penelitian ini adalah melakukan *preprocessing* data tweet. Setelah dilakukan *preprocessing*, sistem akan mencari kalimat yang mengandung frasa opini, sehingga kalimati yang tidak mengandung frasa opini tidak diikutkan dalam proses selanjutnya. Kemudian sistem akan menentukan sentimen tweet tersebut menggunakan metode *lexicon based* apakah positif, negatif atau netral. Untuk menambah keakuratan, sistem akan menentukan kembali sentimen tweet tersebut kali ini dengan *machine learning* yaitu dengan metode Support Vector Machine (SVM). Jika antara metode *lexicon based* dan *machine learning* menghasilkan sentimen yang sama akan sentimen tersebut yang menjadi kesimpulan akhir. Jika kedua metode tersebut menghasilkan sentimen yang tidak sama, misal *lexicon based* menghasilkan sentimen positif dan *machine learning* menghasilkan sentimen negatif maka yang akan digunakan adalah hasil sentimen dari *machine learning*. Tetapi salah satu menghasilkan sentiment positif atau negatif sedang yang lainnya menghasilkan netral maka sentiment bukan netral yang akan dipakai menjadi kesimpulan akhir.

Secara garis besar gambaran umum dari penelitan ini ditunjukan seperti gambar 1:

Crawl tweets

Preprocessing

Lexicon based method

Machine Learning Method

Classifier sentiment

Gambar 1. Gambaran Umum Sistem yang akan dibangun

1. Implementasi

Pada tahap ini akan dilakukan pembuatan kode program sampai proses pembangunan basis data.

1. Pengujian

Pada tahap ini, dilakukan perhitungan akurasi dari implementasi yang telah dilakukan. Metode yang digunakan adalah *precision , recall* dan *F-score*. Menurut Risbergen (1979), *Precision* adalah perbandingan antara jumlah kalimat yang mengandung opini baik sentimen positif maupun negatif yang dapat dikelompokan oleh sistem dengan jumlah seluruh kalimat yang diambil oleh system. Sedangkan *Recall* adalah perbandingan antara jumlah kalimat yang mengandung opini, baik sentimen positif atau negatif yang dapat dikelompokan oleh sistem dengan jumlah kalimat yang mengandung opini, baik positif maupun negatif yang berada dalam koleksi dokumen. *F-score* adalah 2 kali perbandingan antara perkalian nilai *Precision* dan nilai *Recall* dengan penjumlahan antara *Precision* dan *Recall*.

# Jadwal Penelitian

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No. | Kegiatan | Bulan | | | | | | | | | | | | | | | | | | | | | | | |
| Juni | | | | Juli | | | | Agustus | | | | September | | | | Oktober | | | | November | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1. | Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. | Pengumpulan  Corpus |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3. | Analisis &  Perancangan |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. | Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. | Pengujian |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# Daftar Pustaka

Aliandu,P., dan Winarko,E., 2012 *Analisis Sentimen Tweet Berbahasa Indonesia di Twitter*. Universitas Gadjah Mada.

Franky dan Manurung, R., 2008 *Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews*. University of Indonesia.

Go, A., Bhayani,R. dan Huang, L., 2009, *Twitter Sentiment Classification using Distant Supervision*, CS224N Project Report, Stanford.

Harb, A., Plantie, M., dan Dray, G. 2008, *Web Opinion Mining: How to extract opinions from blogs?* inCSTST’08 : International Conference on Soft Computing as Transdisciplinary Science and Technology

Komansilan,E., dan Winarko,E., 2012 *Penambangan Opini Pada situs Review Film Berbahasa Indonesia*. Universitas Gadjah Mada.

Pang B., Lee, L., dan Vaithyanathan S. , 2002, *Thumbs up? Sentiment Classification using Machine Learning Techniques. In Procceedings of the Conference on Empirical Methods in Natural Language Processing*.

Pak,A., dan Paurobek, P., 2010, *Twitter as a Corpus for sentiment Analysis and Opinion Mining*, Universite de Paris-sud, Laboratoire LIMSI-CNRS. Batiment 508, F-91405 Orsay Cedex, France

Quirk, R., GreenBaum, S., Leech, G., dan Svantvik, J., *A Comprehensive Grammar of The English Language*, In London, 1985.

Rasyi,A., Purwarianti,A., 2011, *Sentiment Classification for Indonesian Message in Social Media, School of Electrical and Informatics Engineering*, Bandung Institute of Technology

Risbergen, V. *Information Retrival, 2nd edition*. In Butterworths, London,1979.

Semiocast,semiocast.com/publications/2012\_01\_31\_Brazil\_becomes\_2nd\_country\_on\_Twitter\_superseds\_Japan, diakses 30 Mei 2012

Socialbakers, [www.socialbakers.com/facebook-statistics/](http://www.socialbakers.com/facebook-statistics/), diakses 30 Mei 2012

Zhang, C., Zuo, W., Peng, Tao., dan He, Fengling., 2008 *Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel. In International Conference on Convergence and Hybrid Information Technology*