**TESIS**

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

**OPINION MINING ON TWITTER FOR INDONESIAN USING SUPPORT VECTOR MACHINE AND BASED LEXICON METHOD**

Diajukan untuk memenuhi salah satu syarat memperoleh derajat

Master of Computer Science

****

**JAN KRISTANTO WIBISONO**

**10/306141/PPA/03230**

**PROGRAM STUDI S2 ILMU KOMPUTER**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS GADJAH MADA**

**YOGYAKARTA**

**2013**

**HALAMAN PERNYATAAN**

Dengan ini saya menyatakan bahwa Tesis ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar Master di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 2, Juni 2013

JAN KRISTANTO WIBISONO

**HALAMAN PENGESAHAN**

**tesis**

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

**oleh**

**Nama Mahasiswa**

**Nomor Pokok Mahasiswa (NPM)**

telah dipertahankan di depan Tim Penguji

pada tanggal ........(lengkapi *tanggal bulan tahun*)

**Susunan Tim Penguji**

**Ketua, Sekretaris,**

**Nama .......... Nama...................**

**NIP. .... NIP. ....**

**Penguji Utama, Penguji Sekretaris,**

**Nama .......... Nama...................**

**NIP. .... NIP. ....**

Diterima dan dinyatakan memenuhi syarat kelulusan pada tanggal .....

**Ketua Jurusan .....**

**Fakultas ....**

**Nama.............**

**NIP.................**

# KATA PENGANTAR

Segala puji bagi Allah SWT., Pencipta dan Pemelihara alam semesta, shalawat serta salam semoga terlimpah bagi Muhammad SAW., keluarga dan para pengikutnya yang setia hingga akhir masa.

Atas rahmat Allah SWT., akhirnya Penulis dapat menyelesaikan tesis ini, meskipun proses belajar sesungguhnya tak akan pernah berhenti. Tesis ini sesungguhnya bukanlah sebuah kerja individual dan akan sulit terlaksana tanpa bantuan banyak pihak yang tak mungkin Penulis sebutkan satu persatu, namun dengan segala kerendahan hati, Penulis mengucapkan terima kasih kepada …….. selama melaksanakan studi dan menyelesaikan tesis ini.

…..

Akhirnya, Penulis berharap semoga penelitian ini menjadi sumbangsih yang bermanfaat bagi dunia sains dan teknologi di Indonesia, khususnya disiplin keilmuan yang Penulis dalami.

Yogyakarta, Juni 2013

Penyusun

# DAFTAR ISI

[KATA PENGANTAR iv](#_Toc359969091)

[DAFTAR ISI v](#_Toc359969092)

[DAFTAR TABEL ix](#_Toc359969093)

[DAFTAR GAMBAR x](#_Toc359969094)

[Intisari xi](#_Toc359969095)

[ABSTRACT xii](#_Toc359969096)

[BAB I 1](#_Toc359969097)

[PENDAHULUAN 1](#_Toc359969098)

[1.1 Latar Belakang 1](#_Toc359969099)

[1.2 Rumusan Masalah 2](#_Toc359969100)

[1.3 Batasan Masalah 2](#_Toc359969101)

[1.4 Tujuan Penelitian 3](#_Toc359969102)

[1.5 Manfaat Penelitian 3](#_Toc359969103)

[1.6 Metode Penelitian 3](#_Toc359969104)

[1.7 Sistematika Penulisan 5](#_Toc359969105)

[BAB II 7](#_Toc359969106)

[TINJAUAN PUSTAKA 7](#_Toc359969107)

[BAB III 11](#_Toc359969108)

[LANDASAN TEORI 11](#_Toc359969109)

[3.1 Natural Language Processing 11](#_Toc359969110)

[3.2 Part-of-speech 12](#_Toc359969111)

[3.2.1 Part-of-speech Tagging 12](#_Toc359969112)

[3.2.2 Tagset 13](#_Toc359969113)

[3.3 Opinion Mining 13](#_Toc359969114)

[3.4 Support Vector Machine 15](#_Toc359969115)

[3.4.1 Konsep SVM 15](#_Toc359969116)

[3.4.2 SVM linear 16](#_Toc359969117)

[3.4.3 Hyperplane SVM 18](#_Toc359969118)

[3.4.4 SVM Nonlinear 21](#_Toc359969119)

[3.5 TF-IDF 23](#_Toc359969120)

[BAB IV 25](#_Toc359969121)

[ANALISIS DAN RANCANGAN SISTEM 25](#_Toc359969122)

[4.1 Analisis Sistem 25](#_Toc359969123)

[4.1.1 Deskripsi Sistem 25](#_Toc359969124)

[4.2 Analisis Data 27](#_Toc359969125)

[4.3 Preprocessing 27](#_Toc359969126)

[4.3.1 Clear Invalid UTF-8 28](#_Toc359969127)

[4.3.2 Casefolding 28](#_Toc359969128)

[4.3.3 Remove Symbol 28](#_Toc359969129)

[4.3.4 Processing Numbers 28](#_Toc359969130)

[4.3.5 Remove Repeat 29](#_Toc359969131)

[4.3.6 Replace Slang 30](#_Toc359969132)

[4.3.7 Spelling Correction 30](#_Toc359969133)

[4.4 Penentuan Label Data Training 31](#_Toc359969134)

[4.4.1 POS taging 31](#_Toc359969135)

[4.4.2 Opinion Rules 32](#_Toc359969136)

[4.4.3 Check Negation 33](#_Toc359969137)

[4.5 Pelatihan 34](#_Toc359969138)

[4.5.1 Generate Weight 35](#_Toc359969139)

[4.5.2 Traning SVM 36](#_Toc359969140)

[4.6 Penentuan Label dengan SVM 36](#_Toc359969141)

[4.6.1 Pencarian Tweets 37](#_Toc359969142)

[4.6.2 Cek Frase 38](#_Toc359969143)

[4.6.3 Generate bobot 38](#_Toc359969144)

[4.6.4 Penentuan label dengan SVM 38](#_Toc359969145)

[4.6.5 Pemetaan dalam sentimen 38](#_Toc359969146)

[4.7 Arsitektur Sistem 39](#_Toc359969147)

[4.8 Rancangan Sistem 41](#_Toc359969148)

[4.8.1 Class Diagram 41](#_Toc359969149)

[4.8.2 Sequence diagram proses pencarian data twitter 43](#_Toc359969150)

[4.8.3 Sequence diagram preprocessing 43](#_Toc359969151)

[4.8.4 Sequence diagram proses pemberian label dengan metode lexicon 44](#_Toc359969152)

[4.8.5 Sequence diagram proses pemberian label dengan SVM 47](#_Toc359969153)

[BAB V 48](#_Toc359969154)

[IMPLEMENTASI 48](#_Toc359969155)

[5.1 Preprocessing Data 48](#_Toc359969156)

[5.1.1 clearInvalidUTF8 48](#_Toc359969157)

[5.1.2 Casefolding 49](#_Toc359969158)

[5.1.3 removeSymbol 49](#_Toc359969159)

[5.1.4 ProcessNumbers 50](#_Toc359969160)

[5.1.5 RemoveRepeat 51](#_Toc359969161)

[5.1.6 *Replace* *informal* *word* 52](#_Toc359969162)

[5.1.7 *Spelling* *Corrention* 53](#_Toc359969163)

[5.2 POS taging 54](#_Toc359969164)

[5.3 Pelabelan menggunakan metode *lexicon* *based* 55](#_Toc359969165)

[5.4 Pembobotan data 58](#_Toc359969166)

[5.5 Pelatihan 61](#_Toc359969167)

[5.6 Pencarian *Tweet* 62](#_Toc359969168)

[5.7 Pengecekan frase opini 63](#_Toc359969169)

[5.8 Pelabelan dengan metode SVM 64](#_Toc359969170)

[BAB VI 65](#_Toc359969171)

[HASIL PENELITIAN DAN PEMBAHASAN 65](#_Toc359969172)

[6.1 Perbandingan proses pelabelan data pelatihan 66](#_Toc359969173)

[6.1.1 Hasil perbandingan pelabelan data pelatihan 66](#_Toc359969174)

[6.2 Pemberian Label 67](#_Toc359969175)

[6.3 Akurasi Pemberian Label 70](#_Toc359969176)

[BAB VII 72](#_Toc359969177)

[KESIMPULAN DAN SARAN 72](#_Toc359969178)

[7.1 Kesimpulan 72](#_Toc359969179)

[7.2 Saran 72](#_Toc359969180)

[DAFTAR PUSTAKA 74](#_Toc359969181)

\*) untuk meng-update isi daftar lakukan klik kanan pada salah satu judul, kemudian klik **Update Field,** selanjutnya **Update Entire Table**, lalu klik **OK**.

# DAFTAR TABEL

[Tabel 2.1 Perbandingan penelitian lampau dengan sistem yang akan dibangun 9](#_Toc359972526)

[Tabel 3.1 Contoh Tags untuk bahasa Indonesia 13](#_Toc359972527)

[Tabel 3.2 Fungsi Kernel 23](#_Toc359972528)

[Tabel 4.1 Daftar Konversi 29](#_Toc359972529)

[Tabel 4.2 Contoh Kopnversi Angka 29](#_Toc359972530)

[Tabel 4.3 Contoh menghapus karakter berualng 30](#_Toc359972531)

[Tabel 4.4 Contoh Pengantian kata tidak baku 30](#_Toc359972532)

[Tabel 4.5 Contoh SpellCorrection 31](#_Toc359972533)

[Tabel 4.6 Opinion Rules 33](#_Toc359972534)

\*) untuk meng-update isi daftar lakukan klik kanan pada salah satu judul, kemudian klik **Update Field,** selanjutnya **Update Entire Table**, lalu klik **OK**.

# DAFTAR GAMBAR

[Gambar 1.1 Gambaran Umum Sistem 5](file:///D:\yan\thesis\laporanthesis\Laporan%20Thesis%20Jan.docx#_Toc359972770)

\*) untuk meng-update isi daftar lakukan klik kanan pada salah satu judul, kemudian klik **Update Field,** selanjutnya **Update Entire Table**, lalu klik **OK**.

# Intisari

Tulislah isi abstrak dalam bahasa Indonesia di sini. Tulislah isi abstrak dalam bahasa Indonesia di sini. Tulislah isi abstrak dalam bahasa Indonesia di sini. Tulislah isi abstrak dalam bahasa Indonesia di sini. Tulislah isi abstrak dalam bahasa Indonesia di sini. Tulislah isi abstrak dalam bahasa Indonesia di sini. Tulislah isi abstrak dalam bahasa Indonesia di sini.

# ABSTRACT

Tulislah isi abstrak dalam bahasa Inggris di sini. Tulislah isi abstrak dalam bahasa Inggris di sini. Tulislah isi abstrak dalam bahasa Inggris di sini. Tulislah isi abstrak dalam bahasa Inggris di sini. Tulislah isi abstrak dalam bahasa Inggris di sini. Tulislah isi abstrak dalam bahasa Inggris di sini. Tulislah isi abstrak dalam bahasa Inggris di sini.

# 

# PENDAHULUAN

## Latar Belakang

Opini adalah pendapat pribadi yang tidak obyektif dan tidak melalui proses verifikasi (Quirk et al., 1985). 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](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 ([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.

## Rumusan Masalah

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

1. Bagaimana melakukan *opinion mining* pada *social media* untuk bahasa Indonesia.
2. Bagaimana mengumpulkan data pelatihan secara otomatis.
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 mengembangkan metode untuk melakukan *opinion mining* dalam bahasa Indonesia pada *social media* dengan mengkombinasikan metode *lexicon-based* dan *learning-based*.

## Manfaat Penelitian

Penelitian ini diharapkan dapat menghasil metode yang dapat mesin dapat mengekstrak opini dari *social media* yang dapat menjadi pertimbangan dalam proses pengambilan keputusan.

## Metode 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 mengenai klasifikasi, khususnya SVM (Support Vector Machine).

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 kalimat yang tidak mengandung frasa opini akan langsung dimasukan dalam kelas netral. Kemudian sistem akan menentukan sentimen tweet yang memiliki frase opini tersebut menggunakan metode *lexicon based* apakah positif, negatif atau netral. Hasil dari proses lexicon based ini akan menjadi data training pada proses penentuan sentimen tweet baru yang kali ini akan dilakukan dengan *machine learning* yaitu dengan metode Support Vector Machine (SVM).

Secara garis besar gambaran umum dari penelitan ini ditunjukan seperti Gambar 1.1 :

Gambar 1. Gambaran Umum Sistem

Testing

Training

Classifier sentiment

Crawl tweets

Preprocessing

Lexicon based method

Machine Learning Method

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. Pengujian yang dilakukan adalah perbandingan akurasi antara metode pengumpulan data *training* antara metode *emoticon* dan metode *lexicon* *based*. Selain itu juga dilakukan pengujian akurasi dari metode yang telah di implementasikan pada penelitian ini yaitu metode *lexicon* *based* dan SVM.

## Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini memuat latar belakang masalah, rumusan masalah, batasan masalah,tujuan penelitian, manfaat penelitian, metode penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini memuat penelitian-penelitian terdahulu yang berkaitan dengan penelitian yang dilakukan penulis. Beberapa penelitian terkait Opinion Miningatau *Sentiment analysis* dibahas secara singkat dan dibandingkan dengan penelitian penulis.

BAB III LANDASAN TEORI

Bab ini membahas teori-teori yang berkaitan dengan penelitian. Bagian pertama membahas tentang *Natural* *Language* *Processing*, Postagging, *Opinion* *Mining*, permasalahannya, dan beberapa pendekatan yang digunakan. Bagian kedua membahas tentang *Support* *Vector* *Machine* (SVM), *vector* *space* *model* dan bagaimana model tersebut bekerja.

BAB IV ANALISIS DAN RANCANGAN SISTEM

Bab ini membahas analisis kebutuhan data, analisis model beserta rancangan sistemnya.

BAB V IMPLEMENTASI

Bab ini membahas implementasi opinion mining dengan metode *lexicon* *based* dan *Support* *Vector* *Machine* sesuai rancangan pada bab sebelumnya.

BAB VI HASIL DAN PEMBAHASAN

Bab ini membahas percobaan yang dilakukan pada proses pelatihan dan penentuan kelas beserta uraian mengenai hasil dan perbandingannya.

BAB VII KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan-kesimpulan dari hasil penelitian dan saran-saran yang berguna untuk penelitian selanjutnya.

# 

# 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 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.

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.

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.

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%.

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 2.1.

Tabel 2. Perbandingan penelitian lampau dengan sistem yang akan dibangun

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Penulis & Tahun | Persamaan | Perbedaan |
| 1. | 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. | Harb, et al., 2008 | Metode yang digunakan lexicon based dan pada media online. | Selain *lexicon based* juga menggunakan *machine learning*. |
| 3. | 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*. |
| 4. | 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*. |
| 5. | Go, et al ., 2009 | Menganalisis dari sumber twitter | Penelitian ini dengan tidak hanya dengan *machine learning* saja tetapi dengan *lexicon based* juga |
| 6. | 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* |
| 7. | Rasyid dan Purwarianti .,2011 | Sumber yang digunakan dari *social media* berbahasa indonesia | Penelitian ini metode yang digunakan gabungan antara *lexicon based* dan *machine learning*. |
| 8. | 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. | Aliandu dan Winarko., 2012 | Sumber data yang digunakan dari twiiter.  Analisis pada bahasa Indonesia. | Metode yang digunakan pada penelitian ini gabungan SVM dan lexicon. |

# 

# LANDASAN TEORI

## Natural Language Processing

Natural Language Processing (NLP) adalah salah satu bidang ilmu komputer, kecerdasan buatan, dan bahasa (linguistik) yang berkaitan dengan interaksi antara komputer dan bahasa alami manusia, seperti bahasa Indonesia atau bahasa Inggris. Tujuan utama dari studi NLP adalah membuat mesin yang mampu mengerti dan memahami makna bahasa manusia lalu memberikan respon yang sesuai (Kao dan puteet, 2006). Secara umum, Jenis aplikasi yang bisa dibuat dalam bidang ilmu NLP terbagi dua, yaitu *text-based application* dan *dialogue-based application*.

*Text-based application* adalah segala macam aplikasi yang melakukan proses terhadap teks tertulis seperti misalnya dokumen, e-mail, buku dan sebagainya. Beberapa jenis aplikasi NLP yang berbasis teks : Program yang mampu mengklasifikasi dan mengambil isi dari suatu dokumen berdasarkan kontennya. Seperti spam filtering (pemfilteran pesan sampah), language identification (identifikasi bahasa), dan Machine Translation,Program yang mampu mentranslasi kalimat baik berupa teks maupun suara dari satu bahasa alami ke bahasa lainnya. Contoh : Google Translate.

*Dialogue-based application* idealnya melibatkan bahasa lisan atau pengenalan suara, akan tetapi bisa juga memasukan interaksi dialog dengan mengetikkan teks pertanyaan melalui *keyboard*. Contoh : Perangkat lunak yang mampu melakukan tugas-tugas dan jasa berdasarkan inputan dari pengguna, lokasi, dan memiliki kemampuan untuk mengakses informasi dari berbagai sumber online (seperti cuaca, keadaan lalu lintas, berita, saham, dll). Contohnya adalah Siri pada produk-produk Apple dan S-Voice pada produk-produk seluler Samsung.

Chatbot adalah program komputer yang didesain untuk mensimulasikan sebuah percakapan cerdas dengan satu atau lebih pengguna manusia melalui inputan suara atau teks, utamanya digunakan untuk percakapan kecil. Contoh : Cleverbot, SimSimi, dan begobet.

## Part-of-speech

*Part-of-speech* menunjukkan kelas atau jenis kata dalam sebuah kalimat. Menurut Jurafsky dan Martin (2008), *part-of-speech* dapat dibedakan menjadi dua buah kelas utama, yaitu tipe kelas tertutup dan tipe kelas terbuka. Tipe kelas tertutup terdiri atas *part-of-speech* yang memiliki anggota relatif tetap. Contoh kelas ini adalah preposisi. Tipe kelas terbuka, sebaliknya, terdiri atas kata-kata yang relatif terus mengalami penambahan karena terbentuknya kata baru akibat serapan dari bahasa asing. Ada empat *part-of-speech* tipe kelas terbuka utama yang terdapat dalam hampir semua bahasa di dunia, yaitu kata benda, kata sifat, kata kerja, dan kata keterangan. Kelas kata lainnya merupakan sub bagian dari keempat *part-of-speech* tersebut.

### Part-of-speech Tagging

*Part-of-speech* (POS) *tagging* adalah proses pemberian label pada setiap kata dalam kalimat dengan anotasi *part-of-speech* yang bersesuaian (Manning dan Schütze, 1999). Anotasi POS menunjukkan jenis kata dalam sebuah kalimat. Masukan dalam pemberian anotasi adalah rangkaian kata beserta sekumpulan anotasi yang mungkin. Keluaran dari proses ini adalah anotasi tunggal terbaik yang sesuai untuk setiap kata dalam rangkaian masukan.

Menurut Jurafsky dan Martin (2008), POS *tagging* memiliki beberapa manfaat dalam berbagai penelitian pemrosesan bahasa. POS *tagging* dapat digunakan untuk membuat model bahasa untuk pengenalan wicara (*speech recognition*). Pemberian anotasi juga membantu dalam menentukan pelafalan sebuah kata dalam sistem sintesis wicara. Dalam bidang temu kembali informasi, POS *tagging* dapat membantu proses *stemming* atau pemilihan kata-kata penting tertentu dalam sebuah dokumen.

### Tagset

Hal yang perlu menjadi pertimbangan dalam POS *tagging* adalah *tagset*, yaitu anotasi *part-of-speech* mana saja yang akan digunakan. Ukuran dan isi dari tagset ini perlu diperhitungkan. Menurut Güngör (2010), penggunaan *tagset* berukuran besar membantu mendapatkan pengetahuan tentang struktur morfologi dan morfosintaksis dari kata, namun dapat mempersulit untuk membedakan kemiripan antar anotasi. Perbedaan anotasi, pada kasus-kasus tertentu, sangatlah kecil sehingga terkadang terjadi perbedaan pendapat dalam penggunaan satu anotasi dengan anotasi lainnya. Contoh *tagset* untuk bahasa Indonesia dapat dilihat pada Tabel 3.1.

Tabel 3. Contoh Tags untuk bahasa Indonesia

|  |  |  |
| --- | --- | --- |
| Anotasi | Definisi | Contoh |
| JJ | Kata Sifat | Baik, Kejam |
| NN | Kata Benda | Mobil, Buku |
| VB | Kata Kerja | Pergi, membeli |
| RB | Kata Keterangan | Sementara, Nanti |
| IN | Kata Depan | Di, Ke, Dari |
| MD | Kata Bantu | Masih, Telah |
| UH | Kata Seru | Ayo |
| CC | Kata Sambung Koordinasi | Atau, Bahwa |
| PR | Kata Ganti | Anda, Dia |
| DT | Kata Tunjuk | Ini, Itu |
| CK | Kata sambung khusus | Dengan, Yang |
| RP | Partikel | Doang, Dong |

## Opinion Mining

*Sentiment analysis* atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *text* *mining*. Secara umum, bertujuan untuk menentukan pendapat pembicara atau penulis berkenaan dengan topik tertentu.

Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur / tingkat aspek apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas / aspek bersifat positif , negatif atau netral. Jika ditinjau lebih dalam lagi sentiment analysis dapat menyatakan emosional sedih, gembira, atau marah.

Sebuah metode yang berbeda untuk menentukan sentimen adalah penggunaan sistem skala dimana kata-kata umumnya terkait memiliki sentimen negatif, netral atau positif dengan mereka diberi nomor pada skala -5 sampai +5 (paling negatif hingga yang paling positif) dan ketika sepotong teks terstruktur dianalisis dengan pemrosesan bahasa alami, konsep selanjutnya dianalisis untuk memahami kata-kata ini dan bagaimana mereka berhubungan dengan konsep.

Setiap konsep kemudian diberi skor berdasarkan bagaimana kata-kata sentimen berhubungan dengan konsep, dan skor yang terkait. Hal ini memungkinkan gerakan untuk pemahaman yang lebih canggih dari sentimen berdasarkan skala 11 titik.

Penelitian lain arah berbeda diantaranya adalah identifikasi subjektivitas / objektivitas. Tugas ini biasanya didefinisikan sebagai menggolongkan suatu teks yang diberikan (biasanya kalimat) ke salah satu dari dua kelas: objektif atau subjektif (Pang, B. & Lee, L, 2008). Masalah ini kadang-kadang dapat lebih sulit daripada klasifikasi polaritas. Subjektivitas kata-kata dan frase mungkin tergantung pada konteks dan dokumen objektif mungkin berisi kalimat subjektif (misalnya, sebuah artikel berita mengutip pendapat orang).

Orang sering kali menyatakan lebih dari satu opini "the movie was terrible, but DeNiro's performance was superb, as always", sebuah sarkasme "this is probably the best laptop Dell could come up with", atau menggunakan negasi dan banyak elemen kompleks sehingga sulit untuk diparsing "not that I'm saying this was a bad experience".

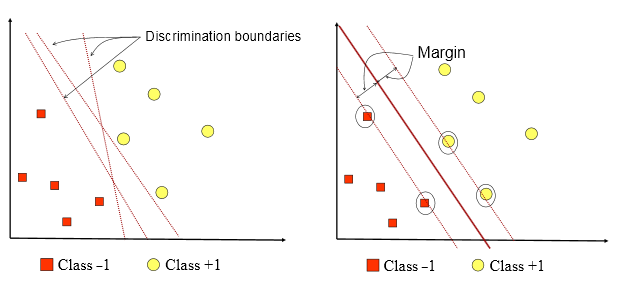
Ekspresi atau sentiment mengacu pada fokus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subject yang berbeda. Sebagai contoh, adalah hal yang baik untuk mengatakan alur film tidak terprediksi, tapi adalah hal yang tidak baik jika ‘tidak terprediksi’ dinyatakan pada kemudi dari kendaraan. Bahkan pada produk tertentu, kata-kata yang sama dapat menggambarkan makna kebalikan, contoh adalah hal yang buruk untuk waktu start-up pada kamera digital jika dinyatakan “lama”, namun jika” lama” dinyatakan pada usia batere maka akan menjadi hal positif. Oleh karena itu pada beberapa penelitian, terutama pada review produk, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses opinion mining.

## Support Vector Machine

Support Vector Machine (SVM) kini banyak dikembangkan dan diterapkan baik untuk klasifikasi maupun regresi. Metode ini berakar dari teori pembelajaran statistik yang hasilnya sangat menjanjikan untuk memberikan hasil yang lebih baik daripada metode lainnya. SVM juga bekerja dengan baik pada set data berdimensi tinggi.

### Konsep SVM

Ide dasar SVM adalah memaksimalkan batas hyperplane, seperti yang diilustrasikan pada gambar 3.1. pada Gambar 3.1 (a) ada sejumlah pilihan hyperplane yang mungkin untuk data set, dan pada Gambar 3.1 (b) merupakan hyperplane dengan margin yang paling maksimal.Hyperplane yang maksimal akan memberikan generalisasi yang lebih baik pada metode klasifikasi.



**Gambar 3.1 Decision Boundary**

Konsep klasifikasi dengan SVM dapat dijelaskan secara sederhana sebagai usaha untuk mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah kelas data pada ruang input (Nugroho, 2007). Gambar 3.1 memperlihatkan beberapa pola yang merupakan anggota dari dua buah kelas data : +1 dan -1. Data yang tergabung dalam kelas -1 disimbolkan dengan bentuk bujur sangkar sedangkan data pada kelas +1 disimbolkan dengan bentuk lingkaran.

Hyperplane pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin hyperplane tersebut dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan data terdekat dari masing-masing kelas. Data yang paling dekat ini disebut *support vector.* Garis solid pada gambar 3.1 menunjukan hyperplane terbaik yaitu yang terleat tepat pada tengah-tengah kedua kelas, sedangkan data lingkaran dan bujur sangkar yang dilewati garis batas margin (garis putus-putus) adalah *support vector*. Usaha untuk mencari lokasi hyperplane ini merupakan inti dari proses pelatihan pada SVM.

### SVM linear

Setiap ada latih dinyatakan oleh ( , ) dimana i = 1,2,...., N , dan = merupakan atribut (fitur) set untuk data latih ke-i € {-1,+1} menyatakan label kelas. Hyperplane klasifikasi linear SVM seperti pada gambar 3.1, dinotasikan dengan

w. + b = 0 ( 3.1 )

w dan b adalah parameter model w. merupakan inner product dalam antara w dan . Data yang masuk kelas -1 adalah data yang memenuhi pertidaksamaan berikut :

w. + b ≤ -1 ( 3.2 )

Sementara yang masuk dalam kelas +1 adalah data yang memenuhi pertidaksamaan berikut:

w. + b ≥ +1 ( 3.3 )

Sesuai pada gambar 3.2, jika ada data dalam kelas -1 misalnya yang bertempat di hyperplane, persamaan (1) akan terpenuhi. Untuk data kelas -1 dinotasikan dengan :

w. + b = 0 ( 3.4 )

sementara kelas +1 misal akan memenuhi persamaan

w. + b = 0 ( 4.5 )

Dengan mengurangkan persamaan 5 dan 4, didapatkan

w.( - ) = 0 ( 4.6 )

- adalah vektor pararel di posisi hyperplane dan diarahkan dari ke . Karena inner product dalam bernilai 0, arah dari 2 harus tegak lurus terhadap hyperplane (Gambar 3.1).

Dengan memberikan label -1 untuk kelas pertama dan +1 untuk label kelas kedua, prediksi semya data uji akan menggunakan formula

Susuai pada gambar 3.2, hyperplane untuk kelas -1 adalah data pada *support vector* yang memenuhi persamaan

w. + b = -1 ( 3.7 )

Sementara hyperplane kelas +1 memenuhi persamaan

w. + b = +1 ( 3.8 )

dengan demikian margin dapat dihitung dengan mengurangi Persamaan 8 dan 7 didapatkan

w. (- . ) = 2 ( 3.9 )

Margin hyperplane diberikan oleh jarak antara dua hyperplane dari dua kelas tersebut. Notasi di atas diringkas menjadi

||w|| x d = 2 atau d = ( 3.10 )

### Hyperplane SVM

Klasifikasi kelas data pada SVM persamaan 2 dan 3 dapat digabungkan dengan notasi

( w. + b) ≥ 1 , i =1,2.....,N ( 3.11 )

Margin optimal dihitung dengan memaksimalkan jarak antara hyperplane dan data terdekat. Jarak ini dirumuskan dengan persamaan 10. Selanjutnya, masalah ini diformulasikan ke dalam quadratic programming (QP) problem, dengan meminimalkan invers Persamaan 10, , dengan syarat

Minimalkan :

( 3.12 )

Syarat

( w. + b) ≥ 1 , i =1,2.....,N ( 3.13 )

Optimalisasi ini dapat diselesaikan dengan Lagrange Multiplier

Lp = - ( 3.14 )

adalah Lagrange Multiplier yang berkorespindensi dengan . Nilai adalah nol atau positif. Untuk meminimalkan Lagrangian, Persamaan 14 harus diturunkan pada w dan b, dan diset dengan nilai 0 untuk syarat optimalisasi diatas.

Syarat 1 :

= 0 => w = ( 3.15 )

Syarat 2 :

= 0 => = 0 ( 3.16 )

N adalah jumlah data yang menjadi support vector. Karena Lagrage Multiplier () tidak diketahui nilainya, persamaan diatas tidak dapat diselesaikan secara langsung untuk mendapatkan w dan b. Untuk menyelesaikan masalah tersebut, modifikasi persamaan 14 menjadi kasus pemaksimalan, syarat untuk dualitasnya menggunakan constraint karush-Kuhn-Tucker (KKT) berikut

Syarat 1 :

( 3.17 )

Syarat 2 :

≥ 0, i=1,2,...,N ( 3.18 )

Dengan menerapakan constraint pada persamaan 17 dan 18, dipastikan bahwa nilai Lagrange Multiple sama banyaknya dengan data pelatihan, tetapi sebenarnya banyak dari data pelatihan yang Lagrange Multiplier-nya sama dengan nol (Karena hanya beberapa saja yang akan menjadi support vector) ketika menerapkan syarat pertama. Constraint diatas menyatakan bahwa Lagrange Multipioer harus nol, kecuali untuk data latih yang memenuhi persamaan.

( w. + b) = 1 ( 3.19 )

Data pelatihan itu, dengan > 0, terletak pada hyperplane atau dan disebut *support vector.* Data pelatihan yang tidak terletak di hyperplane tersebut memiliki = 0. Persamaan 15 dan 16 juga menyarankan parameter w dan b, yang mendefinisikan hyperplane, hanya tergantung pada support vector.

Walaupun bisa diselesaikan, masalah optimalisasi diatas masih sulit karena banyak parameter : w , b, ,untuk menyerderhanakannya, persamaan optimalisasi 14 harus ditransformasi kedalam fungsi Lagrange Multiplier itu sendiri.

Persamaan lagrange Multiplier 14 dapat dijabarkan menjadi

Lp = - – b + ( 3.20 )

Syarat optimal persamaan 16 ada dalam suku ketiga ruas kanan dalam persamaan 20, dan memaksa suku ini menjadi sama dengan nol. Dengan mengganti w dari syarat 15 dan suku = . , persamaan diatas akan berubah menjadi dualitas Lagrange Multiplier yang berupa Ld, didapatkan :

Ld = - ( 3.21 )

Syarat 1 :

= 0 ( 3.22 )

Syarat 2

≥ 0, i = 1,2,.... N ( 3.23 )

Untuk set data yang besar, masalah dualitas optimalisasi tersebut (21,22,23) dapat diselesaikan dengan metode numerik seperti quadratic programming. Begitu didapatkan, persamaan 15 dan 17 bisa digunakan untuk mendapakan solusi w dan b. Hyperplane didapat dengan formula

+ b = 0 ( 3.24 )

N adalah jumalah data yang menjadi support vector, xi merupakan support vector, z merupakan data uji yang akan diprediksi kelasnya dan xi.z merupakan inner product antara xi dan z. Untuk nilai b didapat dari persamaan 17 pada support vector. Karena dihitung dengan metode nnumerik dan mempunyai error numerik, nilai yang dihitung untuk b bisa jadi tidak sama. Hal ini disebabkan oleh support vector yang digunakan dalam persamaan 17, biasanya diambil nilai rata-rata dari b yang didapat untuk menjadi parameter hyperplane. Untuk mendapatkan b, persamaan 15 dapat disederhanakan menjadi

b = 1 – yi(w.xi) ( 3.25 )

Penjelasa diatas berdasarkan asumsi bahwa kedua kelas kelas dapat terpisah sempurna oleh hyperplane. Akan tetapi pada umumnya kedua kelas tersebut tidak dapat terpisah secara sempurna. Hal ini menyebabkan proses optimalisasi tidak dapat diselesaikan karena tidak ada w dan b yang memuhi persamaan 13. Untuk itu, pertidaksamaan tersebut dimodifikasi dengan memasukan variabel slack. Hasilnya adalah

( w. + b) ≥ 1 - ( 3.26 )

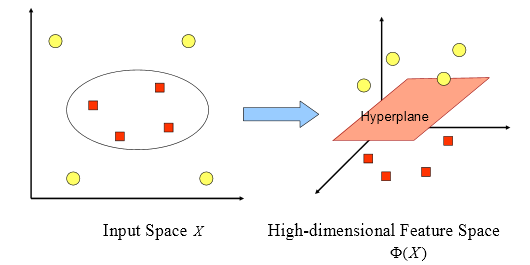
Demikian juga untuk masalah persamaan 12

+ C ( 3.27 )

Parameter C berguna untuk mengontrol pertukaran antara margin dan error klasifikasi. Semakin besar nilai C, semakin besar pula pelanggaran yang dikenakan untuk tiap klasifikasi.

### SVM Nonlinear

Dalam SVM ada linear dan nonlinear (kernel trik). SVM sebenarnya adalah hyperplane linear yang hanya bekerja pada data yang dapat dipisahkan secara linear. Untuk data yang distribusi kelasnya tidak linear digunakan pendekatan kernel pada fitur data awal set data. Kernel dapat didefinisikan sebagai suatu fungsi yang memetakan fitur data dari dimensi awal (rendah) ke fitur lain yang berdimensi lebih tinggi.



**Gambar 3.2 Dimensi Data**

Ilustrasi kernel yang digunakan untuk memetakan dimensi awal set data ke dimensi baru dapat dilihat pada gambar 3.21. Algoritma memetaan kernel seperti berikut

: →

X → (x)

merupakan fungsi kernel yang digunakan untuk pemetaan, D merupakan data training, q merupakan set fitur dalam satu data yang lama, dan r merupakan set fitur yang baru sebagai hasil pemetaan setiap data training. Sementara x merupakan data training, dimana, x1,x2,x3......,xn merupakan fitur-fitur yang dipetakan kefitur berdimensi tinggi r., jadi untuk set data yang digunakan sebagai pelatihan dengan algoritma yang ada dari dimensi fitur yang lama D ke dimensi fitur baru r.

Selanjutnya dilakukan proses pelatihan yang sama sebagaimana pada SVM linear. Proses pemetaan pada fase ini memerlukan perhitungan dot-product dua buah data pada ruang fitur baru. Dot-product kedua buah vektor xi dan xj dinotasikan sebagai (xi) . (xj). Nilai dari dot-product kedua vektor ini dapat dihitung secara tidak langsung, yaitu tanpa mengetahui fungsi transformasi . Teknik komputasi seperti ini kemudian disebut kernel trik, yaitu menghitung dot-product dua vektor diruang dimensi baru dengan memakai komponen kedua buah vektor tersebut diruang dimensi asal, seperti berikut

K (xi.xj) = (xi) . (xj) ( 3.28 )

Dan prediksi pada set data dengan dimensi fitur baru diformulasikan dengan

F( (x)) = sign(w. (z) +b) = sign( ) ( 3.29 )

N adalah jumlah data yang menjadi support vecotr, xi adalah support vector dan z adalah data uji yang akan diprediksi. Untuk pilihan fungsi kernel yang banyak digunakan dalam aplikasi dapat dilihat pada tabel 3.2

Tabel 3. Fungsi Kernel

|  |  |
| --- | --- |
| Nama Kernel | Definisi Fungsi |
| Linear | K (x,y) = x.y |
| Polinomial | K(x,y) = |
| Gaussian RBF | K(x,y) = exp( |
| Sigmoid | K(x,y) = tanh((x.y) + c) |
| Invers Multikuadrik | K(x,y) = |

## TF-IDF

Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (*term*) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu *Term* *frequency* (TF) merupakan frekuensi kemunculan kata (t) pada kalimat (d). *Document* *frequency* (DF) adalah banyaknya kalimat dimana suatu kata (t) muncul. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut.

Bobot kata semakin besar jika sering muncul dalam suatu dokumen dan semakin kecil jika muncul dalam banyak dokumen. Pada Metode ini pembobotan kata dalam sebuah dokumen dilakukan dengan mengalikan nilai TF dan IDF. Pembobotan diperoleh berdasarkan jumlah kemunculan term dalam kalimat (TF) dan jumlah kemunculan term pada seluruh kalimat dalam dokumen ( IDF). Menghitung bobot (W) masing-masing dokumen dengan persamaan di bawah

= x ( 3.30 )

Dimana :

d : kalimat ke-d

t : kata (term)

TF : term frequency

W : bobot

IDF : inverse document frequency

Sedangkan nilai IDF sebuah term dihitung menggunakan persamaan di bawah:

IDF = log ( N / ) ( 3.31 )

Dimana :

N : jumlah kalimat

: jumlah kemunculan kata (term)

# 

# ANALISIS DAN RANCANGAN SISTEM

## Analisis Sistem

*Opinion mining* merupakan sebuah sistem yang dapat digunakan untuk menentukan sentimen atau opini secara otomatis pada sebuah dokumen. Penelitian ini akan mengambil data dari *twitter* yang berbahasa indonesia. Sistem yang dibangun dan dirancang adalah sistem yang memiliki kemampuan untuk menentukan sentimen dari *tweet* hasil *query user* atas *term* pada aplikasi yang terhubung dengan Twitter Search API. Metode yang digunakan untuk menentukan sentimen adalah metode gabungan *unsupersive* yang berbasis pada analisa bahasa dan metode *supervise* yaitu *Support Vector Machine* (SVM).

*Data training* yang digunakan adalah *tweet* yang juga digunakan dalam penelitian paulina. Data tersebut akan dikategorikan ke dalam kelas sentimen positif, negatif dan netral dengan menggunakan analisis bahasa (*unsupervise*). Hasil pengkategorian *data training* digunakan untuk menentukan sentiment pada *tweet* yang baru dengan menggunakan metode *supervise* yaitu *Supoort Vector Machine* (SVM).

### Deskripsi 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 4.1.

Pada proses pelatihan dengan menggunakan masukan berupa corpus yang digunakan pada penelitian paulina 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*.



Gambar . Proses Proses dalam Opinion Mining

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 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 sentimen pada *tweet* yang telah dipreprocessing. Selanjutnya untuk *tweet* yang terindikasi memiliki sentimen, sentimennya akan ditentukan dengan metode SVM.

## Analisis 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 paulina 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.

## Preprocessing

*Preprocessing* sangat menentukan dalam proses pelabelan secara *lexicon* *based*. Dimana pelabelan *lexicon* *based* nantinya akan menentukan proses pelabelan yang dilakukan dengan metode SVM. *Preprocessing* juga digunakan dalam kedua proses utama, baik proses pelatihan maupun proses penentuan label dengan SVM. Jadi *preprocessing* ini cukup penting peranannya dalam penelitian ini.

Tahap *preprocessing* sendiri terdiri atas beberapa proses yang akan dibahas satu per satu secara lebih mendetail. Proses-proses tersebut diantaranya adalah pembersihkan karater yang bukan termasuk UTF-8 (*clear* *invalid* *UTF8*), merubah ke huruf kecil (*casefolding*), menghapus simbol-simbol(*remove* *symbol*), memproses angka-angka(*processing* *numbers*), menghapus perulangan karakter (*remove* *repeat*), melakukan penggantian terhadap kata tidak baku(*replace* *slang*), melakukan spelling correction(*spelling* *correction*).

### Clear Invalid UTF-8

Proses ini adalah bagian dari tahap preprocessing. Pada intinya proses ini akan menghilakan karakter-karakter yang tidak termasuk dalam UTF-8. Secara teknik proses ini dilakukan dengan fitur pada bahasa pemrograman yang sudah disediakan.

### Casefolding

Proses ini hanyalah merupan semua karakter yang ada pada tweet menjadi karakter dalam huruf kecil semua. Hal ini dilakukan karena tidak ada kaitan antara huruf kecil dan huruf besar terhadap sentimen, serta mempermudah proses-proses selanjutnya.

### Remove Symbol

Proses ini adalah proses selanjutnya dari tahap preprocessing. Pada proses ini tanda baca seperti titik(.) , koma(,), seru(!), tanya (?) dan lain sebagainya akan dihilangkan. Tanda/simbol ini dapat dihilangkan karena tidak banyak berpengaruh pada penentuan label. Secara teknik ini akan dilakukan dengan bahasa pemrograman dengan mencari simbol-simbol kemudian menghapusnya.

### Processing Numbers

Proses selanjuatnya dalam tahap preprocessing adalah pemrosesan terhadap angka-angka. Kita tahu dalam sosial media, pengguna sering menggunakan angka-angka baik untuk menyingkat kata, atau untuk bergaya atau untuk menyebutkan nilai. Angka-angka yang terdapat dalam sebuah kata akan di convert menurut tabel 4.1.

Tabel . Daftar Konversi

|  |  |
| --- | --- |
| Angka | Konversi |
| 0 | O |
| 1 | i |
| 2 | Copy karakter-karakter sebelumnya sebelumnya |
| 3 | e |
| 4 | a |
| 5 | s |
| 6 | g |
| 7 | T |
| 8 | b |
| 9 | g |

Tabel 4.2 adalah beberapa contoh kata yang akan diubah oleh proses ini bila ditemukan.

Tabel . Contoh Kopnversi Angka

|  |  |
| --- | --- |
| Masukan | Keluaran |
| ga2l | gagal |
| p3rg1 | pergi |
| Hati2 | Hati-hati |
| Ba9u5 | Bagus |

### Remove Repeat

Proses ini salah satu proses penting dalam tahap prepocessing. Proses ini akan menghapus karakter yang berulang lebih dari dua kali. Seperti diketahui dalam bahasa indonesia perulangan karakter yang berurutan hanya dua karakter, maka jika lebih akan dihapus dan hanya akan disisakan 2 karakter yang sama yang berurut saja. Untuk memperjelas perhatikan tabel 4.3.

Tabel 4.3 Contoh menghapus karakter berualng

|  |  |
| --- | --- |
| Input | Ouput |
| Tidaaaaaaaaaak | Tidaak |
| Cemunguuuuudh | Cemunguudh |
| eaaaaaa | eaa |

### Replace Slang

Proses ini merupakan proses penggantian kata tidak baku menjadi kata baku. Proses ini dilakukan dengan menggunakan bantuan daftar kata-kata informal dan padanannya dalam kata-kata formal. Proses ini akan mengecek kata terdapat dalam daftar kata informal atau tidak. Jika tidak maka kata tidak akan diganti tetapi bila ya maka kata akan diganti dengan padanannya yang terdapat dalam daftar kata baku. Daftar kata baku dan kata tidak baku ini menggunakan daftar yang digunakan pada pelatihan edmond, yang sudah ditambah secara manual berdasarkan pengamatan dalam proses pengembangan sistem ini. Contoh dari proses ini akan ditunjukan pada tabel 4.4.

Tabel . Contoh Pengantian kata tidak baku

|  |  |
| --- | --- |
| Kata tidak baku | Kata baku |
| Nggak | Tidak |
| Lu | Kamu |
| Ngadu | Mengadu |

### Spelling Correction

*Spelling* *Correct* adalah proses terakhir pada tahap *preprocessing*. Proses ini akan memperbaiki kesalahan penulisan yang ada data *tweet*. Seperti kita tahu pengguna sosial media sering sengaja menyingkat kata untuk memenuhi 140 karakter (jumlah maksimal karakter pada twitter) atau secara tidak sengaja salah menuliskan kata. Metode yang digunakan akan dengan memanfaatkan daftar kata baku yang sudah ada dan dengan digabungkan dengan fungsi levenshtein. Fungsi levenshtein adalah fungsi yang dapat menghitung kedekatan/kemiripan antar kata. Secara teknis kata yang telah melalui proses sebelumnya dan tidak terdapat dalam daftar kata baku, maka kata ini akan dikenai proses ini. Tabel 4.5 adalah contoh kata-kata yang dapat ditangani oleh proses ini.

Tabel . Contoh SpellCorrection

|  |  |
| --- | --- |
| Input | Ouput |
| Bapk | Bapak |
| Menghitumg | menghitung |
| Yng | yang |

Walaupun proses ini memberikan manfaat yang cukup besar. Namun terkadang proses ini memberikan kelemahan. Hal ini terjadi ketika proses ini diberikan input suatu frase atau kata serapan baru yang belum terdapat dalam dafatar kata baku. Misalkan BLBI oleh karena BLBI tidak terdapat dalam daftar kata baku maka BLBI akan digantikan dengan kata yang terdekat dengan kata BLBI. Padahal yang dimaksud dengan kata ini adalah kata serapan baru.

## Penentuan Label Data Training

Tahap selanjutnya setelah preprocessing data adalah penentuan label dengan metode *lexicon* *based*. Tahap ini masih dalam proses utama yang pertama yaitu proses pelatihan. Pada tahap penentuan label ini terdapat beberapa proses di dalamnya. Seperti ditunjukan pada gambar, proses-proses tersebut diantaranya *POS taging*, *Opinion* *Rules* dan *check* *negation*.

### POS taging

Setelah tahap *preprocessing* selesai berarti data training sudah bersih dan telah siap untuk dilakukan proses penentuan label secara *lexicon* *based*. Proses pertama yang dilakukan pada tahap ini adalah melakukan *POS taging* pada data tersebut.

POS taging adalah proses untuk menentukan kelas kata dari setiap kata-kata yang terdapat dalam kalimat. Proses *POS taging* dalam penelitian ini menggunakan metode *dictionary*. Misalnya terdapat satu *tweet* “Kasus penembakan yang dilakukan oleh polisi sangat rumit” maka dengan POS taging akan didapatkan “Kasus/NN penembakan/VB yang/CK dilakukan/VB oleh/IN polisi/NN sangat/RB rumit/JJ”.

Langkah pertama adaalah melakukan tokenisasi, yaitu pross yang memecah-mecah *tweet* menjadi satuan kata. Dimana sistem akan melihat setiap spasi yang ada dalam *tweet* dan berdasarkan spasi tersebut maka kata-kata dapat dipecah.

Langkah kedua yang dilakukan adalah melakukan penyesuaian dengan koleksi kata yang terdapat dalam basisdata. Setiap kata dari hasil tokenisasi tersemua kemudian dicek satu-persatu kelas katanya sesuai dengan koleksi kata dalam basisdata. Jika kata tersebut terdapat dalam basisdata maka akan langsung ditentukan kelas katanya. Namun apabila kata tersebut tidak ditemukan dalam database, maka kata tersebut akan dianggap sebagai kata benda(NN).

Apabila dilihat dari cara kerja *POS taging* ini, ada kata-kata yang tidak bisa dideteksi dengan benar, karena karakteristiknya tidak bisa menangani kata-kata yang memiliki konteks dan maknna ganda. Misalkan pada tweet “Mobil polisi menarik mobil sedang yang kecelakaan.” Kata menarik dalam kali ini akan dideteksi menjadi kata sifat padahal bila dilihat dari konteks nya kata ‘menarik’ disini adalah kata kerja.

### Opinion Rules

Seperti yang telah dijelaskan sebelumnya, dalam penentuan label data training akan menggunakan metode *lexicon* *based*, dimana penentuan labelnya menggunakan rule-rule tertentu. Rule-rule ini mengacu pada penelitian yang dilakukan oleh Edmond.

Secara teknis langkah yang dilakukan adalah dengan melihat sekuens dari kata-kata hasil *POS taging* sebelumnya yang sudah menyertakan kelas kata dari masing-masing kata. Rule yang dipakai untuk mendeteksi frasa yang terdiri dari dua kata saja dan kata-kata yang berdiri sendiri. Pada tabel 4.6 ditampilkan daftar rule frasa opini yang digunakan dalam sistem ini.

Tabel 4.6 Opinion Rules

|  |  |
| --- | --- |
| NO | Rule |
| 1 | RB JJ |
| 2 | RB VB |
| 3 | CK VB |
| 4 | NN JJ |
| 5 | NN VB |
| 6 | VB JJ |
| 7 | JJ VB |

Dalam penentuan frase ini terlebih dahulu mendeksi kehadiran dari kata sifat atau kata kerja. Setelah sistem dapat mendeteksi kehadiran dari kata kerja atau kata sifat dalam kalimat, maka sistem akan melihat kata sebelum kata kerja atau kata sifat dan melihat kelas katanya. Sehingga apabila kedua kelas kata dari kedua kata tesebut diurutkan makan akan dicocokan dengan rule yang ada. Jika sama dengan rule yang ada maka frase tersebut bisa dikatakan memiliki opini atau sentimen.

Misal terdapat suatu kalimat “Film itu sangat bagus”, oleh *POS taging* akan ditandai sebagai “Film/NN itu/DT sangat/RB bagus/JJ”. Dari hasil dapat dilihat ada frasa yang sesua dengan rule yaitu ‘sangat bagus’ (RB JJ) karena kata ‘bagus’ memiliki sentimen positif maka kalimat tersebut diberi label positif.

### Check Negation

Proses terakhir pada tahap penentuan rule ini adalah mengecek negasinya. Setalah *tweet* dilakukan *POS taging* dan dicocokan dengan rule-rule yang ada. Apabila terdapat sentimen dari rule-rule dari maka selanjutnya akan dicek apakah terdapat negasi dari hasil rule-rule tadi.

Konsepnya mencari kata negasi sebelum kata sifat atau kata kerja. Dimana daftar kata negasi diambil dari basisdata. Toleransi jarak yang digunakan adalah 5 kata dari kata sifat atau kata benda. Misalkan terdapat tweet “Film itu tidak bagus”. Pertama akan dilakukan POS taging sehingga menghasilkan seperti berikut “Film/NN itu/DT tidak/MD bagus/JJ”. Kemudian akan dilakukan opinion rule yang akan menghasilkan frase positif yang dihasilkan dari kata ‘bagus’. Proses *check* *negation* akan melakukan pengecekan kata-kata sebelum frase tersebut. Jika terdapat kata negasi maka labelnya akan dibalik. Jadi penentuan dari *tweet* tersebut akan menghasilkan label negatif.

## Pelatihan

Pelatihan digunakan untuk membuat model dari data training. Masukan dari proses ini adalah berupa data training yang sudah dilakukan preprocessing dan diberi label dengan menggunakan *lexicon* *based*. Model yang dihasilkan dari proses ini kemudian digunakan untuk proses penentuan label. Pelatihan dilakukan dengan metode SVM dan data training berupa *tweet* diubah kedalam *vector* *space* *model*. Algoritma vector space yang digunakan adalah menggunakan TF-IDF (term frequency inverse document frequency). Tahap pelatihan ini terdiri atas beberapa proses yang diperlihatkan pada gambar 4.2.



Gambar .2 Proses-proses dalam Tahap Pelatihan

### Generate Weight

SVM adalah salah satu metode yang banyak digunakan untuk klasifikasi, tetapi SVM hanya peduli tentang titik dalam ruang, bukan *tweet* atau dokumen. Untuk tujuan ini *vector* *space* *model* digunakan untuk mereprestasikan setiap kata dalam dokumen ke dalam *vector*. *Vector* *space* *model* yang digunakan dalam penelitian ini adalah TF-IDF.

Masalah utama yang dihadapi adalah bagaimana kita menentukan bobot untuk suatu kata dalam *tweet*. Secara sederhana kita dapat menghitung kemunculan suatu kata dalam suatu tweet. Misal kata ‘baik’ yang muncul dalam sebuah tweet sebanyak 2 kali maka kata tweet memiliki bobot 2. Hal ini dikenal dengan istital *term* *frequency* (tf) atau disebut berat lokal. Ini dihitung pada sebuh *tweet* tanpa mengacu pada *tweet* yang lain.

Sedangkan mitra dari berat lokal ini adalah berat global. Yaitu bobot suatu kata pada seluruh *tweet* (dokumen) yang ada. Ada juga yang mengistilahkan dengan berat umum atau *inverse* *document* *frequency*, yang didefinisikan sebagai jumlah dari semua *tweet*(dokumen) dibagi dengan jumlah *tweet*(dokumen) yang mengandung istilah tertentu. Sebagai contoh terdapat 100 dokumen, 9 diantaranya mengandung kata ‘baik’ maka berat global (idf) kata baik adalah 11.1.

Pembobotan ini digabung dengan istilah tf-idf, secara sederhana dengan mengalikannya. Hasil dari pembobotan ini akan disimpan dalam sebuah text dokumen yang nantinya akan digunakan sebagai file input proses pelatihan.

### Traning SVM

Proses pelatihan pada penelitian ini akan dilakukan dengan *library* libsvm. Hasil dari proses pelatihan ini adalah SVM Model yang nantinya akan digunakan dalam proses penentuan label.

## Penentuan Label dengan SVM

Setelah SVM Model berhasil dibuat, tahapan terakhir dalam sistem ini adalah penentuan label dengan menggunakan SVM. Dalam tahap ini terdapat beberapa proses diantaranya, pencarian *tweet*, *preprocessing*, pengecekan frase, generate bobot data testing, penentuan label seperti ditunjukan pada gambar 4.3.



Gambar .3 Proses-proses dalam Tahapan Penentuan Label

### Pencarian Tweets

Pada proses ini sistem akan mencari *tweet* yang terkait dengan kata kunci yang dimasukan oleh pengguna. Proses ini dilakukan dengan bantuan API yang telah disediakan oleh twitter. Berikut adalah format API yang disedikan twitter.

[http://search.twitter.com/search.atom?lang=[bahasa]&rpp=[jumlah]&q=[kata\_kunci](http://search.twitter.com/search.atom?lang=%5bbahasa%5d&rpp=%5bjumlah%5d&q=%5bkata_kunci)]

*Link* tersebut akan mengembalikan kata XML yang berisi dengan *tweet* yang terkait dengan kata kunci. Selanjutnya sistem akan merubah dan menyimpan data XML tersebut ke dalam database.

### Cek Frase

Sebelum Cek Frase sebenarnya dilakukan preprocessing untuk data yang baru saja didapat dari API twitter. Tetapi proses tersebut tidak dibahas disini karena telah dibahas pada subbab sebelumnya. Setelah dipreprocessing data tersebut juga dilakukan *POS taging* yang juga tidak dibahas disini karena prosesnya sama dengan proses *POS taging* pada tahap sebelumnya.

Langkah berikutnya setelah dipreprocessing dan *POS taging* adalah pengecekan apakah terdapat frase yang mengandung opini atau tidak. Jika tidak mengandung opini maka tweet tidak akan ikut diproses pada tahap penentuan sentimen tetapi langsung diberikan label netral. Jika terdapat frase opini maka akan dilakukan proses penentuan label dengan SVM. Proses pengecekan frase dilakukan dengan mencari keberadaan kata sifat atau kata kerja pada tweet.

### Generate bobot

Proses selanjutnya adalah mengenerate bobot untuk *tweet* yang mengadung frase opini. Metode pembobotan yang dilakukan akan dengan menggunakan TF-IDF yang sudah dijelaskan sebelumnya. Proses pembobotan ini akan menghasilkan sebuah berkas yang disebut dengan berkas testing yang berisi bobot dari *tweet* yang memiliki frase opini tadi. Berkas ini selanjutnya akan menjadi masukan bagi proses penentuan label dengan SVM.

### Penentuan label dengan SVM

Pentuan label dengan SVM dilakukan dengan libsvm. Proses ini menggunakan dua buah masukan yaitu SVM Model dan berkas testing. Proses ini akan menghasilkan berkas output yang merupakan hasil pelabelan dari berkas testing.

### Pemetaan dalam sentimen

Seperti diketahui SVM hanya memperdulikan titik pada vector. Output yang dihasilkan juga hanya 1 dan -1. Untuk itu proses ini kan memetakan 1 sebagai sentimen positif dan -1 sebagai sentimen negatif yang kemudian hasil pementaan ini akan disimpan dalam database yang nantinya dapat dianalisa untuk proses pengujian dan perhitungan akurasi sistem.

## Arsitektur Sistem

Arsitektur sistem *Opinion* *Mining* yang dibuat ditunjukan pada gambar 4.4. 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 pelatihan maka akan dilakukan penentuan label pada untuk data *tweet* yang sudah dipreprocessing dengan menggunakan metode *lexicon* *based*. Selanjutnya dilakukan proses pelatihan pada data *tweet* yang sudah diberi label. Pelatihan ini akan menghasilkan SVM model.

Alur proses besar berikutnya adalah proses penentuan sentimen. 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* kurang lebih 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. Jika tidak maka akan diberi label netral. Selanjutnya adalah klasifikasi dengan metode SVM untuk data *tweet* yang telah teridentifikasi memiliki frase opini. Klasifikasi ini melibatkan SVM model yang sudah dibuat pada proses pelatihan sebelumnya. Kemudian setelah selesai klasifikasi, hasilkan akan disimpan dalam basisdata.



Gambar 4 Arsitektur sistem

## Rancangan Sistem

Rancangan sistem akan memberikan penjelasan tentang rancangan untuk proses pelatihan dan pemberian label. Rancangan kedua proses tersebut akan digambarkan menggunakan class diagram dan sequence diagram.

### Class Diagram

Pada penelitian ini terdapat beberapa *class* yang digunakan. *Class* *diagram* dari semua *class* diperlihatkan pada gambar 4.5.

*Class*-*class* pada gambar 4.5 terbagi menjadi beberapa jenis *class*. *Class* *model*, *class* *controller*, *class* *component* dan *class* *view*. Setiap jenis *class* memiliki peranan masing-masing. *Class* *model* berperan melakukan interaksi terhadap basisdata. *Class* *controller* berperan sebagai pengatur *class*-*class* yang lain. *Class* *component* berisi implementasi dari algoritma-algoritma yang dibutuhkan sistem. *Class* *view* berperan untuk menampilkan presentasi kepada pengguna.

*Class*-*class* yang memiliki peran fundamental dalam penelitian ini adalah pada *class* kategori *component*. *Class* MyTwitterComponent yang berfungsi untuk mencari *tweet* pada twitter. *Class* PreprocessingComponent yang memiliki fungsi untuk melakukan *preprocessing*. *Class* JanPOS tagingComponent memiliki peran untuk melakukan *POS taging*. *Class* SpellingCorrectionComponent yang berfungsi untuk melakukan *spelling* *correction* untuk kata-kata yang tidak terdapat dalam corpus. *Class* WeightComponent adalah *class* yang berfungsi untuk menghitung bobot *tweet* agar dalam digunakan dalam SVM. *Class* JanSvmComponent berfungsi algoritma SVM yang didalamnya memanfaatkan library libsvm. *Class* SentimentAnalysisLexiconBasedComponent berperan untuk melakukan analisis sentiment dengan menggunakan metode *lexicon* *based*.



**Gambar 4.5 Class diagram**

### Sequence diagram proses pencarian data twitter



**Gambar 4.6 Sequence diagram proses pencarian data twitter**

Gambar 4.6 menunjukan *sequence* *diagram* proses pencarian *tweet*. User memasukan kata kunci pencarian melalui *class* HuntsView yang kemudian diteruskan pada *class* HuntsController. *Class* HuntsController akan memanggil *method* getAllTweets pada class MyTwitter. Setelah mendapatkan hasilnya *class* HuntsController akan menyimpan hasilnya pada basisdata dengan cara mendelegasikannya pada *Class* Tweet.

### Sequence diagram preprocessing



**Gambar 4.7 Sequence diagram preprocessing**

Gambar 4.7 menunjukan *sequence* *diagram* *preprocessing* data. Data yang sudah disimpan dalam basisdata akan diambil oleh *class* RepositoriesController dengan mengakses *method* find yang ada pada *class* Repository selanjutnya untuk melakukan *preprocessing,* *class* RepositoriesController akan memanggil method doIt pada *class* Preprocessing. Setelah selesai dilakukan *preprocessing* data akan kembali disimpan dalam basisdata malalui *class* CleanTweet.

### Sequence diagram proses pemberian label dengan metode lexicon



**Gambar 4.8 Sequence diagram proses pemberian label dengan metode lexicon**

Gambar 4.8 menunjukan proses pemberian label dengan metode *lexicon based*. Pertama user akan meminta melakukan analisis dengan metode *lexicon* *based* melalui *method* analisis pada *class* CleanRepositoriesView yang akan diteruskan pada *Class* CleanRepositoriesController. *Class* CleanRepositoriesController akan mengambil data yang sudah dilakukan *preprocessing* dengan bantuan *class* CleanRepository. Setelah mendapatkan data yang dimaksud, *class* CleanRepositoriesController akan melakukan *POS taging* pada data tersebut. Proses *POS taging* dilakukan dengan bantuan *class* JanPOS taging dengan cara mengakses *method* posTagDic. Setelah dilakukan *POS taging* maka proses analisis akan dimulai. Proses-proses analisisnya secara berturut turut adalah preliminaryAnalysis, checkNegation dan conclusion yang semuanya dilakukan oleh *class* SentimentAnalysisLexiconBased. Setelah analisis selesai dilakukan data tersebut kembali disimpan kedalam basisdata melalui *class* CleanRepository.



**Gambar 4.9 Sequence diagram proses pemberian label dengan SVM**

### Sequence diagram proses pemberian label dengan SVM

Gambar 4.9 adalah *sequence* *diagram* dari proses pemberian label dengan motode SVM. Langkah pertama adalah pengambilan data *tweet* yang sudah di *preprocessing* dari basisdata. Proses ini dilakukan oleh CleanTweetsController dengan bantuan *class* CleanTweet. Kemudian akan dilakukan pengecekan apakah *tweet* yang sudah di preprocessing netral atau tidak. Proses ini dilakukan oleh *class* CleanTweetController dengan cara mengakses *method* checkSentiment pada *class* SentimentAnalysisLexiconBased. Kemudian untuk data yang tidak netral akan dilakukan pembobotan yang dilakukan oleh *class* Weight. Selanjutnya Proses pemberian label dengan metode SVM dilakukan oleh *class* JanSvm yang didalamnya memanfaatkan *library* libsvm.

# 

# IMPLEMENTASI

BAB V digunakan untuk menjelaskan implementasi dari perancangan yang telah dijelaskan pada Bab IV. Implementasi yang akan dibahas adalah proses pelabelan dengan metode lexicon based dan pelabelan dengan SVM serta beberapa proses yang ada didalamnya seperti pembobotan, *preprocessing* dan *POS* *taging*. Bab ini juga digunakan untuk menjelaskan kendala-kendala yang ditemui baik yang sudah dapat diatasi dengan baik maupun yang belum. Implementasi pada bab ini menggunakan Apache *server*, bahasa pemrograman PHP Cakephp *Framework* dan basisdata MySQL serta libsvm.

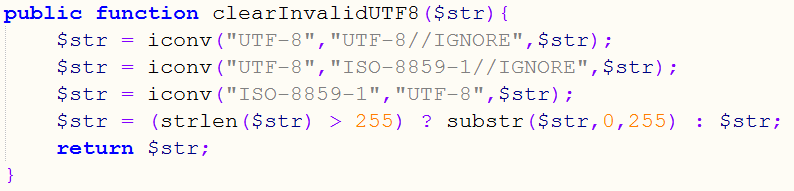
## Preprocessing Data

*Preprocessing* adalah proses yang digunakan oleh kedua proses buat, baik pelabelan dengan metode *lexicon* *based* maupun pelabelan dengan SVM menggunakan metode *preprocessing* data yang sama. *Preprocessing* sendiri berperan penting dalam penentuan sentimen karena *preprocessing* akan menghasilkan data yang sudah bersih dan siap dilakukan analisis.

*Preprocessing* terdiri dari beberapa subproses seperti cleanInvalidUTF8, casefolding, removeSymbol, processNumbers, removeRepeat, *replace* *informal* *word*, *spelling* *correction* yang masing-masing telah dijelaskan pada IV.

### clearInvalidUTF8

Proses clearInvalid UTF8 adalah proses pembersihan karakter yang tidak termasuk dalam UTF8. Dalam sosial media akan banyak ditemu karakter-karakter aneh yang bukan termasuk UTF8 dan karakter tersebut tidak akan dapat tersimpan dengan baik pada MySQL dan tidak banyak mempengaruhi sentimen maka karakter-karakter tersebut akan dihapus. Implementasi dari proses penghapusan tersebut berada dalam class PreprocessingComponent yang ditunjukan pada Gambar 5.1

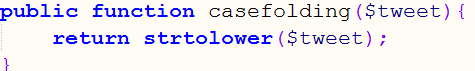


Gambar 5.1 Kode method clearInvalidUTF8

Pada *method* clearInvalidUTF8 digunakan untuk menghapus karakter yang tidak termasuk dalam UTF8. Untuk melakukan hal ini menggunakan fungsi iconv yang disudah disediakan PHP. Iconv merupakan *method* yang disediakan PHP untuk mengkonfersi string kedalam encoding yang ditentukan.

### Casefolding

Casefolding adalah proses untuk merubah karakter kedalam huruf kecil. Karena dalam penentuan sentimen baik huruf kecil maupun besar tidak akan mempengaruhi sentimen. Tetapi dengan mengkonfersi kedalam huruf kecil akan mudah dalam melakukan pencarian pada *corpus*.

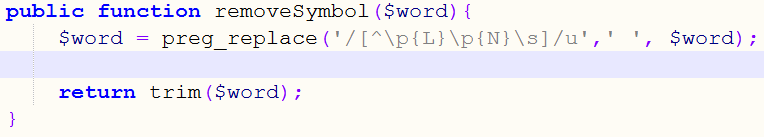


Gambar 5.2 Kode method casefolding

Gambar 5.2 merupakan *method* casefolding yang digunakan dalam penelitian ini. *Method* ini terdapat dalam *class* PreprocessingComponent. Karena dalam PHP *method*-*method* untuk pemrosesan *string* sudah cukup lengkap, proses ini dapat dilakukan dengan mudah hanya dengan memanggil *method* strtolower yang sudah disediakan oleh bahasa PHP.

### removeSymbol

RemoveSymbol adalah sebuah proses untuk menghapus symbol-symbol yang tidak terlalu berpengaruh pada penentuan sentimen. Misalkan simbol #, $,^,%,,\* (,). Gambar 5.3 merupakan kode yang digunakan untuk menghapus symbol.



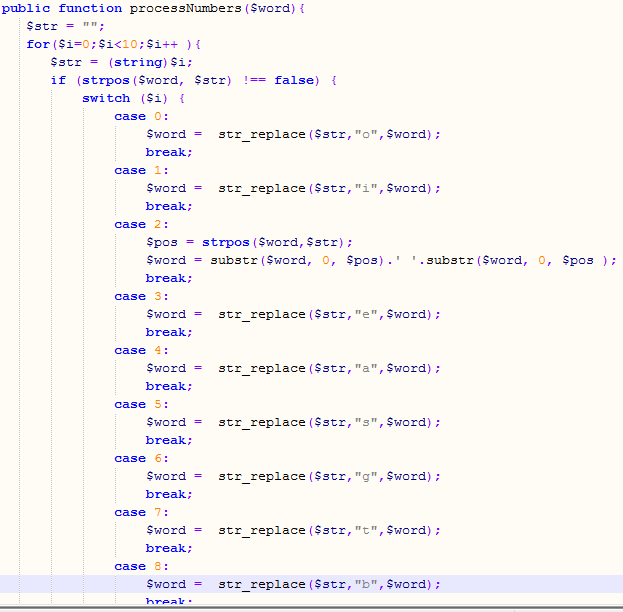
Gambar 5.3 Kode method removeSymbol

*Method* ini juga berada pada *class* PreprocessingComponent. Untuk menghapus symbol-symbol menggunakan *method* preg\_replace atau *method* yang berfungsi untuk melakukan penggantian symbol dengan parameter *regular* *expression*.

### ProcessNumbers

ProcessNumbers adalah proses pengkonfersian angka menjadi huruf. Dalam sosial media atau dalam dunia maya sering digunakan baik untuk mempersingkat kata atau hanya sekedar iseng saja. Untuk melakukan pengkonfersian sesuai yang sudah dijelaskan pada bab IV akan dilakukan oleh *method* processNumbers yang berada pada *class* PreprocessingComponent.

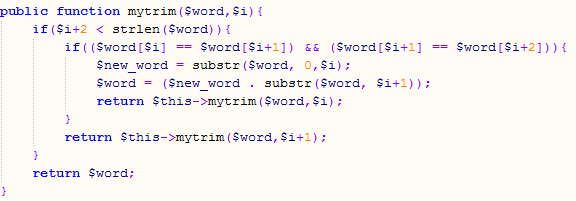
Gambar 5.4 merupakan kode dari *method* processNumbers yang intinya akan mengkonfersi angka menjadi karakter seperti dijelaskan pada bab IV. Caranya adalah dengan mengecek satu persatu, bila ditemukan angka dalam karakter maka akan dikonfersi dengan *method* str\_replace.



Gambar 5.4 Kode method processNumbers

### RemoveRepeat

RemoveRepeat adalah proses untuk menghapus perulangan karakter yang berlebihan. Pada sosial media twitter telah umum digunakan perulangan karakter yang berlebihan. Baik yang tidak sengaja dilakukan karena kesalahan penulisan atau yang sengaja untuk menyangatkan makna. Contohnya adalah misal ingin mengungkapkan setuju terkadang ada yang berlebihan dengan menuliskan “setujuuuuuuuuuuu”. Hal tersebut bila tidak diatasi akan mempengarui penentuan sentimen. Dalam bahasa indonesia perulangan karakter yang berturut-turut hanya dua karakter. Untuk itu proses ini akan melakukan penghapusan karakter berulang berturut-turut yang berlebihan dan hanya disisakan menjadi dua karater yang sama secara berturut-turut.

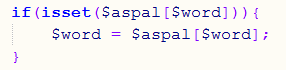


Gambar 5.5 kode dari proses RemoveRepeat

Gambar 5.5 merupakan kode dari proses RemoveRepeat. *Method* ini merupakan *method* *recursive* yang terdapat dalam *class* PreprocessingComponent. Algoritmanya adalah mengecek apakah karakter setelahnya dan setelahnya lagi adalah karakter yang sama dengan karakter sekarang. Jika benar maka akan dihapus satu karakter yang sama. Hal itu dilakukan sampai karakter terakhir pada suatu kata.

### *Replace* *informal* *word*

*Replace* *informal* *word* adalah proses penggantian kata yang ada dalam kamus kata tidak baku diganti dengan kata yang sesuai yang ada pada kamus kata baku. Gambar 5.6 adalah kode proses penggantian kata ini.

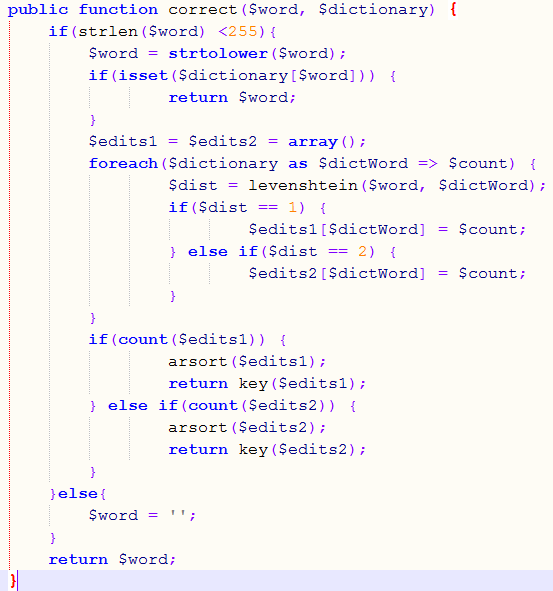


Gambar 5.6 Kode proses penggantian kata tidak baku

Tidak terdapat algoritma yang rumit dalam algoritma ini hanya mengecek apakah kata terdapat dalam kata tidak baku. Jika ada maka akan langsung diganti dengan kata yang baku.

### *Spelling* *Corrention*

*Spelling* *Correction* adalah proses yang apabila bertemu dengan kata baru yang tidak terdapat dalam kamus maka proses ini akan menggantinya dengan kata yang paling dekat dengan kata yang ada dikamus. Proses ini juga dapat dimanfaatkan untuk mengkoreksi dari kesalahan penulisan walaupun tidak memiliki akurasi 100%. Gambar 5.7 adalah kode dari proses *spelling* *correction* yang terdapat dalam *class* SpellingCorrection.



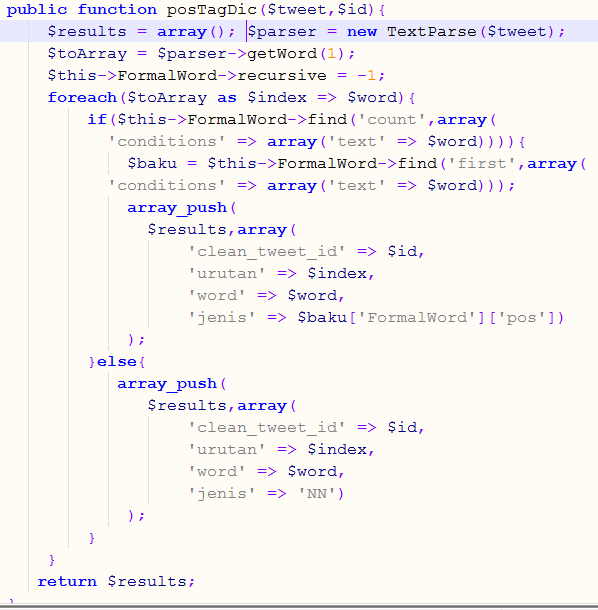
Gambar 5.7 Kode *method* correct pada class SpellingCorrection

Proses *spelling* *correction* tersebut diimplementasikan menggunakan algoritma levenshtein untuk menghitung kedekatan kata baru dengan kata-kata yang ada dalam kamus. Bahasa PHP sudah mengimplementasikan algoritma levenshtein ini hanya dengan menjalankan *method* levenshtein yang sudah disediakan.

## POS taging

POS taging adalah proses untuk pelabelan kata berdasarkan kelas katanya. Proses ini meminta input sebuah kalimat atau dalam penelitian ini adalah sebuah *tweet* dan memberikan *output* sebuah kalimat yang sudah diberi label kelas katanya. Metode POS taging yang digunakan dalam penelitian ini adalah POS taging dengan menggunakan *dictionary* atau kamus.

Proses POS taging ini akan berfungsi untuk menentukan sentimen pada metode *lexicon* *based* dan akan berfungsi untuk mendeteksi kalimat bersentimen atau tidak pada metode SVM. Implementasi prosess ini dilakukan pada *class* JanPOS tagingComponent. Gambar 5.8 menunjukan kode proses POS taging.



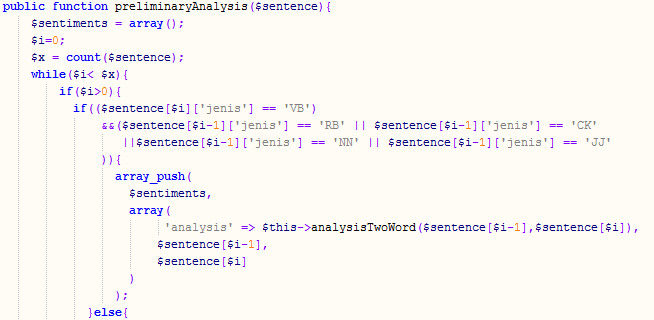
Gambar 5.8 Kode proses POS taging

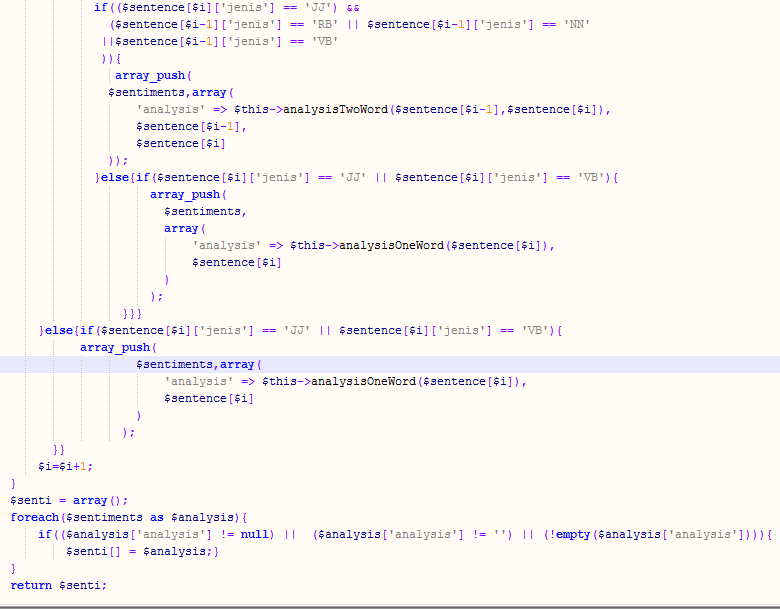
Langkah pertama dari algoritma dari POS taging menggunakan *dictionary*, adalah memecah kalimat atau *tweet* menjadi kata per kata. Kemudian mencocokan masing-masing kata tersebut ke dalam *dictionary* yang sudah terdapat kelas katanya.

## Pelabelan menggunakan metode *lexicon* *based*

Setelah data berhasil dipreprocessing artinya dapat sudah siap untuk dilakukan analisa. Pada tahap pelabelan data pelatihan, metode yang digunakan untuk pemberian label adalah metode *lexicon* *based*. Proses pelabelan dengan metode *lexicon* *based* ini dilakukan oleh *class* SentimentAnalysisLexiconBased. Pada penelitian ini analisis yang dilakukan hanya memperhatikan sampai dua kata yang berdampingan saja.

Proses analisis untuk melakukan pelabelan ini dibagi menjadi beberapa tahap. Tahap pertama adalah pencocokan kelas kata dengan *rule* yang sudah ada. Pada bab sebelumnya telah dijelaskan bahwa sentimen suatu kalimat ditentukan oleh adanya kata benda (NN) atau kata sifat(JJ). Maka tahap pertama akan mencari kata sifat atau kata benda yang ada. Gambar 5.9 adalah kode untuk mencocokan antara kelas dengan frase yang dapat membentuk opini.

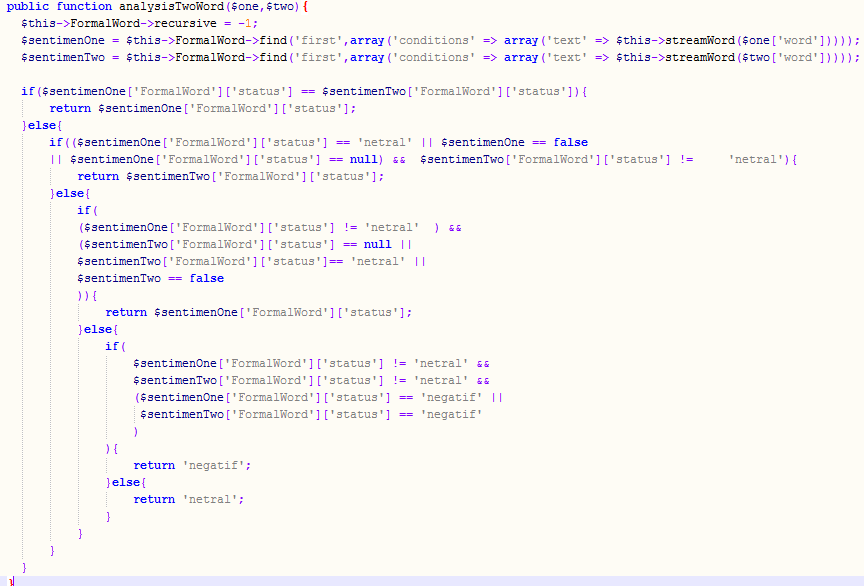




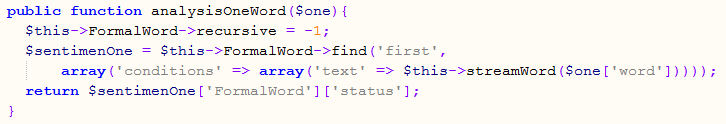
Gambar 5.9 Kode pencocokan kelas kata dan frase

Pada gambar 5.9 *method* prelimeneryAnalysis akan mencari kata benda atau kata sifat. Bila terdapat kata benda atau kata sifat akan dicek kata selanjutnya apakah ada frase yang bersesuai dengan frase yang membentuk opini atau tidak. Jika terdapat frase yang membentuk opini akan dilakukan analisis dua kata. Gambar 5.10 adalah kode *method* analysisTwoWords. Namun apabila tidak ditemukan frase yang bersesuaian tetapi ditemukan kata benda atau kata sifat maka akan dilakukan analisis satu kata. Gambar 5.11 adalah kode *method* analysisOneWord yaitu *method* yang digunakan untuk menganalisis satu kata. Apabila dalam kalimat atau *tweet* tidak terdapat kata benda ataupun kata sifat maka kalimat atau *tweet* tersebut akan langsung diberi label netral.

Setelah ditentukan setimen awal, tahap berikutnya adalah memeriksa apakah sebelum frase yang menunjukan opini terdapat kata negasi atau tidak.



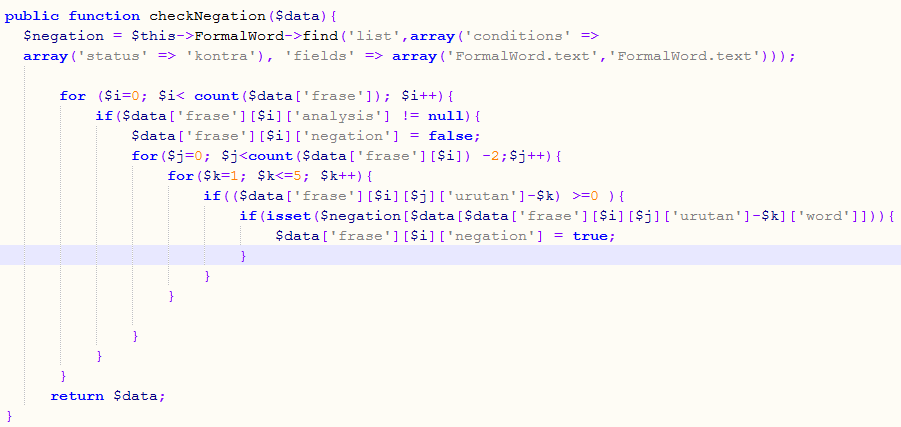
Gambar 5.10 kode method analysisTwoWord



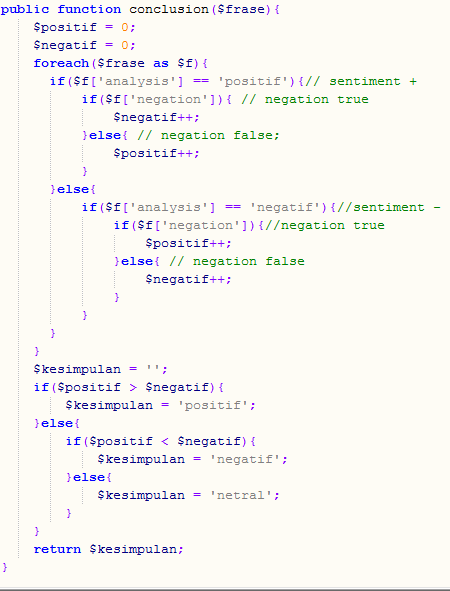
Gambar 5.11 kode method analysisOneWord

Bila terdapat kata negasi sebelum frase maka sentimen dari kalimat atau *tweet* akan diubah. Jika positif akan menjadi negatif dan bila negatif akan menjadi positif. Jarak kata negasi dengan frase diberi toleransi sebanyak lima kata sebelumnya. Gambar 5.12 adalah kode *method* yang digunakan untuk memeriksa negasi.

Setelah tahap pemeriksaan negasi, tahap berikutnya adalah penarikan kesimpulan. Beberapa hal yang perlu diperhatikan adalah sentimen awal, negasi dan menghitung jumlah frase opini yang ada. Kesimpulan diambil berdasarkan perbandingan jumlah opini yang ada. Bila jumlah frase positif lebih banyak maka kalimat akan diberi label positif, apabila frase negatif lebih banyak maka kalimat akan diberi label negatif. Gambar 5.13 menunjukan kode penarikan kesimpulan yang dilakukan.



Gambar 5.12 kode *method* chcekNegation

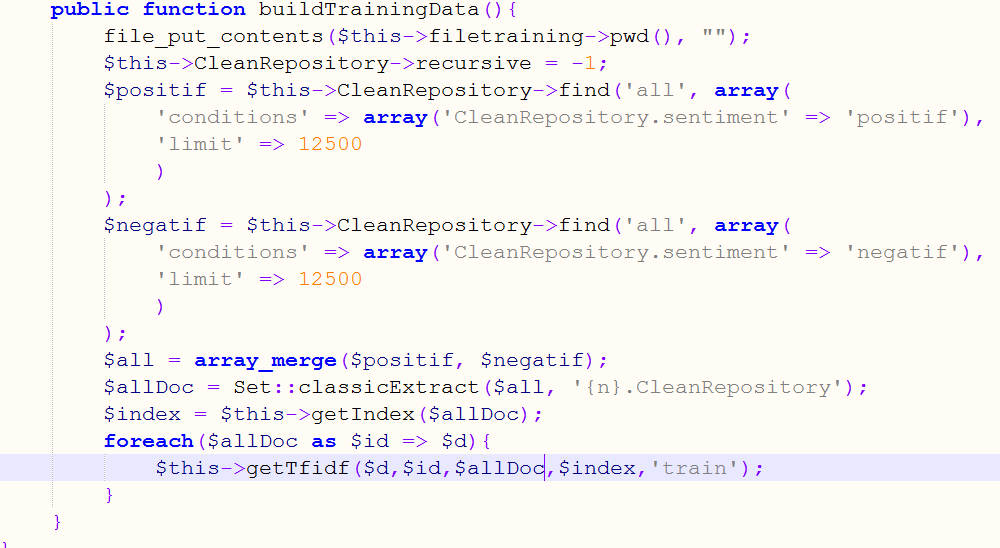


Gambar 5.13 Kode *method* conclusion

## Pembobotan data

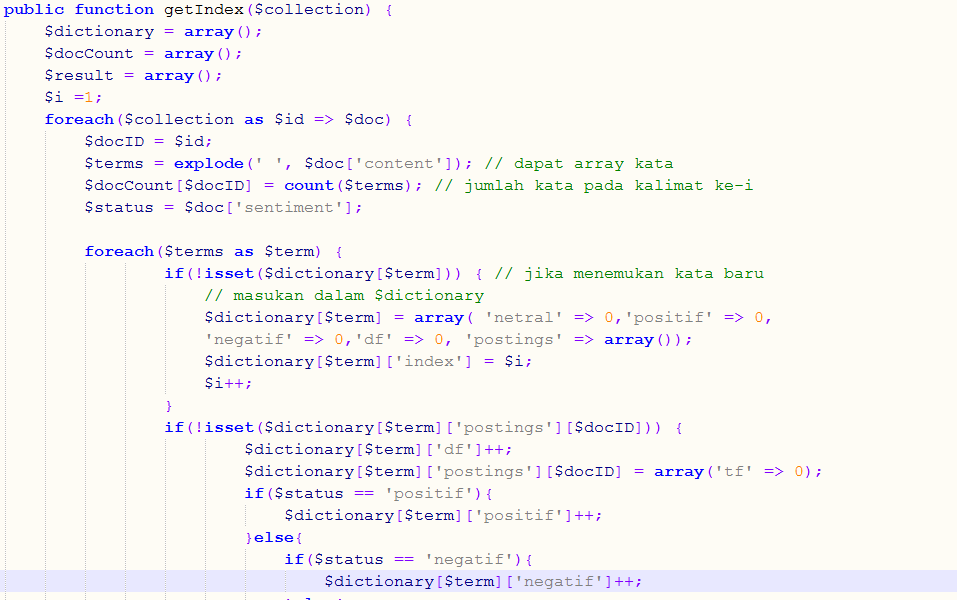
Setelah data berhasil diberi label dengan metode *lexicon* *based*. Data tersebut akan menjadi data training bagi proses pelabelan dengan SVM. Data ini masih berbentuk *tweet* yang tersimpan dalam basisdata sedang SVM hanya mengenali input dan output nya dalam bentuk vector. Untuk itu perlu adanya metode untuk merepresentasikan data ini dalam bentuk vector. Metode tersebut disebut dengan *vector* *space* *model*. Ada beberapa metode *vector* *space* *model*, namun penelitian ini akan menggunakan metode TF-IDF untuk *vector* *space* *model*.

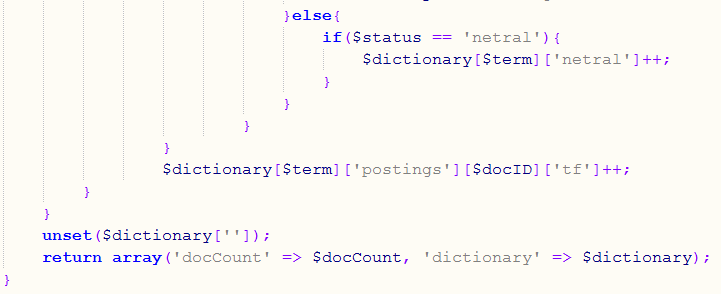
TF-IDF adalah pembobotan yang mengkombinasikan antara berat kata local dan berat global seperti yang sudah dijelaskan pada bab sebelumnya. Metode pembobotan selain digunakan pada proses pembobotan data pelatihan juga digunakan pada proses pembobotan data pengujian. Implementasi dari metode TF-IDF ini dilakukan pada *class* WeightComponent. Gambar 5.14 merupakan kode untuk membangun data pelatihan.



Gambar 5.14 Kode untuk membangun data training

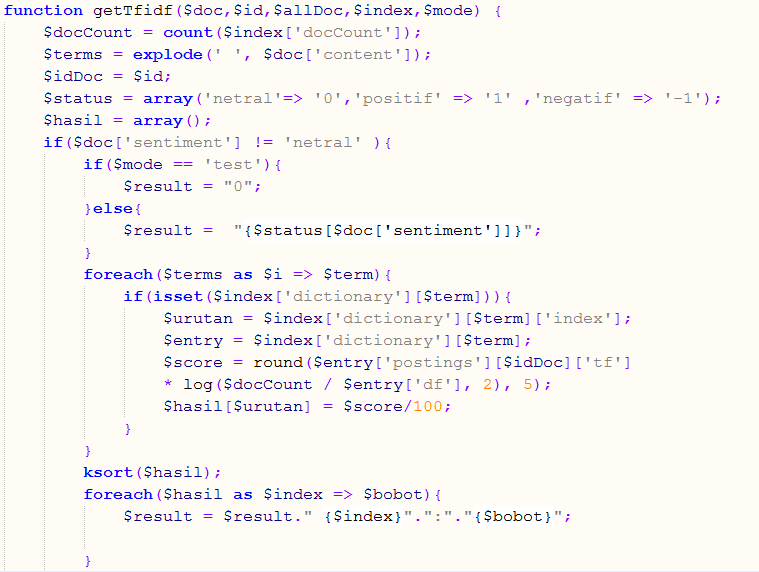
Dalam *method* buildTrainingData seperti pada gambar 5.14. Program akan mengambil data pelatihan sebanyak 25000 *tweet* dari basisdata, yang terdiri dari 12500 data berlabel positif dan 12500 berlabel negatif. Kemudian masing-masing data tersebut akan di-*index*, yaitu menghitung kemunculannya secara global dan lokal. Gambar 5.15 adalah *method* yang digunakan untuk melakukan *indexing*.

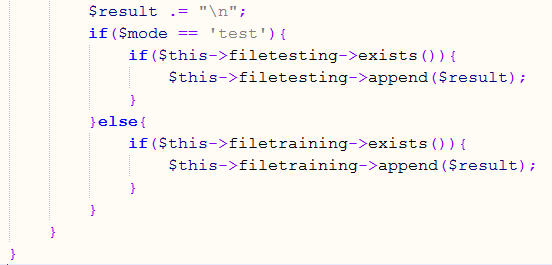




Gambar 5.15 Kode *method* getIndex

Kode pada gambar 5.15 akan menghitung frekuensi kemunculan setiap kata, baik frekuensi kemunculan dalam sebuah dokumen maupun frekuensi kemunculan untuk setiap dokumen. Selanjutnya setelah dilakukan peng-*index*-an akan dilakukan menghitungan nilai TFIDF. Gambar 5.16 merupkan kode untuk menghitung nilai TFIDF.



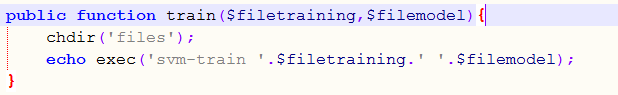


Gambar 5.16 Kode method getTfidf

Kode pada gambar 5.16 adalah kode untuk menghitung nilai TFIDF seperti yang sudah dijelaskan pada bab sebelumnya. Setelah berhasil dihitung nilai-nilai TFIDF yang dihasilkan akan disimpan dalam berkas pelatihan yang nantinya akan dibaca oleh libsvm.

## Pelatihan

Setelah semua siap maka sekarang proses pelatihan sudah dapat dilakukan. Data pembobotan yang sudah dibuat sebelumnya akan disimpan dalam sebuah berkas. Selanjutnya proses pelatihan ini akan implementasikan memanfaatkan *library* libsvm. Pada penelitian ini proses pelatihan ini di implementasikan pada *class* JanSvmComponent. Gambar 5.17 kode yang digunakan untuk memanggil fungsi pelatihan yang sudah diimplementasikan oleh libsvm.

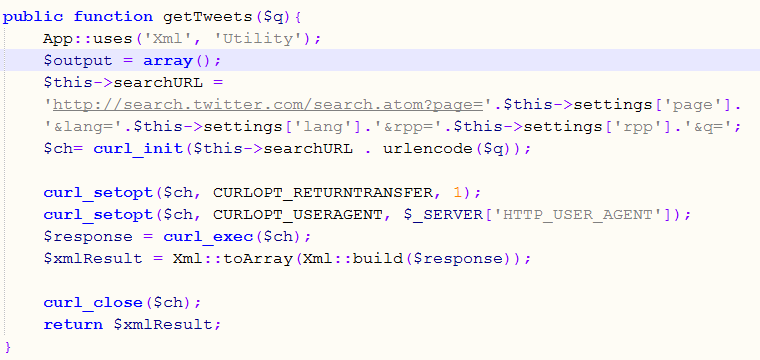


Gambar 5.17 Kode pemanggil *method* pelatihan

Setelah kode pada gambar 5.17 selesai dijalankan maka akan terbentuk SVM model yang disimpan dalam bentuk berkas. Nantinya berkas SVM model tersebut akan dimanfaatkan pada proses *testing*.

## Pencarian *Tweet*

Proses-proses sebelumnya yang sudah dilakukan adalah bagian dari proses utama yaitu proses pelatihan. Pada bagian subbab ini dan seterusnya akan dibahas bagian-bagian dari proses utama yang kedua yaitu *Testing*. Subproses pertama pada proses *testing* ini adalah pengguna akan memasukan kata kunci yang ia akan cari sentimennya. Kemudian sistem akan mencari data *tweet* yang bersesuai dengan kata kunci yang dimasukan oleh pengguna. Proses pencarian ini menggunakan fitur API twitter yang memang sudah disediakan oleh pengembang twitter. Gambar 5.18 adalah kode yang digunakan untuk mencari data pada twitter. *Method* ini terletak pada *class* MyTwitterComponent.



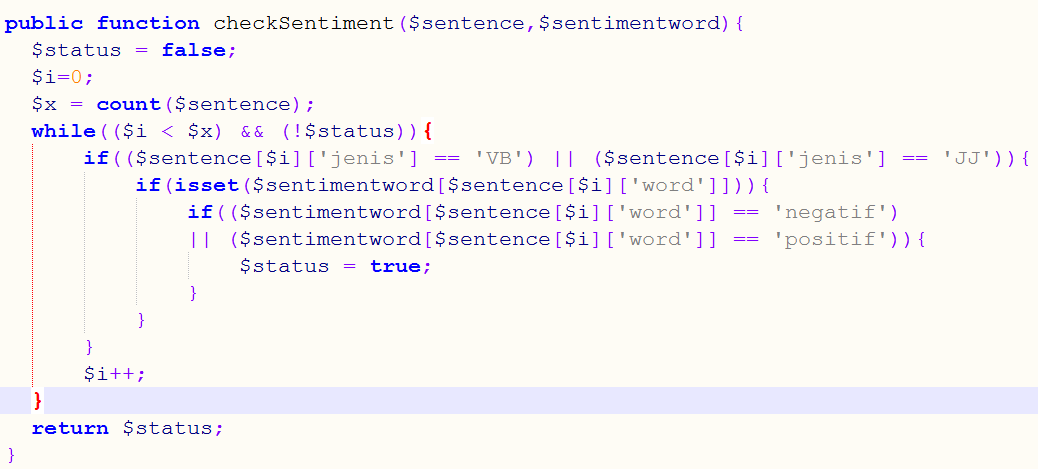
Gambar 5.18 Kode untuk mencari data pada twitter

Pada kode pada gambar 5.18 program akan mencari *tweet* yang bersesuai dengan kata kunci yang dimasukan oleh pengguna. Nilai kembalian dari twitter API dalam bentuk XML akan diubah dalam *array* yang nantinya akan disimpan dalam basisdata sebagai data kotor untuk selanjutnya dilakukan *preprocessing*.

## Pengecekan frase opini

Setelah mendapatkan *tweet* yang bersesuaian dengan kata kunci yang dimasukan dan disimpan dalam basisdata. Tweet-tweet tersebut akan di-*preprocessing*. *Preprocessing* yang dilakukan pada proses *testing* ini sama dengan *preprocessing* yang dilakukan pada proses pelatihan. Kemudian untuk melakukan pengecekan ada atau tidaknya frase yang mengandung opini maka harus dilakukan POS taging pada *tweet*-*tweet* yang sudah di-*preprocessing* tadi. Proses POS taging yang dilakukan pada proses *testing* juga sama dengan POS taging yang dilakukan pada proses pelatihan, jadi tidak akan dijelaskan kembali.

Proses berikutnya adalah proses pengecekan ada atau tidaknya frase opini. Gambar 5.19 merupakan kode untuk melakukan pengecekan ada atau tidaknya frase yang mengandung opini.



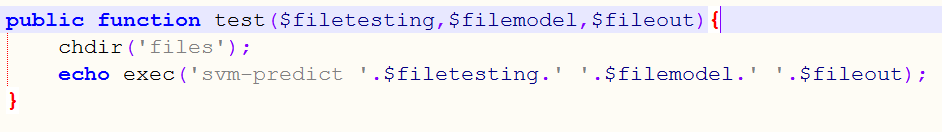
Gambar 5.19 Kode untuk memeriksa frase opini

Dalam menentukan ada atau tidaknya frase opini, program akan memeriksa apakah terdapat kata kerja(VB) atau kata sifat(JJ), jika tidak ada maka dianggap tidak memiliki frase opini tetapi apabila terdapat kata kerja atau kata sifat akan dilakukan pengecekan lagi. Apakah kata kerja atau kata sifat yang ditemukan mengandung sentimen atau tidak. Karena tidak semua kata kerja atau kata sifat yang memiliki sentimen. Untuk *tweet* yang tidak memiliki frase opini akan langsung diberi label netral sedangkan untuk *tweet* yang terindikasi memiliki frase opini akan dilanjutkan dalam proses pelabelan dengan menggunakan metode SVM.

## Pelabelan dengan metode SVM

Setelah berhasil dipisahkan data *tweet* yang mengandung frase opini dan yang tidak. Langkah berikutnya adalah mengkonversi *tweet* yang mengandung frase opini menjadi bentuk *vector* yang dapat dikenali oleh SVM. Metode yang digunakan masih sama dengan yang digunakan pada proses pelatihan yaitu menggunakan TFIDF. Implementasi dari TFIDF juga sudah dibahas sebelumnya.

Proses pelabelan dengan metode SVM dilakukan dengan bantuan SVM *library* yang dipanggil melalui *class* JanSvmComponent. Gambar 5.20 adalah kode implementasi pemanggilan libsvm.



Gambar 5.20 Kode pemanggilan libsvm

Dengan menjalankan kode pada gambar 5.20 maka akan dilakukan proses *testing* oleh libsvm. Proses ini akan beri label 1 atau -1 untuk setiap *vector*. Langkah terakhir adalah menyimpan dalam basisdata dengan label yang bersesuaian dengan dokumennya. Label 1 diartikan positif dan label -1 diartikan negatif. Gambar 5.21 adalah kode untuk menyimpan label yang sudah dihasilkan oleh libsvm ke dalam basisdata.



Gambar 5.21 Kode untuk menyimpan hasil pelabelan

Sampai dengan proses ini maka telah selesai proses pelabelan data dengan metode SVM. Data *tweet* yang baru saja cari sudah diberikan label dan siap ditampilkan kepada pengguna.

# 

# HASIL PENELITIAN DAN PEMBAHASAN

Bab VI memuat hasil penelitian yang terdiri dari tiga bagian. Pertama, perbandingan antara proses pelabelan data pelatihan yang dilakukan paulina 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 paulina. Pada pelatihan yang dilakukan paulina metode yang digunakan adalah dengan metode emoticon. Metode emoticon adalah dengan cara mendeteksi emoticon-emoticon tertentu, yang setiap emoticon memiliki korelasi dengan label.

### Hasil perbandingan pelabelan data pelatihan

Percobaan dilakukan menggunakan 300 data yang diambil secara random yang sebelumnya sudah diberi label pada penelitian paulina. Pada percobaan ini yang dibandingan adalah tingkat akurasi antara metode emoticon yang digunakan pada penelitian paulina 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 0.48666666666667 yang secara detail ditampilkan pada Gambar 6.1.



Gambar 6.1 Perbandingan metode emoticon dan pelabelan manual

Pada pengujian yang dilakukan terhadap metode *lexicon* *based* dan pelabelan manual memiliki akurasi 0.68 yang secara detail ditampilkan pada Gambar 6.2



Gambar 6.2 Perbandingan metode lexicon based dan pelabelan manual

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 0.48667.

## Pemberian Label

Proses pemberian label dimulai dari pengguna memasukan kata kunci topik yang akan dicari kecenderungan sentimennya. Gambar 6.3 merupakan tampilan dari input kata kunci yang diminta oleh sistem.



Gambar 6.3 Tampilan input kata kunci

Pada gambar 6.3 menunjukan tampilan input kata kunci. Pengguna sedang akan mencari tahu kecenderungan sentimen dengan kata kunci “telkomsel”. Selanjutnya pengguna harus menekan tombol Search untuk melanjutkan proses. Sistem akan mencari pada twitter, *tweet* yang didalamnya terdapat kata “telkomsel”. Proses pencarian ini menggunakan API yang sudah disediakan oleh pengembang twitter. Setelah ditemukan *tweet* yang mengandung kata “telkomsel” maka tweet yang sudah ditemukan akan disimpan dalam basis data.

Proses berikutnya adalah melakukan preprocessing pada data tweet yang sudah ditemukan. Gambar 6.4 menunjukan pengguna akan meminta sistem untuk melakukan preprocessing. Sistem akan melakukan preprocessing setelah pengguna menekan link “lakukan proprocessing”.



Gambar 6.4 link untuk melakukan preprocessing

Setelah proprocessing selesai dijalankan. Data *tweet* baru yang sudah dilakukan preprocessing akan kembali simpan sebagai data bersih didalam basis data. Data bersih yang baru saja disimpan akan ditampilkan kepada pengguna. Gambar 6.5 merupakan contoh data bersih yang ditampilkan.



Gambar 6.5 sistem menampilkan data bersih

Pada gambar 6.5 data yang ditampilkan adalah data yang sudah dilakukan preprocessing dan yang belum dilakukan preprocessing. Jadi sangat mudah untuk melihat perbedaan antara data yang sudah dilakukan preprocessing dan yang belum. Selain menampilkan data *tweet* , pada halaman ini juga terdapat link untuk melakukan proses-proses berikutnya. Misal, “lakukan analisis awal”,”generate test” dan “analysis”

Selanjutnya ada dilakukan proses analisis awal, yaitu memisahkan antara *tweet* yang mengandung frase opini dan yang tidak. *Tweet* yang mengandung frase opini akan dimasukan dalam proses pelabelan dengan SVM sedangkan yang tidak mengandung opini akan langsung diberikan label netral. Gambar 6.6 menunjukan data tweet setelah dilakukan analisis awal.



Gambar 6.6 Data tweet setelah dilakukan analisis awal

Pada gambar 6.6 terlihat beberapa *tweet*  sudah diberi label netral itu menandakan *tweet* tersebut tidak mengandung frase opini. Sedangkan yang lainnya belum diberi label, hal itu menunjukan dalam *tweet* tersebut mengandung frase opini. Seperti telah dijelaskan sebelumnya *tweet* yang mengandung frase opini akan masukan dalam proses pelabelan dengan SVM. Langkah pertama adalah dengan melakukan perhitungan bobot untuk *tweet* yang mengandung frase opini. Pengguna dapat menekan link “Generate test” untuk melakukan pembobotan, seperti ditunjukan pada gambar 6.7.



Gambar 6.7 Pembobotan tweet

Pada gambar 6.7 terlihat setelah sistem melakukan pembobotan akan muncul notifikasi bahwa pembobotan data testing sudah selesai dilakukan. Hal ini berarti proses pelabelan dengan SVM sudah siap untuk dilakukan. Pengguna dapat melakukan pelabelan dengan SVM dengan menekan *link* “Analysis”.

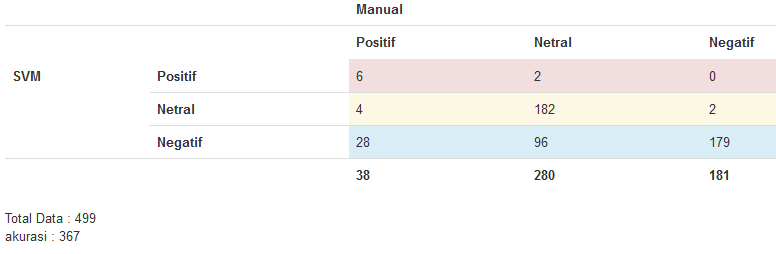


Gambar 6.8 Hasil pelabelan dengan metode SVM

Gambar 6.8 menunjukan hasil analisis pelabelan *tweet* yang memiliki frase opini dengan metode SVM. Hasil pelabelan ini akan langsung dipisahkan antara *tweet* yang bersentimen positif, negatif dan netral. Hasil dari pelabelan ini juga disimpan dalam basis data.

## 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 6.9 menunjukan hasil perbandingan yang sudah dilakukan. Pada gambar 6.9 ditunjukan dari 499 *tweet* , jumlah *tweet* yang diberi label oleh sistem dengan benar berjumlah 367 *tweet*. Berdasarkan data tersebut, akurasi dari penelitian ini adalah 0.735



Gambar 6.9 Akurasi sistem

# 

# KESIMPULAN DAN SARAN

## Kesimpulan

Penelitian dan percobaan telah dilakukan, menghasilkan kesimpulan sebagai berikut :

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 48,7%. 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 77,2%. 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

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.

Güngör, T., 2010, *Part-of-Speech Tagging*, Indurkhya N. dan Damerau F.J., *Handbook of Natural Language Processing*, edisi 2, CRC Press, Florida.

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

Jurafsky, D. dan Martin, J.H., 2008, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, edisi 2, Prentice Hall, New Jersey.

Kao,A., dan Poteet,S., 2006 *Natural Language Processing and Text Mining* , Springer, New York.

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

Pang, B. & Lee, L. 2008. *Subjectivity Detection and Opinion Identification. Opinion Mining and Sentiment Analysis*. Now Publishers Inc

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.

Manning, C. dan Schütze, H., 1999, *Foundations of Statistical Natural LanguageProcessing*, MIT Press, Cambridge.

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*