Opinion Mining Pada Twitter Untuk Bahasa Indonesia Dengan metode Support Vector Machine dan Metode Berbasis Lexicon

**Jan Kristanto\*1, Edi Winarko2**

1Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta

2Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta

e-mail: **\***[**1jan\_kristanto@ugm.ac.id**](mailto:1jan_kristanto@ugm.ac.id), [2ewinarko@ugm.ac.id](mailto:2ewinarko@ugm.ac.id)

***Abstrak***

*Dalam proses pengambilan keputusan opini atau pendapat orang lain sering dijadikan acuan. Dengan internet dan sosial media yang ada sekarang banyak orang sering berpendapat,mengeluh atau mengkritik tentang sesuatu hal. Sehingga untuk mengetahui pendapat orang lain tentang sesuatu hal kita tidak perlu bertanya kepada mereka satu per satu. Kita dapat mengembangkan metode yang dapat mencari dan menganalisa teks pada sosial media yang menghasilkan output berupa pendapat tentang sesuatu hal.*

*Metode untuk mendapatkan informasi berupa pendapat yang terdapat dalam suatu teks adalah “Opinion Mining” atau “Sentiment Analysis”. Pada penelitian ini menggunakan metode gabungan Unsupervised yang berbasis pada lexicon berbahasa indonesia dan Supervised yang berbasis pada Support Vector Machine.*

*Hasil yang diperoleh berupa polaritas sentimen apakah negatif , positif, ataupun netral. Dari hasil pengujian pada penelitian ini didapatkan akurasi 77,7%.*

***Kata kunci***— opinion mining, sentiment analysis, lexicon, support vector machine

***Abstract***

*Opinions in the decision-making process are often used as a reference. Nowadays, with the internet and social media there are a lot of people often argue, complain or criticize about something. So as to know what others think of something that we do not need to ask them one by one. We can develop a method that can search and analyze text on social media which produces output in the form an opinion about something.*

*The method to obtain information in the form of opinions contained in the text is an "Opinion Mining" or "Sentiment Analysis". In this research, using a combined Unsupervised method based on Indonesian Lexicon and Supervised based on Support Vector Machine.*

*Results obtained in the form of sentiment is negative polarity, positive, or neutral. From the test results obtained in this study 77,7% accuracy.*

***Keywords***— opinion mining, sentiment analysis, lexicon, support vector machine

1. PENDAHULUAN

O

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

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

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

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

2. METODE PENELITIAN

## 2.1 Pengumpulan Data

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

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

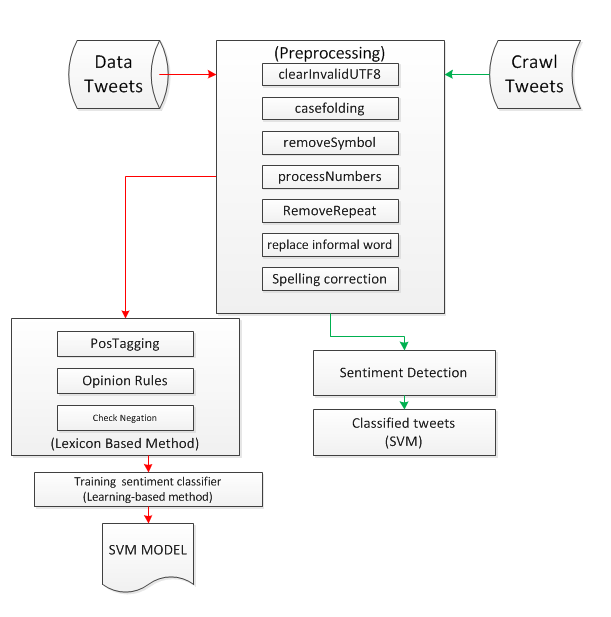
## 2. 2 Analisis Sistem

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

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

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

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



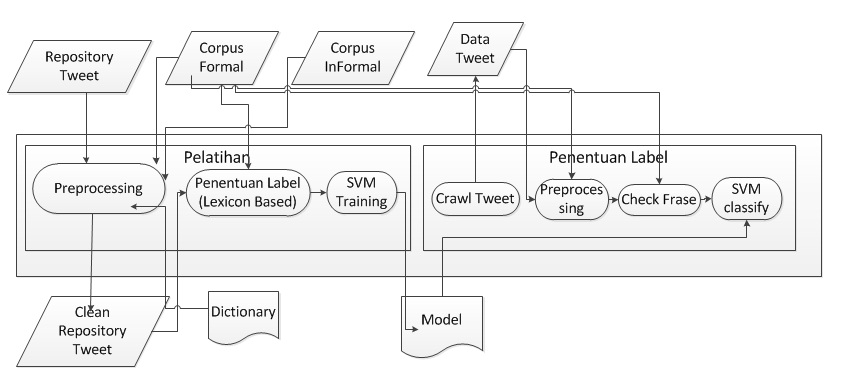
Gambar . Proses Opinion Mining

## 2. 3 Garis Besar Sistem

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

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

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



Gambar . Arsitektur Sistem

## 2.4 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).

## 2.4.1 clear invalid UTF8

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.

## 2.4.2 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.

## 2.4.3 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.

## 2.4.4 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 1.

Tabel . Konversi Angka

| **Angka** | **konversi** |
| --- | --- |
| 0 | o |
| 1 | i |
| 2 | Copy karakter – karakter sebelumnya |
| 3 | e |
| 4 | a |
| 5 | s |
| 6 | g |
| 7 | t |
| 8 | b |
| 9 | g |

## 2.4.5 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 2.

Tabel . Contoh Menghapus Karakter Berulang

| **Input** | **Output** |
| --- | --- |
| Tidaaaaak | Tidaak |
| Cemunguuuuudh | Cemunguudh |
| Eaaaaaa | eaa |

*2.4.6 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 penelitian Komarsilam dan Winarko [9], yang sudah ditambah secara manual berdasarkan pengamatan dalam proses pengembangan sistem ini. Contoh dari proses ini akan ditunjukan pada tabel 3.

Tabel . Contoh Penggantian Kata Tidak Baku

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

*2.4.7 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 algoritma levenshtein. Algoritma levenshtein adalah algoritma yang dapat menghitung kedekatan atau 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 adalah contoh kata-kata yang dapat ditangani oleh proses ini.

Tabel . Contoh Spelling Correction

| **Input** | **Output** |
| --- | --- |
| Bapk | Bapak |
| Menghitumg | menghitung |
| Yng | yang |

*2.5 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*.

*2.5.1 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.

*2.5.2 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 Komarsilam dan Winarko [9].

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 5 ditampilkan daftar rule frasa opini yang diambil dari penelitian Komarsilam dan Winarko [9].

Tabel . Opinion Rules

| **Rule** | **Keterangan** |
| --- | --- |
| RB JJ | Kata Keterangan, Kata Sifat |
| RB VB | Kata Keterangan, Kata Kerja |
| CK VB | Kata Sambung, Kata Kerja |
| NN JJ | Kata Benda, Kata sifat |
| NN VB | Kata Benda, Kata Kerja |
| VB JJ | Kata Kerja, Kata Sifat |
| JJ VB | Kata Sifat, Kata Kerja |

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.

*2.5.2 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.

*2.6 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).

*2.7 Generate bobot*

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 yang dirumuskan pada persamaan (1).

= x (1)

Dimana :

*d : kalimat ke-d*

*t : kata (term)*

*: term frequency pada kalimat ke-d*

*: bobot kata (term) dalam kalimat ke-d*

*: inverse document frequency suatu term*

Sedangkan nilai IDF sebuah term dihitung menggunakan persamaan (2)

IDF = log ( N / ) (2)

Dimana :

N : jumlah kalimat

: jumlah kemunculan kata (term)

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.

*2.8 Training dengan SVM*

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

*2.9 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.

3. HASIL DAN PEMBAHASAN

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

*3.1 Perbandingan proses pelabelan data pelatihan*

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

Percobaan dilakukan menggunakan 300 data yang diambil secara random yang sebelumnya sudah diberi label pada penelitian Alliandu dan Winarko[3]. Pada percobaan ini yang dibandingan adalah tingkat akurasi antara metode emoticon yang digunakan pada penelitian Alliandu dan Winarko [3] dan pelabelan manual dibandingkan dengan metode *lexicon* *based* yang dilakukan pada penelitian ini dan pelabelan manual. Pada pengujian yang dilakukan terhadap metode emoticon dan pelabelan manual memiliki akurasi 0.48666666666667 yang secara detail ditampilkan pada Tabel 6.

Tabel . Perbandingan Metode Emoticon dan Manual

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

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

Tabel . Perbandingan Metode Lexicon dan Manual

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

Berdasarkan pada kedua percobaan yang sudah dilakukan, pelabelan yang dilakukan dengan metode lexicon based ternyata memiliki akurasi yang lebih baik yaitu 0.68 dibandingkan pelabelan yang dilakukan dengan metode emoticon yaitu hanya 0.48667.

*3.2 Akurasi Pemberian label*

Penghitungan akurasi dari penelitian ini adalah dengan membandingkan hasil pelabelan yang dilakukan dengan metode kombinasi antara *lexicon* *based* dan SVM dibandingkan dengan pelabelan yang dilakukan secara manual. Data yang digunakan untuk perbandingan adalah data dari 3 kali percobaan dengan masukan kata kunci yang berbeda-beda. Berdasarkan 3 percobaan yang dilakukan didapatkan 499 data *tweet.* Gambar 8 menunjukan hasil perbandingan yang sudah dilakukan. Pada gambar 6.9 ditunjukan dari 499 *tweet* , jumlah *tweet* yang diberi label oleh sistem dengan benar berjumlah 388 *tweet*. Berdasarkan data tersebut, akurasi dari penelitian ini adalah 0.777

Tabel . Akurasi Sistem

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

4. KESIMPULAN

Berdasarkan hasil pengujian yang telah diperoleh, dapat ditarik kesimpulan:

1. Pada proses pengumpulan data training metode lexicon based memberikan akurasi yang lebih baik dari pada metode emoticon. Dalam penelitian ini metode lexicon based mencapai akurasi 68% sedangkan metode emoticon hanya mencapai akurasi 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 77,7%. Pengujian ini berdasarkan 499 tweet random berbahasa indonesia yang diambil langsung dari twitter.

5. SARAN

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

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

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

DAFTAR PUSTAKA

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

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

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

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

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

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

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

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

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