# Sentiment Analysis Berbasis Big Data
## *Sentiment Analysis Based Big Data*

Petrix Nomleni[1)], Mochamad Hariadi[2)], I Ketut Eddy Purnama[3)]

[1,2,3)]*Jurusan Teknik Elektro, Institut Teknologi Sepuluh Nopemeber*
*Kampus ITS Keputih, Sukolilo, Surabaya, 60111*
*INDONESIA*
*Email: petrix13@mhs.ee.its.ac.id*

**Abstrak**

Pemerintah sebagai pelayan masyarakat memiliki peran yang sangat besar dalam meningkatkan kesejahteraan masyarakat.Maka perlu diadakan suatu perbaikan secara bertahap guna meningkatkan pelayanan masyarakat (*public services*) sebagai tugas utama pemerintah, untuk itu perlu adanya sikap keterbukaan dari pemerintah untuk dapat menerima setiap keluhan masyarakat mengenai kebijakan / program yang langsung menyentuh kepentingan masyarakat. Media Center merupakan sistem pelayanan informasi yang terintegrasi kepada masyarakat Kota Surabaya untuk ikut berpartisipasi dalam pembangunan dengan berbagai cara seperti ide, pengaduan, keluhan, kritik, saran dan pertanyaan. Untuk itu perlu adanya klasifikasi untuk *sentiment analysis* keluhan masyarakat informasi yang masuk ke media center sehingga pengelola dapat memberikan informasi yang efisien dan tepat kepada masyarakat dan pemerintah dapat mengetahui bidang mana yang perlu dibenahi dalam pembangunan. *Sentiment analysis* merupakan proses klasifikasi dokumen tekstual ke dalam beberapa kelas seperti sentimen *positif* dan *negatif* serta besarnya pengaruh dan manfaat dari *sentiment analysis* tersebut. Pada penelitian ini dibahas klasifikasi keluhan masyarakat terhadap pemerintah pada media sosial facebook dan twitter sapawarga data berbahasa Indonesia menggunakan metode *Support Vector Machine (SVM)* yang dijalankan dalam komputasi terdistribusi dengan menggunakan *Hadoop*. Pengujian dilakukan dengan perhitungan precision, kecepatan, akurasi.Hal ini bertujuan untuk mengetahui sejauh mana kehandalan metode yang diusulkan untuk mencapai peningkatan kecepatan dan akurasi klasifikasi.

Kata kunci: *Media Center, Hadoop, Support Vector Machine, klasifikasi, sentiment analysis*

## 1. PENDAHULUAN

Dalam era perkembangan teknologi informasi yang semakin pesat di Indonesia saat ini, keterbukaan atau transparansi merupakan suatu hal yang sangat penting dalam rangka melaksanakan fungsi pengontrolan.Seperti kita ketahui bahwa kehadiran pemerintah sebagai pelayan masyarakat memiliki peran yang sangat besar dalam meningkatkan kesejahteraan masyarakat.Sistem birokrasi yang ada sekarang ini yang dianggap sebagai sarang korupsi, kolusi dan nepotisme (KKN), penghambat investasi dan lain-lain. Untuk itu perlu diadakan suatu perbaikan secara bertahap guna meningkatkan pelayanan masyarakat (*public services*) sebagai tugas utama pemerintah, maka terlebih dahulu perlu adanya sikap keterbukaan dari pemerintah untuk dapat menerima setiap kritik, saran ataupun keluhan masyarakat mengenai kebijakan/program yang langsung menyentuh kepentingan masyarakat misalnya penyelewengan kebijakan dan lain-lain. Hal ini juga dianggap sangat penting untuk mengaktifkan peran masyarakat, LSM dan lain-lain sebagai suatu fungsi kontrol terhadap setiap kebijakan pemerintah.

Perkembangan teknologi yang sangat cepat, tentunya membuka peluang untuk mewujudkan harapan baru.Dengan adanya konsep *e-government* sebagai salah satu upaya yang dikembangkan untuk memperbaiki sistem birokrasi tentunya perlu dimanfaatkan semaksimal mungkin. Untuk itu Pemerintah Kota Surabaya membangun media center yang merupakan sistem pelayanan informasi yang terintegrasi kepada masyarakat Kota Surabaya untuk ikut berpartisipasi dalam pembangunan dengan berbagai cara seperti ide, pengaduan, keluhan, kritik, saran dan pertanyaan. Sebelum adanya media center masyarakat memberikan keluhannya secara manual dengan mendatangi langsung ke Dinas Kominfo Kota Surabaya, tetapi dengan adanya media komunikasi media center memberikan kesempatan kepada masyarakat untuk memberikan keluhannya tanpa harus datang ke media center.

Namun dengan adanya media center yang menampung keluhan masyarakat ini menimbulkan pekerjaan baru bagi petugas pengelola data media center untuk memproses dan memilah-milah data keluhan masyarakat yang masuk. Metode yang digunakan saat ini dengan mencetak isi dari *website* media center tersebut, baru kemudian meneruskan keluhan yang sudah dipilah-pilah secara manual. Dengan meningkatnya jumlah keluhan masyarakat yang masuk, maka proses memilah-milah keluhan secara manual menjadi sangat tidak efisien untuk dilakukan.

Sebagai solusi permasalahan, maka diperlukan sebuah proses yang berjalan secara otomatis untuk melakukan perklasifikasian dan *sentiment*

*analysis*data keluhan masyarakat yang masuk. Manfaat *sentiment analysis* sangat penting untuk mengetahui sejauh mana data keluhan masyarakat terhadap pembangunan serta digunakan sebagai alat bantu untuk melihat respon masyarakat terhadap pembangunan kota Surabaya. Mengingat jumlah data keluhan yang masuk begitu besar maka diperlukan sebuah proses analisa data yang mampu menangani hal ini. Salah satu alternatif yang tersedia saat ini adalah menggunakan analisa *big data*.Karakteristik data sumber dari analisa big data adalah data yang memiliki 3 karakteristik yaitu *volume*(ukuran data yang besar), *variety*(tipe datanya bervariasi dari data tidak terstruktur dan data terstruktur) dan *velocity*(transaksi data dalam jumlah yang besar). Ini sesuai sekali dengan profil dari data *website* media center Pemkot Surabaya.

## 2. DASAR TEORI

### 2.1 *Sentiment Analysis*

*Sentiment analysis* adalah studi komputasi mengenai sikap, emosi, pendapat, penilaian, padangan dari sekumpulan teks yang fokusnya adalah mengekstraksi, mengindentifikasi atau menemukan karakteristik sentimen dalam unit teks menggunakan metode NLP(*Natural Language Processing*), statistik atau *machine learning* Sentiment analysis merupakan proses klasifikasi dokumen tekstual ke dalam beberapa kelas seperti sentimen *positif* dan *negatif* serta besarnya pengaruh dan manfaat dari sentimen analisis menyebabkan penelitian ataupun aplikasi mengenai *sentimet analysis*.

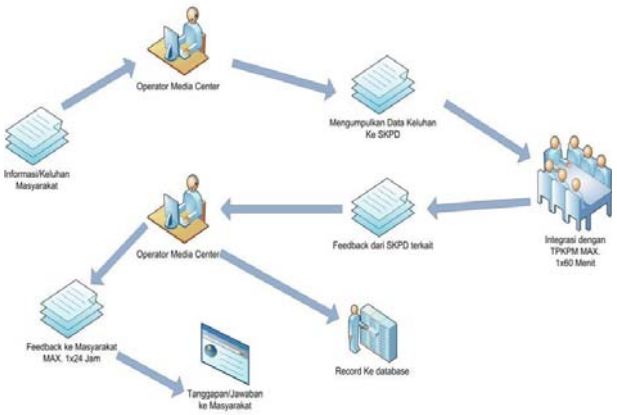
Saat ini perkembangan penelitian sentimen analisis mempunyai perkembangan yang sangat pesat bahkan di Amerika lebih dari 20 sampai 30 perusahaan memfokuskan pada layanan *sentiment analysis*. Pada dasarnya *sentiment analysis* merupakan klasifikasi, namun dalam implementasinya tidak mudah karena seperti proses klasifikasi biasa dikarenakan terkait penggunaan bahasa dimana terdapat ambigu dalam penggunaan kata, tidak adanya intonasi dalam sebuah teks, dan perkembangan dari bahasa itu sendiri.

*Sentiment analysis* bermanfaat juga dalam dunia usaha seperti melakukan analisa tentang sebuah produk yang dapat dilakukan secara cepat serta digunakan sebagai alat bantu untuk melihat respon konsumen terhadap produk tersebut, sehigga dapat membuat langkah-langkah strategis pada tahapan-tahapan berikutnya.

### 2.2 Media Center

Media Center adalah sistem pelayanan informasi terintegrasi yang memberikan kesempatan bagi masyarakat Surabaya yang ingin berpartisipasi dalam perkembangan pembangunan kota Surabaya dan bentuk partisipasi masyarakat terwujud dalam keluhan, pengaduan, kritik, saran dan pertanyaan yang terkait dengan proses pembangunan dan pelayanan yang dilaksanakan oleh Pemerintah Kota Surabaya. Sebelum adanya media center keluhan masyarakat langsung ke Dinas Kominfo dimana menjadi pusat media center, tetapi dengan adanya media komunikasi media center maka masyarakat dapat memberikan keluahan, saran, kritikan melalui media komunikasi tanpa harus langsung ke Dinas Kominfo.

Sejak awal pembangunannya Media Center mempunyai tiga karakteristik yaitu responsif (merespon setiap data keluhan masyarakat yang masuk kedalam Media Center) integratif(menggintegrasikan data keluhan masyarakat yang masuk ke Media Center) dan infomatif (memberikan informasi yang ter*update* kepada masyarakat). Dalam sistem kerjanya media center menerima informasi atau keluhan masyarakat melalui media komunikasi kemudian operator mengumpulkan informasi tersebut dan memberikan kepada SKPD terkait setelah itu integrasi data dengan TPKPM maksimal 1x60 menit setelah itu *feed back* dari SKPD terkait ke operator media center dan *feed back* ke masyarakat maksimal 1x24 jam dan kemudian data disimpan ke dalam database.[2]




**Gambar 2.1 SOP *Media Center***

### 2.3 *Big Data*

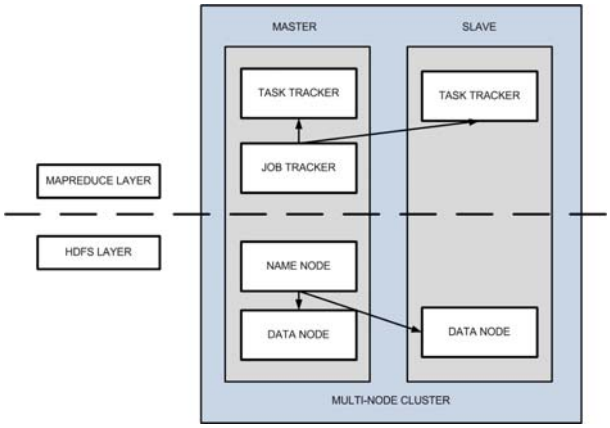
Saat ini proses pengolahan data baik dalam sistem pemerintahan maupun perusahaan swasta sudah menggunakan data center dan setiap bidang atau unit kerja sudah mempunyai *data center* dan hampir semuanya sudah terhubung antar satu dengan yang lainnya dan setiap hari datanya akan semakin bertambah dan semakin banyak variasi data yang yang disimpan serta jumlah transaksi data yang semakin besar maka diperlukan perangkat komputer yang sangat mahal dan membutuhkan tenaga IT yang sangat baik untuk mengoperasikannya.

Untuk itu diperlukan proses analisa *Big Data* yang dalam pengertiannya sebagai pemecahan masalah ketika teknologi lama tidak lagi mampu melayani proses pengolahan data yang sangat besar. *Big data* mempunyai tiga karakteristik yaitu *volume*(ukuran data yang besar dan terdistribusi di banyak server), *variety* (tipe data bervariasi dari data terstruktur hingga dari tidak terstruktur), dan *velocity* (jumlah transaksi data yang besar sehingga perubahan ukuran data juga akan semakin besar). Prinsip kerja big data yaitu tidak membuang atau menghapus sebuah data dikarenakan data tersebut menjadi penting dalam kurun waktu tertentu, proses

data secara real-time dan mampu mengekstrasi dan transformasi sebuah data tanpa menghapus data awalnya.

### 2.4 *Hadoop*

Hadoop merupakan sebuah *software framework* teknologi terbaru berbasis *Java* dan sangat mudah didapatkan karena hadoop merupakan software open source. Hadoop diciptakan untuk pengolahan data yang sangat besar hingga petabyte dimana pengolahan data-data tersebut dilakukan dengan cara mendistribusikan data-data tersebut kedalam beberapa komputer yang telah di cluster dan komputer-komputer tersebut terhubung satu dengan yang lainnya.




**Gambar 2.2 Hadoop Model**

Dalam perancangannya terdapat bagian seperti *Common Hadoop* yang fungsinya untuk menyediakan akses ke filesystem dan *Common Hadoop* berisi paket *file* dan *skrip* yang dibutuhkan *Hadoop* untuk memulai pekerjaannya. Paket ini menyediakan kode sumber, documen dan bagian kontribusi yang cakupannya sangat besar dan waktu penjadwalan kerja yang efektif.*File* sistem *Hadoop* harus kompatibel karena wajib memberikan lokasi jaringan yang dipakai agar node dapat bekerja.

Salah satu contoh ketika *cluster Hadoop* kecil yang didalamnya terdapat sebuah *master node* dan beberapa *node* untuk bekerja atau lebih dikenal dengan *slave node*. *Master node* terdiri dari beberapa bagian yaitu *jobtracker*, *tasktracker, name node,* dan *data node*.*Node* untuk bekerja terdiri dari *data node* dan *tasktracker*, walaupun hanya untuk mendapatkan pekerja node data, dan hanya pekerja node menghitung.

Pada sistem cluster yang sangat besar, *file* sistem *HDFS* dikerjakan dengan server *name node* diperuntukan pada host indeks *file* sistem, dan sebuah name node sekunder dapat menghasilkan snapshot dari struktur memori namenode, sehingga mencegah korupsi sistem file dan mengurangi hilangnya data. Demikian pula, *serverjobtracker* dapat mengelola penjadwalan kerja secara mandiri. Dalam *cluster Hadoop MapReduce* mesin digunakan meng*cloudfile* sistem alternatif, *name node* itu, name *node sekunder* dan arsitektur data *node* dari *HDFS* digantikan oleh setara *file* sistem-spesifik. Dalam sistem inti kerjanya *Hadoop* terdiri atas 2 bagianyaitu :

### 2.4. *HDFS(Hadoop Distributed File System)*

HDFS Merupakan sebuah *file* sistem yang fungsinya untuk menyimpan data yang sangat besar jumlahnya dengan cara mengdistribusi data-data tersebut kedalam banyak komputer yang saling berhubungan satu dengan yang lainnya. Cara kerjanya yaitu file yang masuk kemudian dipecah-pecah dalam bentuk blok sebesar 64 MB atau bisa dikonfigurasi sendiri besarnya. Kemudian data direplikasi kedalam beberapa *node*(biasanya 3 *node*), dan disimpan dalam beberapa rak yang berbeda dengan tujuan agar menjaga reability dari *HDFS*.

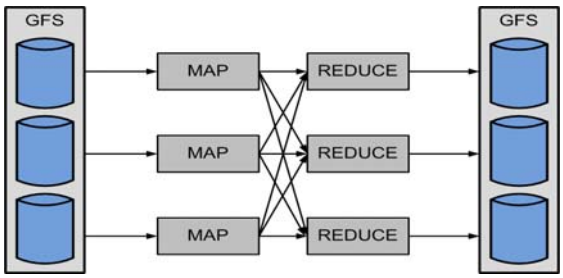
Untuk itu file system sangat membutuhkan *server* induk atau *master node* yang berfungsi untuk menyimpan metadata dari data yang ada di *HDFS* dan data-data tersebut disimpan dalam *server-server (datanode)* yang dapat diakses melalui protokol HTTP serta data nodenya saling terkait satu dengan lainnya untuk menjaga konsistensi datagan menggunakan protokol HTTP. Data node ini bisa saling berkomunikasi satu sama lain untuk menjaga konsistensi datadan memastikan proses replikasi data berjalan dengan baik.

HDFS mempunyai kelemahan yaitu *master node* bersifat *Single Point of Failure* yang akan membuat data akan hilang apabila *server master node* mati. Walaupun dalam *HDFS* ada *secondaryname node* tetapi tetapi secondary name *node* hanya menyimpan informasi terbaru dari struktur direktori pada name node. Untuk itu untuk mengatasi kelemahan yang ada maka dibuatkan *cloning* dari *server name node* ke beberapa *server* yang berbeda sehingga terjadi gangguan terhadap *name node* maka akan langsung digantikan oleh *cloning*nya.

Keuntungan dari *HDFS* adalah *jobtracker* dan *tasktracker* yang membuat jadwal dan peta serta mengurangi pekerjaan untuk *tasktrackers* pada lokasi-lokasi data.Sebagai contoh jika data pada *node* A (x, y, z) dan data yang terdapat *node* B (a, b, c).*jobtracker* akan jadwal node B untuk melakukan peta / mengurangi tugas pada (a, b, c) dan node A akan dijadwalkan untuk melakukan peta/mengurangi tugas pada (x, y, z). maka akan mengurangi jumlah lalu lintas yang berjalan di atas jaringan dan mencegah *transfer* data yang tidak perlu. *Hadoop* ketika digunakan dengan *file* sistem lain keunggulan ini tidak ada. Dan memberikan dampak yang signifikan terhadap waktu penyelesaian pekerjaan yang dapat ditunjukkan waktu data dijalankan dengan pekerjaan intensif.

### 2.4.2 *MapReduce*

Merupakan *software framework* yang digunakan untuk mendukung distribusi computing dengan menjalankan data yang sangat besar dan pertama kali diperkenalkan oleh *Google*.dalam proses kerjanya terdiri atas dua proses yaitu :




**Gambar 2.3 MapReduce**

#### 2.4.2.1 *Map*

Sebuah proses ketika *master node* menerima masukkan berupa data atau *file*, kemudian masukan tersebut dipecah menjadi beberapa bagian permasalahan yang kemudian didistribusikan ke *worker nodes*. *Worker nodes* ini akan memproses beberapa bagian permasalahan yang diterimanya untuk kemudian apabila masalah tersebut sudah diselesaikan, maka akan dikembalikan ke *master node*.

#### 2.4.2.2 *Reduce*

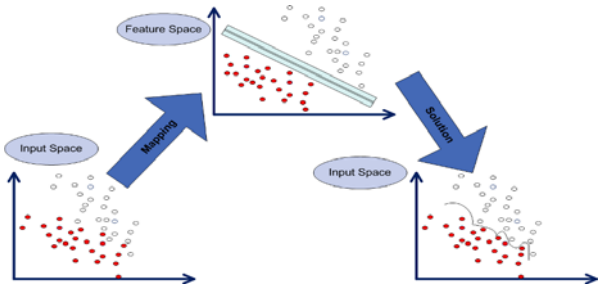
Sebuah proses ketika *master node*menerima jawaban dari semua bagian permasalahan dari banyak data *nodes*, kemudian menggabungkan jawaban-jawaban tersebut menjadi satu jawaban besar untuk menghasilkan penyelesaian dari permasalahan utama. Keuntungan *MapReduce* adalah proses *map* dan *reduce* dijalankan secara terdistribusi. Setiap proses *mapping* sifatnya *independen* yang membuat proses dijalankan secara simultan dan paralel. Begitu juga dengan proses *reduce*r dilakukan secara paralel pada waktu yang bersamaan, selama *output* dari operasi *mapping* mengirimkan *key value* yang sesuai dengan proses *reducer*nya. Dalam proses *MapReduce* dapat diaplikasikan di *cluster server* dengan jumlah yang banyak sehingga dapat mengolah data dalam jumlah besar hanya dalam beberapa jam saja.

Dalam kerja *hadoop*, *mapreduce engine* ini terdiri dari satu *jobtracker* dan satu/banyak *tasktracker*.*JobTracker* merupakan *server* penerima *job* dari *client*, kemudian mendistribusikan *jobs* tersebut ke *tasktracker* yang akan mengerjakan *sub job* sesuai yang diperintahkan *jobtracker*. Sistem kerja ini mendekatkan pengolahan data dengan data itu sendiri, sehingga ini akan sangat signifikan mempercepat proses pengolahan data. Dalam kerjanya *HDFS file* sistem bukan hanya diperuntukan untuk *map/reduce* tetapi saat ini ada beberapa *project* lain yang *related* dengan hadoop yang dapat dijalankan diatas *HDFS* seperti *HBase, Pig, Hive,* dll.

### 2.5 *Support Vector Machine(SVM)*

*Support Vector Machine (SVM)* pertamakali dikembangkan oleh *Boser, Guyon*,dan*Vapnik*.Pada tahun 1992 ketika diadakan di *Annual Workshop onComputational Learning Theory.Support Vector Machine (SVM)* merupakan sistem pembelajaran yang pengklasifikasiannya menggu-nakan ruang hipotesis berupa fungsi-fungsi linear dalam sebuah ruang fitur (*feature space*) berdimensi tinggi.Dalam konsep *SVM* berusaha menemukan fungsi pemisah (*hyperplane*) terbaik diantara fungsi yang tidak terbatas jumlahnya.

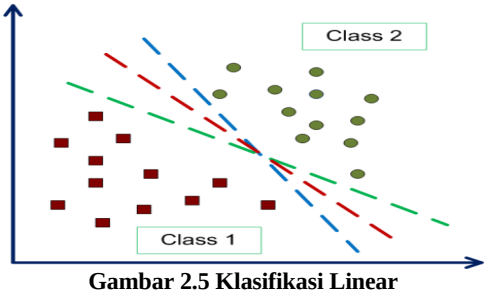
*Hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. Pada awalnya prinsip kerja dari *SVM*yaitu mengklasifikasi secara linear(linear classifier), kemudian SVM dikembangkan sehingga dapat bekerja pada klasifikasi non linear. Formulasi optimasi SVM untuk masalah klasifikasi dibedakan menjadi dua kelas yaitu klasifikasi *linear* dan klasifikasi *non-linear*.



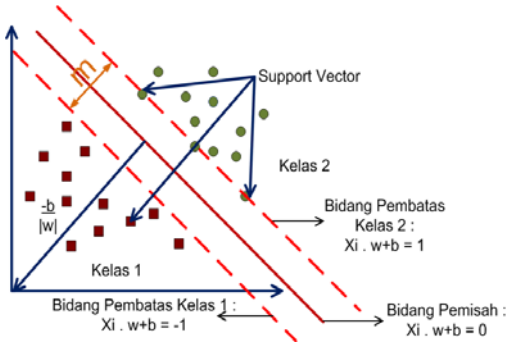

**Gambar 2.4*Support Vector Machine(SVM)***

### 2.5.1 Klasifikasi*Linear*

Dalam kerjanya SVM pada konsepnya dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada *input space*. Dua kelas, +1 dan -1, beserta masing-masing pattern. Dalam mengklasifikasi untuk mendapat hasil yang baik hyperplane digunakan untuk memisahkan menjadi dua kelas dengan mengukur margin *hyperplane* tersebut dan mencari titik maksimalnya, *margin* adalah jarak antara *hyperplane* terdekat dengan pattern terdekat dari masing-masing kelas dan pattern yang paling dekat dengan hyperplane disebut *support vector*.




**Gambar 2.5 Klasifikasi Linear**

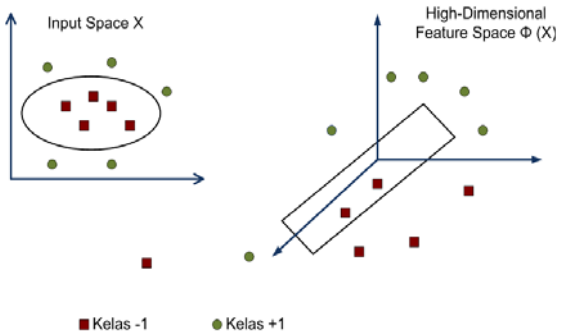



**Gambar2.6 Klasifikasi *Linear***

### 2.5.2 Klasifikasi Non Linear

Pada klasifikasi *non-linear* data yang berada dalam ruang sebuah *vector* awal harus dipindahkan ke ruang vector baru yang berdimensi lebih tinggi.Misal fungsi pemetaan dinotasikan sebagai x. Pemetaan ini bertujuan untuk merepresentasikan data ke format yang *linearly separable* pada ruang vector baru.Prosesnya optimisasi pada *fase* ini diperlukan perhitungan *dot product* dua buahcontoh pada ruang vector baru.Dot product kedua buah vector xi dan xj dinotasikan sebagai xi.xj.

Nilai dot product kedua vector ini dapat dihitung secara tidak langsung,yaitu tanpa mengetahui proses transformasi.




**Gambar 2.7 Klasifikasi *Non Linear***

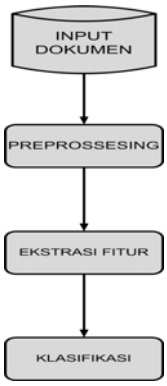
## 3. METODOLOGI PENELITIAN

### 3.1 Dasar Penelitian

Saat ini Pesan teks bahasa Indonesia dari media sosial seperti *Twitter* atau *Facebook*, orang cenderung menggunakan kata-kata tidak formal daripada yang formal seperti menggunakan angka untuk mengganti alfabet, karakter berulang vokal, dan menggunakan kata-kata informal yang umum untuk menggantikan kata-kata resmi. Untuk memproses kata-kata seperti itu, maka harus dilakukan tahapan-tahapan *preprocessing* seperti berikut:

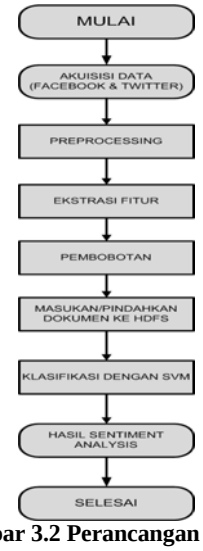
- *Converse* karakter numerik ke dalam alfabet, seperti ’du2k’ menjadi ’duduk’
- Hapus pengulangan vokal, seperti ’aduuuh’ menjadi ’aduh’
- Terjemahkan kata informal menjadi kata-kata resmi menggunakan kamus, seperti ’cemungudh’ ke ’Semangat’.

Meskipun beberapa kata informal salah eja dari kata-kata formal, tetapi beberapa lainnya benar-benar berbeda leksikal dari kata-kata formal, karena itu strategi adalah untuk membangun sebuah kamus dan menggunakannya untuk menerjemahkan kata informal menjadi kata-kata formal.[1]




**Gambar 3.1 Arsitektur Sistem Sentiment Analysis**

### 3.2 Perancangan Sistem




**Gambar 3.2 Perancangan Sistem**

### 3.2.1 Akuisisi Data

Pada tahapan awal ini data yang diakusisi atau dikumpulkan dari situs jejaring sosial Twitter dan Facebook yang terhubung langsung melalui *API (Application Programing Interface)* dan menambahkan proses deteksi bahasa untuk mendapatkan data atau dokumen yang berhasa indonesia.

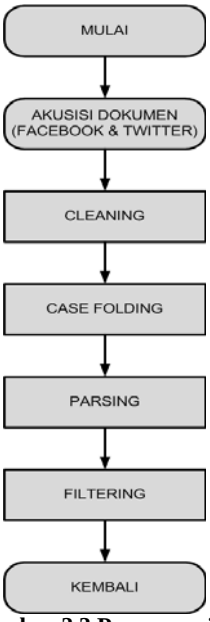
### 3.2.2Preprocessing

Pada tahapan ini dokumen yang diakusisi kemudian dimasukkan kedalam sistem. Dalam proses *presprocessing* ada beberapa tahapan yaitu:

- ***Cleaning***
  Tahapan atau proses membersihkan dokumen teks yang masuk dari kata-kata yang tidak diperlukan untuk mengurangi *noise* pada proses klasifikasi seperti karakter html, kata kunci,ikon,*hashtag* dan lain-lain.
- ***Case folding***
  Tahapan atau proses untuk menyeragamkan bentuk huruf serta menghilangkan tanda baca.

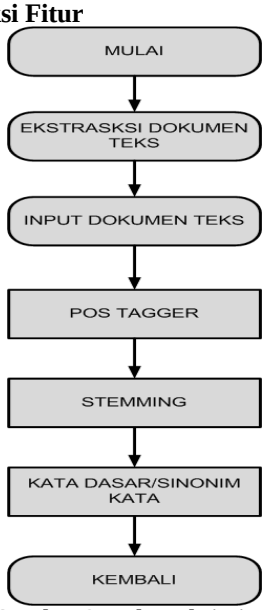
untuk penelitian ini hanya menerima huruf latin A sampai Z.

- ***Parsing***
  Tahapan atau proses membagi atau memecah dokumen menjadi sebuah kata dengan melakukan analisa terhadap kumpulan kata dengan memisahkan kata tersebut dan menentukan struktur sintaksis dari tiap kata tersebut.
- ***Filter***
  Tahapan untuk memilih data pada *twitter* dan *facebook* yang berbahasa Indonesia saja dan apabila ditemui kata berbahasa Indonesia tidak baku maka diganti dengan sinonimnya berupa kata baku yang sesuai dengan Kamus Besar Bahasa Indonesia(KBBI).




**Gambar 3.3 Preprocessing**

### 3.2.3 Ekstraksi Fitur




**Gambar 3.4 Ekstraksi Fitur**

- ***Part of Speech (POS)Tagger***
  Tahapan atau proses pada ekstrasi fitur yang bertujuan untuk memberikan kelas pada kata. Kelas kata yang dipilih adalah kata sifat (*adjective*), kata keterangan (*adverb*), kata benda (*noun*) dan kata kerja (*verb*), untuk diketahui dari empat jenis kata diatas merupakan jenis kata yang banyak mengandung sentimen. Penentuan kelas kata berdasarkan Kamus Besar Bahasa Indonesia(KBBI).
- ***Stemming***
  Tahapan atau proses dari ekstraksi fitur yang bertujuan mengurangi variasi kata yang memiliki kata dasar sama. Proses *stemming* juga menggunakan bantuan Kamus Besar Bahasa Indonesia(KBBI).

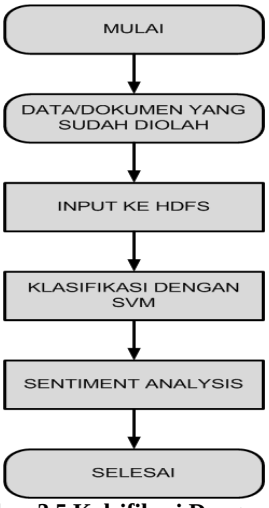
### 3.2.4 Pembobotan

Pada tahapan pembobotan ini fitur yang digunakan adalah unigram dengan pembobotan menggunakan Term Presense (TP), Term Frequency (TF), Term Frequency-Inverse Document Frequency.Pada tahapan ini kata dan simbol direpresentasi kedalam bentuk vektor dan TF-IDF, kata dan simbol direpresentasi ke dalam bentuk vektor dan tiap kata atau simbol dihitung sebagai satu fitur. Untuk perhitungan bobot digunakan rumus sebagai berikut:

- ***Term Presense (TP)***
  *ni(d)=1*, jika fitur *fi* ada didokumen *d*
  *ni(d)=0*, jika fitur *fi* tidak ada di dokumen *d*
- ***Term Frequency (TF)***
  $\rightarrow d:=(n1(d), n2(d), \cdots nm(d))$
- ***Term Frequency Inverse Document Frequency (TF-IDF)***
  $ni(d) = dfi.logD/dfi$

dimana :*dfi* adalah banyak data/dokumen yang mengandung fitur i(kata) yang dicari D adalah jumlah dokumen. Setelah perhitungan bobot tiap term dilakukan, selanjutnya proses penentuan kelas sentimen yang memberikan argumen maksimum dengan membandingkan nilai dari ketiga kelas *sentiment*tersebut.

### 3.2.5 Klasifikasi Dokumen




**Gambar 3.5 Kalsifikasi Dengan *SVM***

Pada penelitian ini dokumen yang sudah diolah kemudian dimasukkan kedalam *Hadoop Distributed File Sistem (HDFS)* yang kemudian diolah data tersebut menggunakan algoritma *SVM*.Untuk mendapatkan hasil klasifikasi terbaik, diujikan menggunakan tiga kelas sentimen kemudian membandingkan nilai dari tiga kelas tersebut.

### 3.2.6 Hasil Sentimen Analisis

Proses akhir menggunakan *3-fold cross validation* dengan membandingakn tiga kelas sedangkan untuk klasifikasi ditabulasi dalam tabel *confusion matrix*.

## 4. ANALISA HASIL

Dalam Pengujian data yang sudah diakuisisi dari *twitter* dan *facebook*, kemudian diberi nilai untuk mengetahui setiap atribut yang diuji pada setiap pengujian. Pengujian ini dilakukan dengan beberapa kali percobaan dengan data latih atau test yang berbeda. Pengujian ini untuk mengetahui seberapa besar kehandalan dari system untuk memberikan sebuah hasil yang lebih akurat. Pengujian pertama adalah pengujian untuk mengetahui bagaimana kemampuan sistem terhadap berbagai macam variasi data latih dan data uji dengan berbagai macam nilai.Pengujian kedua adalah pengujian untuk mengetahui pengaruh besarnya data latih terhadap akurasi sistem.Pengujian ketiga adalah pengujian untuk mengetahui akurasi metode klasifikasi SVMdalam melakukan klasifikasi pada data yang tidak seimbang.

Pengujian pengujian ini menggunakan 200 data yang dilakukan menggunakan data sebanyak 200 data.

**Tabel 4.1 Dataset Pengujian**

| Pengujian | Perbandingan | Jumlah Data Latih | Jumlah Data Uji |
|---|---|---|---|
| 1 | 60% : 40% | 120 | 80 |
| 2 | 50% : 50% | 100 | 100 |
| 3 | 40% : 60% | 80 | 120 |

Pada Pengujian pertama dengan data latih 120 data uji 80 akurasi pengukuran yang didapatkan sebesar 72,5% dan dari pengujian pertama didapatkan hasil seperti tabel dibawah ini.

**Tabel 4.1 Data Hasil Pengujian 1**

| Prediksi | Data Target | | | TP | FP | Precision | Recall | F-Maesure |
|---|---|---|---|---|---|---|---|---|
| | Positif | Netral | Negatif | (%) | (%) | (%) | (%) | (%) |
| Positif | 87 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Netral | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Negatif | 20 | 0 | 0 | 1 | 1 | 0.725 | 1 | 0.841 |
| weighted Avg | | | | 0.73 | 0.725 | 0.526 | 0.725 | 0.609 |

Pada pengujian kedua dengan data latih 100 dan data uji 100 akurasi pengukuran yang didapatkan sebesar 71% dan dalam pengujian kedua didapatkan hasil seperti tabel dibawah ini

**Tabel 4.2 Data Hasil Pengujian 2**

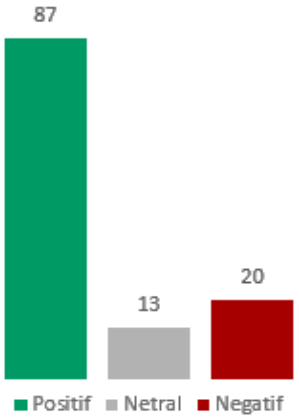
| Prediksi | Data Target | | | TP | FP | Precision | Recall | F-Maesure |
|---|---|---|---|---|---|---|---|---|
| | Positif | Netral | Negatif | (%) | (%) | (%) | (%) | (%) |
| Positif | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Netral | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Negatif | 19 | 0 | 0 | 1 | 1 | 0.71 | 1 | 0.83 |
| weighted Avg | | | | 0.71 | 0.71 | 0.504 | 0.71 | 0.59 |

Pada Pengujian ketiga dengan data latih 80 dan data uji 120 akurasi pengukuran yang didapatkan sebesar 70% dan dalam pengujian kedua didapatkan hasil seperti table dibawah ini

**Tabel 4.2 Hasil Pengujian 3**

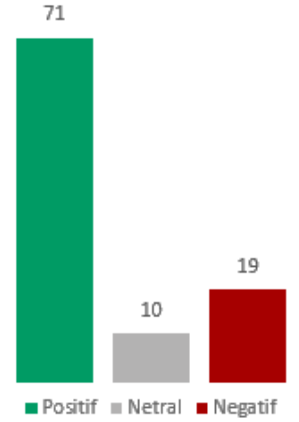
| Prediksi | Data Target | | | TP | FP | Precision | Recall | F-Maesure |
|---|---|---|---|---|---|---|---|---|
| | Positif | Netral | Negatif | (%) | (%) | (%) | (%) | (%) |
| Positif | 56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Netral | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Negatif | 16 | 0 | 0 | 1 | 1 | 0.7 | 1 | 0.824 |
| weighted Avg | | | | 0.7 | 0.7 | 0.49 | 0.71 | 0.576 |

Hasil Pengujian pertama dengan data latih 120 dapat digambarkan dalam grafik seperti dibawah ini



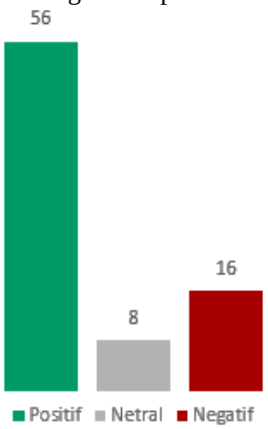

**Gambar 4.1 Grafik Pengujian 1**

Hasil Pengujian kedua dengan data latih 100 dapat digambarkan dalam grafik seperti dibawah ini




**Gambar 4.2 Grafik Pengujian 2**

Hasil Pengujian kedua dengan data latih 80 dapat digambarkan dalam grafik seperti dibawah ini




**Gambar 4.3 Grafik Pengujian 3**

## 5. KESIMPULAN

Dari hasil perancanngan, pengujian dan pembahasan diatas maka dapat disimpulkan :

1. Analisa terhadap setiap keluhan yang masuk ke *media center* Sapawarga Kota Surabaya dapat diklasifikaskan menggunakan metode klasifikasi *Support Vector Machine* untuk mendapatkan nilai sehingga diketahui polaritas keluhan yang masuk yang dpat menjadi acuan dan masukkan yang bernilai untuk tim media center dalam mengelola dan meningkatkan pelayanan terhadap masyarakat.
2. Keakuratan klasifikasi akan semakin tinggi apabila data latih yang dipergunakan semakin banyak tetapi tidak menutup kemungkinan jika kata-kata yang masuk bermakna ganda atau mengalami pembiasan maka akan mengurangi pembiasan.
3. HDFS sangat penting dalam penyimpanan pengelolaan data yang besar karena data tidak bias di hapus dan dapat dijalankan secara real time sehingga data yang lama dapat ditemukan kembali dalam waktu yang relatif singkat.

## UCAPAN TERIMA KASIH

Terima kasih kepada Tuhan Yesus yang selalu menjaga dan melindungiku dan kepada istriku serta semua keluargaku tercinta yang selalu berdoa untukku.

Ucapan terima kasih juga penulis kepada :

- Bapak Mochamad Hariadi, ST., M.Sc., Ph.Ddan Bapak DR. I Ketut Eddy Purnama, ST, MT sebagai dosen pembimbingku
- Staf Pengajar Jurusan Teknik Elektro Institut Teknologi Sepuluh Nopember
- Teman-teman CIO ITS angkatan 2013
- Semua pihak yang membantu saya baik moril substansi maupun materi untuk dapat menyelesaikan paper ini walaupun masih baik kekurangan.

*God Bless You All*

## REFERENCES

Edwin Lunando, Ayu Purwarianti,2013, *Indonesian Social Media Sentiment Analysis with Sarcasm Detection,.*

Media Center Pemerintah Kota Surabaya(Dinas Komunikasi & Informatika Kota Surabaya)

Neethu M S, Rajasree R, 2013 *Sentiment Analysis in Twitter using Machine Learning Techniques,*

Javier Conejero, Peter Burnap, Omer Rana, Jeffrey Morgan, 2013. *Scaling Archived Social Media Data Analysis using a Hadoop Cloud*

Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas ByShoba G, 2012, *Sentiment Analysis on Social Media*

Sang-Hyun Cho,Hang-Bong Kang Statistical Text *Analysis and Sentiment Classification in Social Media*

Visalakshi P and Karthik TU,2011 *MapReduce Scheduler Using Classifiers for Heterogeneous Workloads,*

Richard Colbaugh, Kristin Glass, 2011 *Agile Sentiment Analysis of Social Media Content for Security Informatics Applications*

Kazunari Ishida,2010*Periodic Topic Mining from Massive Amounts of Data,*

ang-Hyun Cho, Hang-Bong Kang, 2012, *Statistical Text Analysis and Sentiment Classification in Social Media*

Keke Cai, Scott Spangler, Ying Chen, Li Zhang, 2006, *Leveraging Sentiment Analysis for Topic Detection*

Muhamad Yusuf Nur dan Diaz D. Santika, 2011, Analisa Sentimen Pada Dokumen Bahasa Indonesia Dengan Pendekatan (*Support Vector Machine*),

Ni Wayan Sumartini Saraswati, 2013, *Naive Bayes Classifier* dan *Support Vector Machines* Untuk *Sentiment Analysis*,

Noviah Dwi Putranti dan Edi Winarko, 2013, Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan *Maximum Entropy* dan *Support Vector Machine,*

Imam Fahrur Rozi, Sholeh Hadi Pramono, Erfan Achmad Dahlan, 2012,Implementasi Opinion Mining (Analisis Sentimen)Implementasi untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi,

Judith Hurwits, Alan Nugent, Dr. Fern Helper, Marcia Kaufman, 2013 , *Big Data For Dummies A Wiley Brand,*

Irwin King, Michael R. Lyu, Haiqin Yang, 2013 *Online Learning for Big Data Analytics,*

Budi Santosa Tutorial *Support Vector Machine*