JURNAL TEKNOLOGI TECHNOSCIENTIA ISSN: 1979-8415
Vol. 9 No.1 Agustus 2016

APLIKASI OPINION MINING DAN SENTIMENT ANALYSIS UNTUK MERANCANG MESIN PENCARI OPINI PADA KUESENER MAHASISWA

Amir Hamzah[1], Naniek Widyastuti[2]

[1,2]Jurusan Teknik Informatika, Institut Sains & Teknologi AKPRIND Yogakarta

Masuk : 12 Juli 2016, revisi masuk : 1 Agustus 2016, diterima : 18 Agustus 2016

ABSTRACT

The measurement of academic services using questionnaires with multiple choice answers generally provide comments and advice columns. In the data analysis results, comments and suggestions made by the thousands of students can not be utilitized due to the lack of analysis tools. Whereas comments and suggestions can actually contain student opinions on various things, such as facilities, faculty, library and others. Opinion mining and sentiment analysis as a new tool in text mining can be applied to the data to utilitize comments and suggestions. This research applied HMM-POS Tagger to give automatically POS TAG to the sentence based on training POS TAG data by using the Hidden Markov Model. By implementing POS TAG pattern the comments can then be determined whether it was opinion or not. Morever if it were opinion it can be determinied its target and also the orientation of the opinion whether it is positive or negative. The data used was 1,000 comments given POS-TAG manually and 1,000 comments as test data. Sentiment analysis is applied using four methods of classification, namely SVM, NBC, ME and KM-Clustering. The result showed that the accuracy of POS-Tagger was 0.95 and the avarage of accuracy of four classification method was 0.85.

Keywords: *HMM POS_Tagger, opinion, classification*

INTISARI

Pengukuran pelayanan akademis menggunakan kuesener dengan jawaban multiple choice umumnya menyediakan kolom komentar dan saran. Pada analisis data, komentar dan saran yang berjumlah ribuan ini tidak dapat dimanfaatkan karena tidak adanya alat analisis. Padahal Komentar dan saran ini sebenarnya memuat opini mahasiswa tentang berbagai hal, misalnya sarana, dosen, perpustakaan dan lain-lain. Opinion mining dan sentiment analysis sebagai alat baru dalam text mining dapat diterapkan untuk memberdayakan data komentar dan saran. Penelitian ini menerapkan HMM-POS Tagger untuk memberikan POS-TAG otomatis pada data komentar berdasarkan data training POS-TAG menggunakan teknik Hidden Markov Model. Dengan menerapkan pola POS-Tag selanjutnya dapat ditetapkan apakah suatu kalimat komentar itu opini atau bukan. Lebih jauhnya jika ia opini, maka selanjutnya dapat dicari apakah target dari opini tersebut dan apakah orientasi opini tersebut positif atau negatif. Data yang digunakan adalah 1.000 komentar yang diberikan POS-TAG manual dan 1.000 komentar sebagai data uji. Sentiment analysis diterapkan dengan menggunakan 4 metode klasifikasi, yaitu SVM, NBC, ME dan KM-Clustering. Hasil menunjukkan bahwa POS Tagger memiliki akurasi 0.95 sedangkan rata-rata akurasi dari 4 metode klasifikasi adalah 0.85.

Kata Kunci :HMM POS_Tagger, opini, klasifikasi

PENDAHULUAN

Opini dan orientasi opini adalah bagian terpenting dalam pengambilan keputusan untuk suatu kebijakan. Keputusan yang tepat sangat dipengaruhi oleh analisis opini dari berbagai sumber yang terkait dengan pengambilan keputusan. Sebagai contoh pada dunia

bisnis, penambahan produk oleh manajer produksi sangat memerlukan analisis dari *review* produk barang yang ada di pasaran. Contoh lain misalnya pada manajemen pelayanan pendidikan di perguruan tinggi, pengukuran tentang tingkat kepuasan layanan pembelajaran dapat diukur dari opini mahasiswa tentang proses pembelajaran. Opini muncul pada berbagai situasi, misalnya yang dengan sengaja diminta oleh suatu alat penjajagan opini melalui permintaan saran dalam aktivitas kuesener, atau muncul secara alami dari suatu forum *on line* yang disediakan oleh situs resmi perguruan tinggi. Volume opini *on line* yang berupa teks bebas ini semakin hari semakin banyak dan umumnya tidak dimanfaatkan karena bentuknya yang tidak terstruktur.

Keberadaan internet dan sumber informasi *on-line* lainnya berkembang sangat pesat. Data dan informasi *online* dari perusahaan dan organisasi pada umumnya berbentuk tidak terstruktur dan umumnya berbentuk teks yang mencapai 80% (Grimes,2013). Ditemukannya media sosial seperti *Facebook (2004)* dan *Tweeter* (2006) telah mendorong kegiatan seperti *review*, forum diskusi, blog, *micro-blog*, komentar, dan posting yang melipatgandakan keberadaan dokumen teks di internet. Hal ini karena media sosial tersebut telah digunakan baik oleh individu maupun organisasi untuk berbagai kepentingan di dalam melakukan kegiatan sharing informasi. Kondisi ledakan informasi ini semakin menyulitkan proses *data mining* sebagaimana jauh hari telah diprediksi (Putten, et.al.,2002). Untuk itu pengembangan penelitian di bidang *opinion mining* menjadi topik yang sangat penting d isamping juga topik-topik sebelumnya, yaitu *data mining dan text mining.*

Salah satu cabang riset yang kemudian berkembang dari situasi ledakan informasi di internet adalah *sentiment analysis* dan *opinion mining*. Opinion mining menjadi riset yang menantang karena didalamnya terdapat akumulasi dari berbagai tantangan riset, yaitu dari bidang *Information Retrieval* (IR) : *information extraction*, *information summarizatian*, *document classification* dan dari bidang *Natural Language Processing* (NLP) seperti *Named Entity Recognition* (NER) dan *document subjectivity analysis* (Pang and Lee, 2002). Cabang riset ini mengkaji bagaimana seseorang mengekstrak opini dari media *on line* dan melakukan analisis terhadap opini tersebut. *Sentiment Analysis* atau *opinion mining* adalah studi komputasional dari opini-opini orang, *appraisal* dan emosi melalui entitas, *event* dan atribut yang dimiliki (Liu, 2010).

Aplikasi *sentiment analysis* dan *opinion mining* untuk melakukan evaluasi kebijakan dan pengambilan keputusan menjanjikan cara yang lebih praktis dan ekonomis dibandingkan dengan metode klasik menggunakan pendekatan kuesener. Kritik terhadap metode kuesener sebagai metode yang lama dan mahal, disamping juga memberikan hasil yang kadang kurang dapat menangkap problem yang sebenarnya. Kuesener dan interview dinilai lemah karena pada umumnya orang kurang suka menjawab pertanyaan survei yang kadang bertele-tele. Dalam posisi ini *opinion mining* menjawab persoalan penggalian opini dengan cara mendengar (*by listening*) dari pada dengan bertanya seperti kuesener (*by asking*), sehingga lebih akurat mencerminkan realitas sebenarnya (Shelke, et.al.,2012). Bahkan lebih jauh *opinion mining* memungkinkan untuk menangkap emosi pemilik opini (Loia and Senatore, 2014). Contoh bagus dalam masalah ini adalah penelitian Greaves et.al. (2013) di *English National Health Service website* yang menangkap 6.412 *comment* bebas dari pasien yang dirawat. Analisis tentang comment terkait dengan kebersihan, pelayanan rumah sakit dan berbagai aspek tanggung jawab rumah sakit memberikan hasil kesesuain antara 81% sampai 89% dibandingkan dengan metode rating kuantitative yang diberikan melalui kuesener.

Institut Sains dan Teknologi AKPRIND sebagai lembaga pendidikan tinggi senantiasa ingin meningkatkan layanan dalam manajemen untuk pembelajaran. Untuk maksud tersebut pada setiap akhir semester bagian administrasi akademik mengadakan evaluasi layanan pembelajaran yang

menggunakan instrumen kuesener dengan butir-butir jawaban yang telah disediakan. Selama ini ada data kuesener yang tidak dapat dimanfaatkan dan dianalisis yaitu data **saran** mahasiswa. Data ini jumlahnya mencapai ribuan **saran** atau lebih tepatnyua **opini** yang berasal dari seluruh seluruh peserta dari seluruh mata kuliah. Saran/opini dapat mengenai suasana akademik, dosen, ruang kuliah, AC, OHP, atau fasilitas kampus lainnya. Dalam beberapa tahun data ini semakin menumpuk.

Permasalahan dalam penelitian ini adalah bagaimana membangun suatu prototype perangkat lunak yang dapat melakukan ekstrasi opini dari suatu koleksi dokumen teks komentar, kemudian menentukan target opini dan orientasi opini.

Adapun tujuan dari penelitian ini adalah melakukan kajian penerapan teknik *opinion mining* dan *sentiment analysis* untuk menganaliasa data-data saran/opini mahasiswa. Penelitian ini juga dirancang untuk menciptakan prototipe perangkat lunak *opinion mining* dan *sentiment analysis* yang dapat melakukan ekstraksi opini, menganalisis opini, memetakan target opini dan menetapkan orientasi opini.

Metode yang digunakan untuk melakukan ekstraksi opini adalah dengan menerapkan *HMM-POS Tagger*. Metode ini diterapkan pada koleksi data latih yang merupakan kalimat opini dan kalimat non opini yang telah diberikan POS-TAG atau tanda *Part of Speech* dari kata-kata tersebut pada setiap kata. Program diharapkan dapat mengekstrak opini dari teks komentar kuesener, sekaligus mencari objek opini. Untuk orientasi opini dicobakan empat pendekatan metode klasifikasi yaitu menggunakan NBC (*Naïve Bayes Classifier*) dan SVM (*Support Vector Machine*), ME (*Maximum Entrophy)* dan KMC (*K-Means Clustering*).

HMM-POS Tagger adalah metode untuk melakukan POS-tagging pada suatu kalimat secara otomatis berdasarkan suatu analisis dan karakteristik data POS-tag dari koleksi data training. *Part-of-Speech (POS) tagging* atau dikenal dengan *grammatical tagging* adalah proses untuk memberikan suatu POS-tag pada suatu kata dalam suatu teks kalimat. *Part-of-speech* adalah kategori gramatikal kata-kata dalam suatu kalimat, misalnya kata kerja (VERB), kata benda (NOUN), kata sifat (ADJECTIVE), dan lain-lain. POS *tagging* merupakan alat penting dalam banyak aplikasi pengolahan bahasa alami seperti proses disambiguasi, *parsing*, sistem *question-answer,*dan *machine translation*. Dikarenakan menetapkan *part-of-speech* tag untuk kata-kata dalam kalimat secara manual sangatlah mahal, melelahkan, dan memakan waktu, maka telah muncul minat yang luas dalam otomatisasi proses *POS tagging* (Cutting et.al,1992).

Ada beberapa pendekatan untuk POS *tagging* otomatis, yaitu berdasarkan aturan (*rule base*), probabilistik, dan pendekatan berbasis transformasional. POS tagger berbasis aturan menetapkan tag ke kata berdasarkan beberapa aturan linguistik manual yang dibuat, misalnya kata adalah diberi tag NOUN jika mengikuti AJECTIVE. Pendekatan probabilistik menentukan tag kata dari suatu token berdasarkan pada probabilitas konteks tag token-token yang disekitarnya yang telah ditentukan secara manual dari suatu corpus. Pendekatan berbasis berbasis transformasional menggabungkan pendekatan berbasis aturan dan probabilistik untuk secara otomatis menurunkan aturan simbolik dari corpus (Pisceldo dkk, 2009). Penggunaan *Hidden Markov Model* untuk POS Tagger Bahasa Indonesia diteliti oleh Wicaksono dan Purwarianti (2010) dengan akurasi 96,2% dan Widhiyanti dan Harjoko (2012), yang menghasilkan akurasi 92,2%.

Apabila dimiliki suatu kalimat yang terdiri dari n buah kata (wi : i=1,..,n), dan akan ditetapkan POS-tag untuk setiap kata yang menyusun kalimat tersebut (ti : i=1,..,n), maka persoalan ini dapat dirumuskan sebagai mencari nilai maksimum dari :

$$\hat{t} = \arg\max_{t_1^n} P(t_1^n \mid w_1^n) \qquad (1)$$

Dengan menerapkan teorema Bayes dalam probabilitas bersyarat, maka (1) dapat ditulis menjadi :

$$\hat{t} = \arg\max \frac{P(w_1^n \mid t_1^n)P(t_1^n)}{P(w_1^n)} \quad (2)$$

Karena nilai penyebut selalu sama untuk setiap kalimat, maka (2) dapat ditulis menjadi :

$$\hat{t} = \arg\max P(w_1^n \mid t_1^n)P(t_1^n) \quad (3)$$

Dengan membuat dua asumsi, maka persamaan (3) dapat dituliskan :

(1) Probabilitas sebuah kata hanya tergantung pada POS-tagnya.

$$P(w_1^n \mid t_1^n) \approx \prod_{i=1}^{n} P(w_i \mid t_i) \quad (4)$$

(2) Probabilitas sebuah POS-tag hanya tergantung pada POS-tag sebelumnya.

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \quad (5)$$

Dengan menerapkan (4) dan (5) pada persamaan (3) akan diperoleh :

$$\hat{t} = \arg\max \prod_{i=1}^{n} P(w_i \mid t_i)\ P(t_i \mid t_{i-1}) \quad (6)$$

Selanjutnya dari hasil HMM-POS Tagger akan dihasilkan kalimat yang telah diberikan POS-tag. Pada langkah berikutnya berdasarkan pola-pola POS-Tag inilah akan ditetapkan apakah suatu teks itu opini atau bukan opini. Dengan pola-pola POS-Tag juga dapat ditentukan suatu objek dari opini.

Untuk menentukan orientasi opini diajukan 4 metode klasifikasi, yaitu metode *Naive Bayes Classifier (NBC), Support Vector Machine (SVM), Maximum Entrophy(ME) dan K-Means Clustering (KMC).*

Metode NBC mengasumsikan koleksi dokumen opini sebagai D= $\{d_1,d_2,\ldots,d_{|D|}\}$ dan koleksi kategori C= $\{c_1,c_2,\ldots,c_{|C|}\}$. Klasifikasi NBC dilakukan dengan cara mencari probabilitas $P(C=c_j \mid D=d_i)$, yaitu probabilitas category c_j jika diketahui dokumen d_i. Dokumen di dipandang sebagai tuple dari kata-kata dalam dokumen, yaitu <w1, w2,…,wn>, yang frekuensi kemunculannya diasumsikan sebagai variable random dengan distribusi probabilitas Bernoulli (McCallum and Nigam, 1998). Selanjutnya klasifikasi dokumen adalah mencari nilai maksimum dari :

$$V_{MAP} = \arg\max_{c_j \in C} P(C_j \mid w_1, w_2, \ldots, w_n) \quad (7)$$

Dengan menerakan teorema Bayes didapat :

V_{MAP}=

$$\arg\max_{c_j \in C} \frac{P(w_1, w_2, \ldots, w_n \mid c_j)P(c_j)}{P(w_1, w_2, \ldots, w_n)} \quad (8)$$

Dikarenakan nilai penyebut bersifat konstan untuk suatu dokumen, dan dengan mengasumsikan bahwa setiap kata adalah independen satu sama lain maka persamaan (8) dapat ditulis :

$$V_{MAP} = \arg\max_{c_j \in C} \prod_{i=1}^{n} P(w_i \mid c_j)P(c_j) \quad (9)$$

Secara praktis perhitungan $P(c_j)$ didekati dengan :

$$P(c_j) = \frac{|doc_j|}{|contoh|} \quad (10)$$

dimana $|doc_j|$ adalah banyaknya dokumen kategori j dan |contoh| banyaknya dokumen contoh (*training*). Sedangkan $P(w_i|c_j)$ didekati dengan :

$$P(wi|cj) = \frac{|n_i + 1|}{n + |vocabulary|} \quad (11)$$

dimana ni adalah frekuensi kemunculan kata w_i dalam kategori c_j, dan n adalah frekuensi kata dalam dokumen kategori c_j dan |vocabulary| banyaknya kemunculan seluruh kata dalam koleksi dokumen contoh.

Support Vector Machine (SVM) pertama kali dikembangkan oleh Boser et.al. (1992) dan dilanjutkan dengan uraian yang lebih detail oleh Cortes dan Vapnik (1995). Konsep SVM dapat dijelaskan sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space. Untuk koleksi dokumen dalam bentuk :

$$D = \{(xi, yi) \mid xi \in Rp\ ,\ yi \in \{-1,1\}\} \quad (12)$$

dimana yi adalah 1 atau -1, menunjukkan kelas mana titik x_i itu berada. Masing-masing xi adalah vektor nyata p-dimensi. Akan dicari *hyperplane* maksimum

margin yang membagi poin untuk poin yang memiliki $y_i = 1$ dari yang memiliki $y_i = -1$.

Hyperplane apapun dapat ditulis sebagai himpunan titik-titik x memuaskan w. x – b = 0 dimana titik (.) menunjukkan dot product. Vektor w adalah vektor normal: adalah tegak lurus *hyperplane* tersebut. Parameter ||w|| menentukan offset *hyperplane* dari asal sepanjang vektor normal w. Akan diupayakan memilih w dan b untuk memaksimalkan margin, atau jarak antara *hyperplane* paralel yang terpisah sejauh mungkin memisahkan data.

Hyperplanes ini dapat digambarkan oleh persamaan :

$$w \cdot x - b = 1 \quad (13)$$

dan

$$w \cdot x - b = -1 \quad (14)$$

Jarak antara kedua *hyperplane* adalah 2/||w||, jadi kita ingin meminimalkan ‖w‖. Untuk mencegah titik data jatuh ke dalam margin, maka harus ditambahkan batasan berikut: untuk setiap i baik

$w \cdot x_i - b \geq 1 : x_i$ kelas pertama (15)

atau

$w \cdot x_i - b \leq -1 : x_i$ kelas kedua (16)

Hal ini dapat ditulis ulang sebagai :

$$y_i (w. x_i - b) \geq 1 ,$$

untuk semua $1 \leq i \leq n$ (17)

Sehingga problem mencari hyperplane maksimum adalah masalah optimasi:
Minimalkan ‖ w ‖ dengan kendala untuk setiap i = 1,…, n

$$y_i (w. x_i - b) \geq 1 \quad (18)$$

Klasifikasi dengan *Maximum Entropy* (ME) menerapkan teori informasi. *Entropy* merupakan rata-rata himpunan informasi yang terkandung dalam suatu kumpulan kejadian X={x1,x2,…,x3} yang dapat dinyatakan dalam :

$$H(p)= \sum_{x \in X} p(x) \log_e \left(\frac{1}{p(x)} \right) \quad (19)$$

Dengan nilai H(p) adalah merupakan himpunan informasi dari kumpulan kejadian X, dan p(x) adalah probabilitas kejadian x dalam himpunan X. Metode *Maximum Entropy* (ME) adalah metode untuk memaksimalkan nilai H(p). Nilai H(p) maksimal akan diperoleh jika nilai X seragam sehingga p(x)=1/|X| dengan |X| merupakan kardinalitas dari himpunan X.

Penerapan Metode ME untuk klasifikasi dokumen dilakukan dengan pendekatan probabilitas kondisional dari suatu klas dokumen apabila dimiliki suatu dokumen. Misalkan himpunan klas dokumen adalah $A=\{a_1,a_2,\ldots,a_c\}$ dan himpuna koleksi dokumen adalah $D=\{d_1,d_2,\ldots,d_n\}$. Penentuan klas a dari dokumen d akan dilihat dengan menentukan nilai probabilitas kondisional p(a|d) yang bernilai maksimum dari distribusi probabilitas dengan entropy maksimum.

Dalam menentukan distribusi yang seragam untuk setiap pasangan $a \in A$ dan $d \in B$, pencarian ini haruslah memenuhi dari batasan-batasan yang timbul dari fakta yang ada. Fakta dari data training dapat dinyatakan dengan fungsi *feature* $f_j(a,d) \rightarrow \{1,0\}$ yang diambil dari koleksi dokumen D, dengan ketentuan :

$$f_j(a,d)=\begin{cases} 1, & jika\ f_j\ muncul\ di\ dok\ D\ pada\ kelas\ a \\ 0, & jika\ f_j\ tdk\ muncul\ di\ dok\ D\ pada\ kelas\ a \end{cases} \quad (20)$$

Proses klasifikasi dokumen termasuk dalam klas tertentu dilakukan dengan menganggap dokumen sebagai vector yang berisi kemunculan dari fitur-fitur $f_j(a,d)$ lalu mencari probabilitas klas a dari dokumen tersebut. Dokumen diputuskan masuk dalam kelas a dengan memilih nilai p(a,d) yang paling maksimum. Nilai p(a,d) yang paling besar juga merupakan nilai p(a|d) yang paling besar. Hal ini dikarenakan nilai p(d) dalam koleksi dokumen besarnya adalah tetap. Dengan demikian klas a* hasil klasifikasi adalah kelas yang memaksimumkan nilai entropy p(a,d) :

$$p^* = \arg\max_{a \in A} p(a,d) \quad (21)$$

Metode K-means *clustering* melakukan pendekatan clustering dengan mengggunakan pusat cluster sebagai

kriteria pengelompokan. Pusat kluster adalah nilai rata-rata objek seluruh anggota kluster tersebut. Misalnya kita memiliki koleksi dokumen $D=\{d_i \mid i=1,2,\ldots|D|\}=\{d_1,d_2,\ldots,d_{|D|}\}$ yang akan dikluster menjadi K buah kluster. Dalam hal ini d_i adalah vector bernilai real yang mewakili dokumen. Vektor tersebut memiliki dimensi n, yang merupakan banyaknya kata unik dalam koleksi dokumen. Koleksi dokumen dapat diwakili oleh matrik berukuran n x |D|, yaitu $[X_{ij}]$, dengan elemen x_{ij} merepresentasikan *Term Frequency* (TF) yaitu frekuensi kemunculan term (kata) ke-i dalam dokumen ke-j. Untuk mendapatkan akurasi yang lebih baik dalam komputasi, matrik yang berisi nilai frekuensi term diubah menjadi matrik berelemen real yang memperhitungkan fekuensi kemunculan dokumen yang memuat kata ke-i dengan pembobotan Invert *Document Frequency* (IDF). Selanjutnya juga diupayakan agar panjang vector dokumen adalah senantiasa 1 dengan cara melakukan normalisasi vector dokumen. Pembobotan yang menggabungkan TF dengan IDF kemudian dikenal dengan pembobotan TF-IDF yang dapat dirumuskan sebagai :

$$w_{ij}= \frac{(\ln(f_{ij})+1).\log\left(\frac{N}{n_i}\right)}{\sqrt{\left((\ln(f_{ij})+1).\log\left(\frac{N}{n_i}\right)\right)^2}} \qquad (22)$$

Algoritma K-Means *clustering* dilakukan dengan mengambil K buah vector sebagai benih (*seed*) dari pusat kluster. Selanjutnya seluruh vector dokumen dihitung jarak terhadap setiap busat kluster. Vektor dokumen yang berjarak paling dekat dengan suatu pusat kluster maka, dokumen tersebut ditetapkan menjadi anggota baru dari kluster tersebut. Demikian seterusnya diulang sampai tidak ada lagi dokumen yang berpindah kluster. Pusat kluster adalah vector rata-rata dari seluruh vector dalam suatu kluster tertentu. Pusat kluster dirumuskan sebagai persamaan (23) berikut ini:

$$m_i = \frac{1}{n_i}\sum_{j=1}^{ni} d_{ij} \qquad (23)$$

dimana m_i adalah pusat kluster ke-i dan d_{ij} adalah dokumen k-j dalam pusat kluster ke i. Kesamaan dokumen dengan pusat kluster lebih digunakan fungsi similaritas *cosine*, sebagai persamaan (24) berikut :

$$Sim(d_j,m_i)= \sum_{k=1,n} (d_{jk} * m_{ik}) \quad (24)$$

Selanjutnya algoritma K-means standard dapat dituliskan sebagai berikut :
Step:

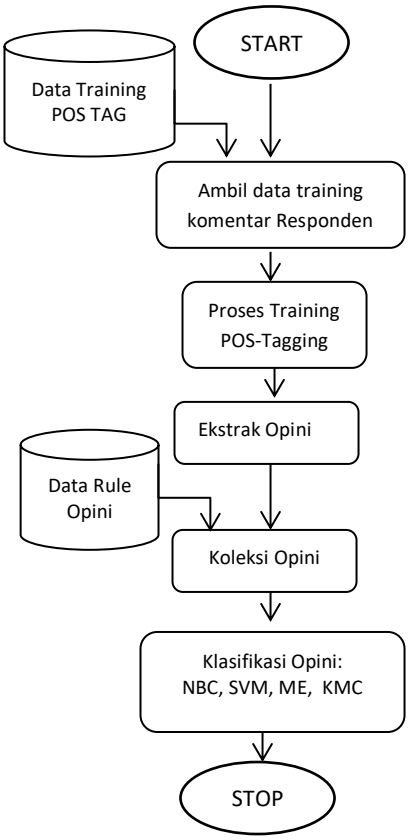
1. Ambil K objek sebagai seed dari K pusat kluster
2. Untuk semua objek: cari kluster dengan jarak terdekat, dan tetapkan objek masuk dalam kluster tersebut.
3. Hitung ulang pusat kluster dengan rata-rata objek dalam kluster tersebut
4. Hitung fungsi kriteria dan lakukan evaluasi. Jika fungsi kriteria berubah cukup kecil algoritma berhenti.

METODE

Skema langkah penyusunan prototipe adalah seperti pada Gambar 1. Data penelitian diambil dari data komentar pada kuesener mahasiswa mahasiswa IST AKPRIND Yogyakarta selama 4 semester, yaitu Semester 1 Tahun 2013/2014 sebanyak 3.801 komentar, Semester 2 Tahun 2012/2013 sebanyak 2.551 komentar, Semester 1 Tahun 2012/2013 sebanyak 3.663 komentar dan Semester 2 Tahun 2011/2012 sebanyak 1883 komentar. Data dipilih sebanyak 1000 komentar sebagai data training dan 1000 komentar sebagai data uji.

Untuk data pelatihan POS-Tagger dibuat format seperti pada Gambar 2. Pemberian POS-TAG dilakukan secara manual pada komentar menggunakan daftar POS TAG dari Tabel 1. Data training diperlukan untuk menentukan parameter parameter dari POS-Tagger yang selanjutnya parameter ini akan diuji menggunakan data test. Data POS TAG ini akan dijadikan dasar dalam pelatihan dalam menentukan nilai probabilitas suatu POS TAG jika diberikan suatu rangkaian kata dalam kalimat. Nilai

probabilitas ini digunakan dalam algoritma untuk menentukan POS TAG yang mana harus diberikan pada suatu kata berdasarkan POS TAG kata sbelumnya.



Gambar 1 . Skema Perancangan

<DOC-0001>Tolong/VBI gedungnya/NNG dibersihkan/VBP</DOC>
<DOC-0002>Birokrasi/NNU kampus/NNC yang/PR jelek/JJ membuat/VBT mahasiswa/NNC merasa/VBT dikhianati/VBP</DOC>
<DOC-0003>Kursi/NNU di/IN dalam/IN ruangan/NNU kurang/RB mengenakkan/JJ</DC>
<DOC-0004>lebih/RB ditingkatkan/VBP lagi/RB</DOC>
....
<DOC0999>LCD/NNC tolong/VBI diperbaiki/VBP </DOC>

Gambar 2. Format dokumen POS-Tag

TABEL 1. KOLEKSI POS TAG

No	POS	POS Name	Contoh
1	OP	Open Parenthesis	({[
2	CP	Close Parenthesis	)}]
3	GM	Slash	/
4	;	Semicolon	;
5	:	Colon	:
6	“	Quotation	" and '
7	.	Sentence terminator	.
8	,	Comma	,
9	-	Dash	-
10	…	Ellipses	…
11	JJ	Adjective	Baik, Bagus
12	RB	Adverb	Sekali, sangat
13	NNC	Countable Noun	Kursi, Kulkas
14	**NNU**	**Uncountable Noun**	**Gula, hujan**
15	NNP	Proper Noun	Toyota, Sony
16	NNG	Genetive Noun	Motornya
17	VBI	Intransitive Verb	Pergi
18	VBT	Transitive Verb	Membeli
19	**VBP**	**Passive Verb**	**ditingkatkan, diperbaiki**
20	IN	Preposition	Di, Dari, Ke
21	MD	Modal	Bisa
22	CC	Coor-Conjunction	Dan, Atau, tetapi
23	SC	Subor-Conjunction	Jika, Ketika
24	DT	Determiner	Para, Ini, Itu
25	UH	Interjection	Wah, Aduh, Oi
26	CDO	Ordinal Numerals	Pertama, Kedua, Ketiga
27	CDC	Collective Numerals	Berdua
28	CDP	Primary Numerals	Satu, Dua, Tiga
29	CDI	Irregular Numerals	Beberapa
30	PRP	Personal Pronouns	Saya, Mereka
31	WP	WH-Pronouns	Apa, Siapa, Dimana
32	PRN	Number Pronouns	Kedua-duanya
33	PRL	Locative Pronouns	Sini, Situ
34	NEG	Negation	Bukan, Tidak
35	SYM	Symbols	#,%,^,&,*
36	RP	Particles	Pun, Kah
37	FW	Foreigns Words	Word

Jika tahapan pelatihan POS-Tagger selesai maka tahap berikutnya adalah pemberian POS TAG pada data uji POS Tag. Dari data uji POS TAG didapatkan koleksi komentar yang telah diberikan POS TAG.

TABEL 2. RULE UNTUK DETEKSI OPINI

No	Rule	Examples
1	RB JJ	sangat buruk, dengan bagus , memang jelek
2	RB VB	semoga berjalan, jika memilih
3	NN JJ	LCDnya jelek, alatnya bagus
4	NN VB	Ngajarnya membosankan, perkataannya menjengkelkan
5	JJ VB	mudah difahami, cepat memahami
6	CK JJ	bagus atau baik, tetapi malas
7	JJ BB	sama bagus
8	VB VB	membuat pusing, membikin bosan
9	JJ RB	indah sekali, bagus sekali
10	VB JJ	membikin bingung
11	NEG JJ	tidak seindah, tidak semudah
12	NEG VB	tidak mengerti, tidak memahami, bukan mengajar
13	PRP VBI	saya menyukai, kita suka
14	PRP VBT	kita suka
15	VBT NN	memiliki kedekatan, memiliki kepekaan
16	MD VBT	Perlu mengambil referensi
17	MD VBI	Perlu dikembangkan
18	**UH VBP**	**Tolong dicat, tolong diperbaiki**
19	**JJ VBP**	**Mudah diterima, sulit dipahami**

TABEL 3. RULE UNTUK DETEKSI TARGET

No	Rule	Examples
1	NN	ac, lcd, internet
2	NNG	laboratoriumnya, lcdnya
3	**NNP**	**pak joko, bu yuli, Pengok**
4	NN NN	kantin kampus, ac pengok
5	NN CC NN	kampus dan lab
6	NN IN NN	ac di klas

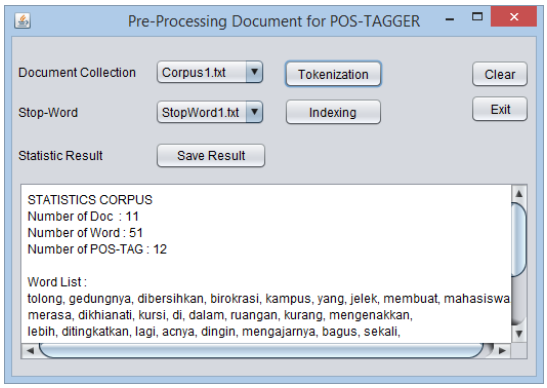
Jika suatu koleksi kalimat telah diberikan POS TAG pada setiap kata penyusun kalimat, maka untuk menentukan apakah suatu kalimat itu opini atau bukan digunakan suatu RULE seperti pada Tabel 2 dan Tabel 3. Suatu kalimat ditetapkan sebagai opini apabila pola seperti pada Tabel 2 muncul dalam kalimat tersebut. Sedangkan untuk menetapkan target dari opini digunakan RULE pada Tabel 3.

Penentuan kinerja dalam HMM POS Tagger dan kinerja klasifikasi digunakan parameter akurasi. Untuk kinerja HMM POS-Tagger akurasi dimaknai sebagai rasio banyaknya kalimat yang diberikan POS TAG dengan benar dibandingkan dengan total banyaknya kalimat yang diuji. Untuk proses klasifikasi akurasi dihitung dengan menghitung rasio banyaknya opini yang diklasifikasi dengan benar dibandingkan dengan total banyaknya opini yang diklasifikasi. Rumus akurasi adalah seperti persamaan (16) berikut.

$$Akurasi = \frac{c(relevan\_document\_disovered)}{c(document\_discovered)} \quad (16)$$

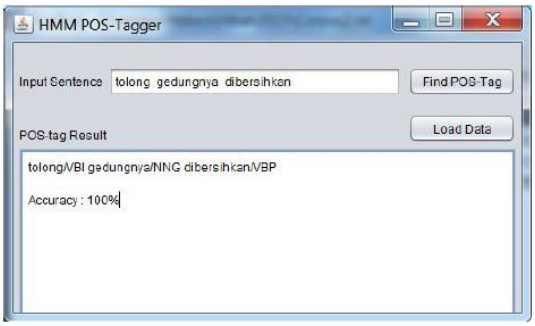
PEMBAHASAN

Sebelum proses pelatihan dilakukan terlebih dahulu dilakukan tahap pre-*processing*, yaitu langkah untuk menghitung banyaknya kata dalam kalimat, banyaknya POS TAG untuk setiap kata. Contoh antar muka untuk preproessing POS Tagger adalah seperti gambar 3.



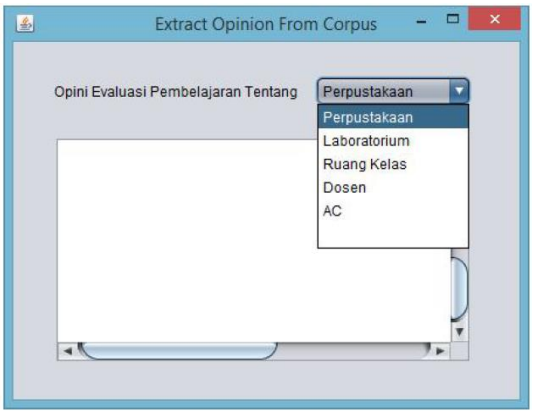
Gambar 3. Format dokumen POS-Tag

Proses pelatihan dilakukan untuk mendapatkan pola berdasarkan statistik yang dihitung dari data latih sebanyak 1000 komentar. Dari pelatihan akan didapatkan statistik kemunculan masing-masing POS-TAG pada setiap kata. Contoh pemberian POS TAG pada kalimat yang diinputkan adalah seperti pada Gambar 4. Dengan statistik tersebut POS-Tagger dapat menentukan besarnya nilai probabilitas penempatan POS-TAG pada data uji menggunakan persamaan (4),(5) dan (6) untuk mendapatkan *tagging* pada suatu kalimat komentar.

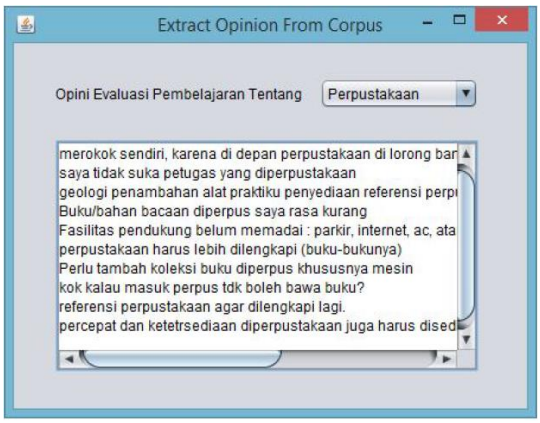


Gambar 4. HMM POS-Tagger

Kalimat yang sudah diberikan POS-TAG kemudian dapat ditentukan apakah termasuk opini atau bukan berdasarkan rule pada Tabel 2. Lebih jauhnya dapat juga dilacak objek opini berdasarkan rule pada Tabel 3. Sebagai gambaran penerapan ekstrak opini berdasarkan RULE dan ekstrak objek opini adalah antar muka seperti Gambar 5 dan Gambar 6 berikut.



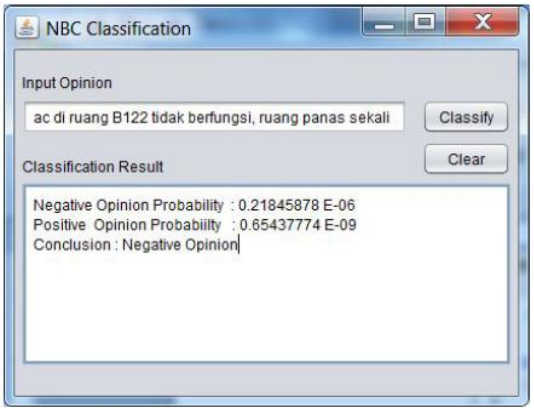
Gambar 5. Ekstrak opini berdasar objek



Gambar 6. Hasil opini perpustakaan

Setelah koleksi opini diperoleh, selanjutnya adalah melakukan klasifikasi opini apakah termasuk opini positif atau negatif. Program *classifier* yang digunakan menggunakan 4 algoritma, yaitu *Naïve Bayes Classifier*, *(NBC), Support Vector Machine (SVM), Maximum Entrophy(ME) dan K-Means Clustering (KMC).*

Contoh hasil implementasi klasifikasi opini seperti pada Gambar 7 yaotu klasifikasi opini dengan NBC. Opini "**ac di ruang b22 tidak berfungsi, ruangan panas sekali**" didapatkan hasil opini negatif.



Gambar 7. NBC Classification

Rangkuman hasil ekstraksi opini menggunakan data uji memiliki akurasi 0.95. Untuk klasifikasi opini pengujian dengan empat metode klasifikasi menghasilkan akurasi seperti pada tabel 4. Rata-rata akurasi dari 4 metode klasifikasi adalah 0.85.

TABEL 4. AKURASI KLASIFIKASI

	Metode Klasifikasi			
	NBC	*SVM*	*ME*	*KMC*
Akurasi	0.84	0.83	0.84	0.88
Rata-rata	0.85			

KESIMPULAN

Dari perancangan dan pengujian prototipe dapat ditarik beberapa kesimpulan antara lain sebagai berikut : 1.Penggunaan HMM POS-Tagger untuk memberikan POS-TAG pada kata-kata

penyusun opini telah berhasil dengan baik. Akurasi yang diperoleh pada pengujian data tes adalah 0.95. 2. Hasil POS TAG telah dapat digunakan untuk mengekstrak opini dari koleksi komentar yang terdiri dari opini dan bukan opini. 3. Klasifikasi opini menjadi positif atau negatif menggunakan 4 metde klasifikasi telah berhasil dengan baik. 4. Rata-rata akurasi dari klasifikasi opini adalah sebesar 0.85

DAFTAR PUSTAKA

Boser,B.E.,Guyon, I.M. and Vapnik,V.N.,1992, "A Training Algorithm for Optimal Margin Classifiers",Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory,1992,pp. 1171-1183.

Cortes,C. and Vapnik,V., 1995,"Support-Vector Networks", Machine Learning, 20, pp.273-297

Cutting, D., Kupiec,J., Pesderson,J. and Sibun,P., "A Practical Part-ofspeech Tagger, Xerox Palo Alto Research Center", in Proceeding of the third Conference on applied Natural Language Processing , 1992, pp.133-140.

Greaves, F., D.R. Cano, C. Millet, A.Darzi, and L. Donaldson, "Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments", Journal of Medical Internet Research, 2013, 15:11, e239. Online publication date: 1-Jan-2013

Grimes, S., 2013, *Unstructured Data and the 80 Precent Rule*, Clarabridge Bridgepoints.

Liu,B.,2010, "Sentiment Analysis: Multi Facet Problem", IEEE Intelligence System, 25 (3),pp:76-80

Loia, L. and Senatore, S.,"A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content", Knowledge-Based Systems 58, 2014, pp. 75-85 Online publication date: 1-Mar-2014

McCallum, A. and Nigam, K., 1998, "A Comparison of Event Models for Naive Bayes Text Classification", AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48

Pang, B., Lee, L. and Vaithyanathan, S., 2002, Thumbs up?: sentiment classification using machine learning techniques", Proceeding EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, pp: 79-86

Putten,P.V.D., Kok, J. and Gupta,A., 2002, *Why the Information Explosion can be Bad for Data Mining, and How Data Fusion Provides a Way Out*, Proc.of the 2nd SIAM International Conference on Data Mining,pp:11-13

Pisceldo. F., Manurung, R. and Adriani,M. "Probabilistic Part-of-Speech Tagging for Bahasa Indonesia", Third International MALINDO Workshop, colocated event ACL-IJCNLP, Singapore, 2009

Shelke,N.M., Deshpande,S. and Thakre,V., "Survey of Techniques for Opinion Mining," International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012

Wicaksono, A.F. and Purwarianti,A., 2010, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia", Proceedings of the 4th International MALINDO Workshop, Jakarta.

Widhiyanti, K and A. Harjoko, "POS Tagging for Bahasa Indonesia dengan HMM dan Rule Based", INFORMATIKA Vol.8.,No.2., November 2012, pp.151-167