PERINGKASAN PADA REVIEW PRODUK MENGGUNAKAN METODE CRF DAN KNN
SUMMARIZATION ON PRODUCT REVIEW USING CRF AND KNN METHOD

Pursita Kania Praisar[1], Warih Maharani[2], Nungki Selviandro[3]

[1,2,3]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
[1]pursitakania@yahoo.com,[2]wmaharani@gmail.com,[3]selviandro@gmail.com

Abstrak

Dalam perkembangan internet pada saat ini, banyak orang yang memanfaatkannya dengan membuat sebuah tampilan website online berisi informasi yang dibutuhkan oleh para konsumen. Salah satu diantaranya adalah website yang berisi tentang konten-konten belanja secara online. Selain memudahkan dalam berbelanja, pada website belanja online juga sering ditemukan fitur review product atau tanggapan dari barang yang dijual pada website belanja online tersebut.

Tanggapan mengenai sebuah barang pada satu website online sering dijadikan sebagian acuan terhadap sebuah kualitas barang. Banyaknya tanggapan terhadap sebuah barang dalam satu website online, menjadikan kesulitan tersendiri untuk menyimpulkan hasil dari tanggapan barang tersebut. Maka dari itu untuk mempermudah dalam menyimpulkan hasil tanggapan dari sebuah barang, perlu dibuat sebuah sistem peringkasan untuk menganalisa hasil tanggapan dari konsumen dalam suatu website belanja online.

Sistem peringkasan ini dilakukan menggunakan metode CRF (Conditional Random Fields) untuk ekstraksi aspeknya dan K-NN untuk klasifikasinya. Parameter yang dibutuhkan pada sistem ini adalah persentase data training, penggunaan lemmatization pada preprocessing, nilai standar deviasi pada ekstraksi aspek, nilai learning rate pada ekstraksi aspek, threshold pada ekstraksi opini, dan nilai k pada klasifikasi.

Penelitian ini dilakukan untuk mencari nilai efektif parameter inputan. Hasil yang didapatkan dari penelitian ini adalah nilai efektif presentase data training adalah 70%, digunakannya tahap lemmatization pada preprocessing, nilai efektif standar deviasi adalah 1.75, nilai efektif learning rate adalah 0.01, nilai efektif threshold adalah 0.5 dan nilai efektif k adalah >7.

Kata kunci: Summarization,CRF, K-NN

Abstract

In the internet development at the moment, many people take advantage of it to make an online website which provides the information needed by the customer. One of them is a website which has content for online shopping. Besides making the shopping easy, the online shopping websites are also carrying common features, such as product reviews or feedback from the goods sold on the online shopping website.

The response to an item on online shopping websites is often used as a reference for goods quality. The number of responses to an item in an online shopping website makes the process of summarizing the results of the responses of the goods become difficult. Therefore, to make it easier concluding the response to an item, we need to make a summarization system to analyses the results of customer feedback in an online shopping website.

The summarization system is conducted using the method of CRF (Conditional Random Fields) for extract its aspects and K-NN for classification. Parameter required in this system is the percentage of training data, the use of lemmatization in the preprocessing, the standard deviation value on the extraction aspect, the value of learning rate on the extraction aspect, the threshold on extract opinions, and the value of k in the classification.

This research was conducted to find effective value input parameter. The results obtained from this research is the effective value of the percentage of the training data by 70%, the use of lemmatization in the preprocessing stage, the effective value of the standard deviation is 1.75, the effective value of learning rate is 0.01, the effective threshold value is 0.5 and the effective k value is> 7.

Keywords: Summarization, CRF, K-NN

1. Pendahuluan

Perkembangan teknologi informasi dan internet banyak dimanfaatkan oleh perusahaan dan perorangan untuk menjalankan beberapa bisnis secara *online*. Salah satu diantaranya adalah dengan membuat sebuah halaman *website* belanja *online*. Selain memudahkan dalam berbelanja, pada *website* belanja *online* juga sering ditemukan fitur *review product* atau tanggapan dari barang yang dijual pada *website* belanja *online*.

Tanggapan mengenai sebuah barang pada satu *website online* sering dijadikan sebagian acuan terhadap sebuah kualitas barang. Bahkan menurut sumber dari *Invesp More Conversions*, 90% dari pengguna *wesite online* membaca tanggapan dari sebuah produk, dan 80% dari mereka mempercayai hasil dari tanggapan terhadap suatu produk **[1]**. Namun demikian, banyaknya tanggapan terhadap sebuah barang dalam satu *website online*, menjadikan kesulitan tersendiri untuk menyimpulkan hasil dari tanggapan barang tersebut.

Maka dari itu untuk mempermudah dalam menyimpulkan hasil tanggapan dari sebuah barang pada satu *website online*, perlu dibuat sebuah *system* untuk menganalisa hasil tanggapan dari konsumen yang telah dikirim. Sistem akan menganalisa aspek yang terdapat pada kalimat dan menentukan apakah termasuk kedalam karakter positive atau negative. Dengan demikian hasil yang didapat bisa dijadikan kesimpulan terhadap tanggapan sebuah barang dalam suatu *website online*.

Sistem peringkasan ini dibangun menggunakan metode CRF (*Conditional Random Fields*) untuk ekstraksi aspeknya dan K-NN untuk klasifikasinya. Metode CRF dan KNN dipilih karena metode ini dapat digunakan untuk mengolah data berupa text dan menghasilkan hasil yang baik.[4,10] Parameter yang dibutuhkan pada sistem ini adalah persentase data training, penggunaan *lemmatization* pada *preprocessing*, nilai standar deviasi pada ekstraksi aspek, nilai *learning rate* pada ekstraksi aspek, *threshold* pada ekstraksi opini, dan nilai k pada klasifikasi. Penelitian ini dilakukan untuk mencari nilai efektif parameter inputan.

2. Landasan Teori

2.1 *Text Mining*

Text mining adalah pencarian pola yang menarik dari sekumpulan data teks. *Text mining* digunakan untuk menyelesaikan permasalahan dari terlalu banyaknya informasi dengan menggabungkan beberapa teknik seperti *data mining, machine learning, natural language processing, information retrieval*, dan *knowledge management*. Proses dalam *text mining* terdiri dari *preprocessing* pada dokumen, proses penyusunan *intermediate representation*, teknik untuk menganalisa *intermediate representation* dan visualisasi hasil pengolahan. [2]

Text biasanya berupa dokumen yang tidak terstruktur. Maka pada proses *text mining*, dokumen diolah sehingga menjadi dokumen yang terstruktur. [3]

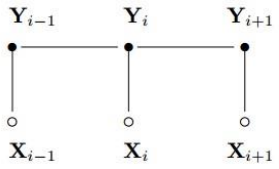
2.2 *Lemmatization*

Lemmatization adalah perubahan kata menjadi kata dasar atau betuk *lemma*-nya [6] Pada dokumen, biasanya terdapat kata-kata yang sudah terjadi perubahan seperti ditambahnya imbuhan seperti –ing dan -s. Sehingga pada proses ini, kata-kata tersebut dikembalikan lagi menjadi kata dasarnya. Proses *lemma* terjadi menggunaakan kamus dan analisis dalam melakukan perubahan pada kata [7].

2.3 *Conditional Random Field*

Conditional Random Fields (CRF) adalah model probabilitas yang digunakan untuk melakukan pelabelan pada data yang berurutan [10]. *Input* dari CRF berasal dari data *training* yang dinotasikan dengan dan *output*nya adalah probabilitas dari label yang mungkin dengan notasi .

CRF memiliki kelebihan daripada metode probabilitas lainnya seperti *Hidden Markov Model* (HMM) karena CRF dapat menentukan banyaknya *feature function* berdasarkan seluruh isi data trainingnya. Berbeda dengan HMM yang bergantung hanya pada label pada kata sebelumnya.



Gambar 1 *Linear Chain Conditional Random Fields*

Persamaan probabilitas CRF y terhadap x adalah sebagai berikut.

$$| \quad — \Pi$$

$_c(y_{c,}, x)$ adalah fungsi potensial. Z(x) adalah fungsi normalisasi dengan persamaan sebagai berikut.

$$\Sigma \Pi$$

2.4 Nearest Opinion Word

Nearest Opinion Word adalah proses ekstraksi opini untuk menentukan opini pada aspek dengan cara memasangkan opini dengan aspek pada jarak terdekat **[11]**. Cara kerja dari metode neares opinion word ini dengan mengukur jarak (dist(f,w)) antara setiap kata yang memiliki postag "JJ"dan aspek. Lalu hitung nilai invers dist(f,w) dengan persamaan sebagai berikut.

$$\_\_\_\_\_$$

Untuk mengambil opini suatu aspek, maka ditentukan nilai threshold sebagai batas nilai minimum rel(f,w). Maka kata yang memiliki nili rel(f,w) lebih besar sama dengan nilai threshold, akan termasuk sebagai opini pada aspek tersebut.

2.5 K-NN

K-Nearest Neighbor (KNN) merupakan suatu metode yang dilakukan dengan cara mencari kelompok k objek dalam data *training* yang paling dekat atau mirip dengan objek pada data baru atau data testing [12]. *Algoritma K-Nearest Neighbor* adalah sebuah metode yang dilakukan untuk klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dan kasus lama yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada [13].

Untuk mendefinisikan jarak antara dua titik yaitu titik pada data *training* (x) dan titik pada data testing (y) maka digunakan rumus *cos similarity* seperti yang ditunjukkan pada persamaan

$$\_\_\_ \quad \frac{\Sigma}{\sqrt{\Sigma} \quad \sqrt{\Sigma}}$$

Pada *fase training*, *algoritma* tersebut hanya menyimpan *vektor-vektor fitur* dan klasifikasi data *training sample*. Pada *fase* klasifikasi, *fitur-fitur* yang sama akan dihitung untuk *testing data* (yang klasifikasinya tidak diketahui). Jarak dari *vektor* baru yang ini terhadap seluruh vektor *training sample* dihitung dan sejumlah (k) buah yang paling dekat diambil [14].

2.6 Evaluasi

Dalam melakukan sebuah penelitian, sebagai acuan pengukuran terhadap metode yang dipakai adalah dengan mengukur performansi dari metode tersebut. Metode pengukuran yang digunakan dalam penelitian peneitian ini diantaranya adalah *Precision*, *Recall*, dan *F-Score*.

Precision adalah presentase dari nilai aspek yang terbukti benar dengan nilai aspek yang terdeteksi. Nilai precision digunakan untuk mengukur seberapa tepat sistem melakukan predisi[9].

Recall adalah presentadi dari nilai aspek yang terbukti benar dengan nilai aspek yang sebenarnya. *Recall* digunakan untuk mengukur seberapa banyak aspek yang memang diprediksi benar[9].

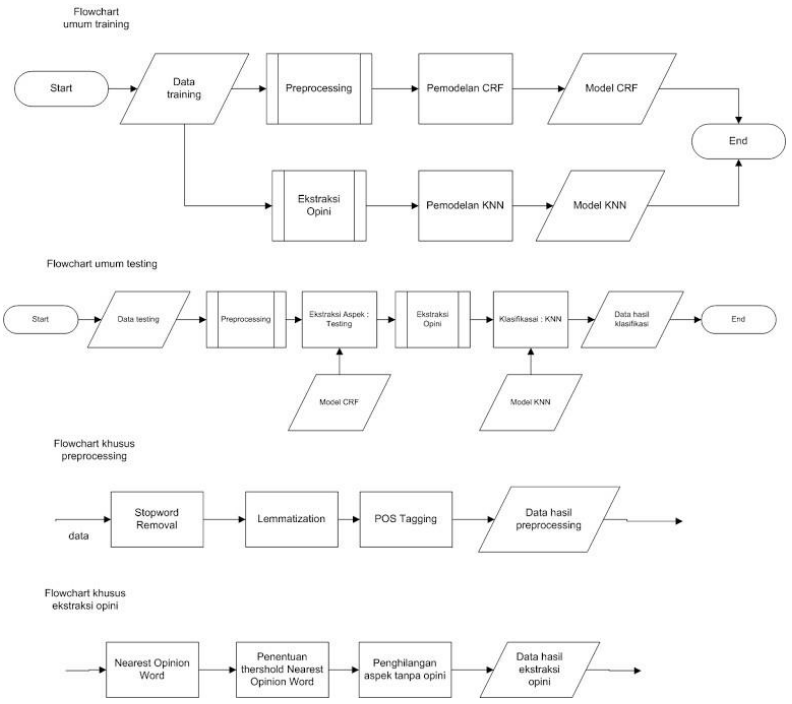
F-score adalah standar pengukuran yang digunakan untuk mengukur kombinasi nilai *precision* dan *recall*[9].

Perhitungan evaluasi ekstraksi fitur dilakukan pada level dokumen. Yang dimaksud level dokumen adalah penghitungan *precision* dan *recall* dilakukan secara langsung menggunakan keseluruhan secara sekaligus. Perhitungan dilakukan dengan rumus berikut:

3. Pembahasan

3.1 Gambaran Umum Sistem

Sistem peringkasan yang menjadi objek penelitian p ini adalah sistem peringkasan untuk menganalisa *review* produk dari konsumen. Tahapan pembentukan sistem ini terbagi menjadi yaitu (1) *preprocessing* data (2) ekstraksi aspek dan opini (3) penentuan sentimen pada fitur menggunakan klasifikasi (4) peringkasan hasil klasifikasi.



Gambar 1 Gambaran Umum Sistem

3.2 Ekstraksi Aspek menggunakan CRF

Data training dilakukan untuk mendapatkan model CRF yang akan digunakan pada proses *testing*. Model CRF tersebut adalah nilai optimal parameter *feature function* yang dilabelkan sebagai λ. Sehingga banyaknya nilai λ sama dengan banyaknya *feature function*.Nilai λ akan terus berubah-ubah pada setiap, tergantung pada fitur dan label yang terandung pada sebuah baris. Perulangan perubahan nilai λ dilakukan sebanyak jumlah baris pada suatu data. Pada iterasi pertama, seluruh nilai λ bernilai 0.

Langkah pertama yang dilakukan untuk pembentukan model CRF adalah mencari *node potensial* dan *edge potensial*. Persamaan *node potensial* dan *edge potensial* didapat dari persamaan CRF. Persamaan untuk mencari *node potensial* dan *edge potensial* adalah sebagai berikut.

Persamaan *node potensial* :

$$(\Sigma \qquad)$$

Persamaan *edge potensial* :

$$(\Sigma \qquad)$$

= *feature function node*

) = *feature function edge*

Langkah selanjutnya adalah dengan menghitung *forward-backward* pass yang berasal dari metode *Stochastic Gradient*. Metode ini dipakai untuk mengolah data *sekuens*. Persamaan *forward pass* adalah sebagai berikut.

$$[\ ] \quad \Sigma \quad [\ \ ]$$

dimana > 0 adalah faktor skala yang ditentukan sehingga Σ [ ] .

Dan persamaan *backward pass* adalah sebagai berikut

$$[\ ] \quad \Sigma \quad [\ \ ]$$

dimana > 0 adalah faktor skala yang ditentukan sehingga Σ [ ] .

Setelah setiap data x memiliki nilai *forward pass* dan *backward pass*, maka dapat dilakukan perhitungan probabilitas lokal *node* untuk setiap label yang mungkin pada data x dan perhitungan probabilitas lokal *edge*. Persamaan probabilitas lokal *node* dan probabilitas lokal *edge* adalah sebagai berikut.

Probabilitas lokal *node*

$$| \qquad [\ ] \qquad\qquad [\ ]$$

dimana adalah faktor normalisasi agar Σ |

Probabilitas lokal *edge*

$$| \qquad [\ ] \qquad\qquad\qquad [\ \ ]$$

Dimana adalah faktor normalisasi agar Σ Σ |

Setelah mendapatkan nilai probabilitas lokal *node* dan probabilitas lokal *edge*, maka dihitung nilai untuk setiap *feature function node* dan *feature function ege*. Pesamaan nilai didapat dari turunan *loglikelihood*. Persamaan untuk menghitung nilai memiliki persamaan sebagai berikut.

Untuk *Feature function node*

$$\Sigma \qquad \Sigma \quad | \qquad \_\_$$

[ ]

Untuk feature function edge

$$\Sigma \qquad\qquad \Sigma \qquad |$$

[ ] __

Dimana variabel adalah standar deviasi dari Distribusi *Gauss* yang nilainya ditetapkan berdasarkan *trial and error*. Pada penelitian-penelitian pelabelan, nilai yang digunakan lebih besar dari 0 [8].

Langkah terakhir dalam proses *training* adalah memperbaharui nilai untuk setiap *feature function node* dan *feature function edge*. Persamaan nilai adalah sebagai berikut.

dimana adalah *learning rate* dengan [5].

Nilai adalah model dari CRF yang nantinya akan digunakan pada proses *testing*. Input dari proses *testing* adalah data *testing*, *feature function* dan nilai yang sebelumnya sudah diapatkan pada proses *training*.

Tahapan pertama yang dilakukan adalah mencari *node potensial* dan *edge potensial* dengan persamaan sebagai berikut.

Persamaan *node potensial* :

( )

Persamaan *edge potensial* :

( )

= *feature function node*
) = *feature function edge*

Kemudian hitung *Maximal Forward Pass* dengan persamaan:

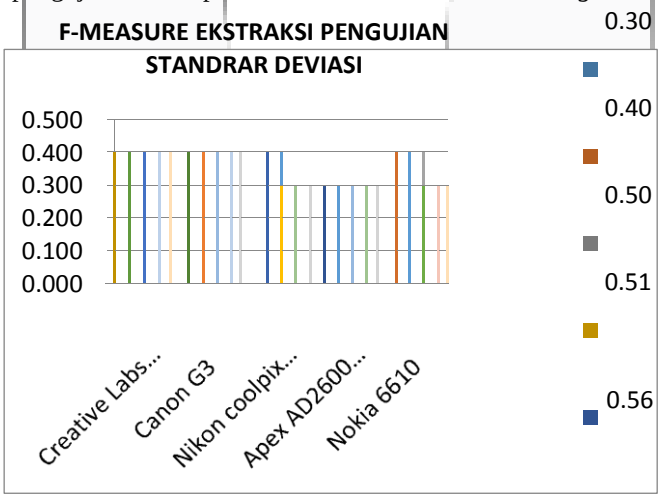
[ ] [ ]

Setelah itu dilakukan *Backtracking* yang bertujuan untuk menelusuri kembali dan meninjau label optimal pada tiap data dengan persamaan :

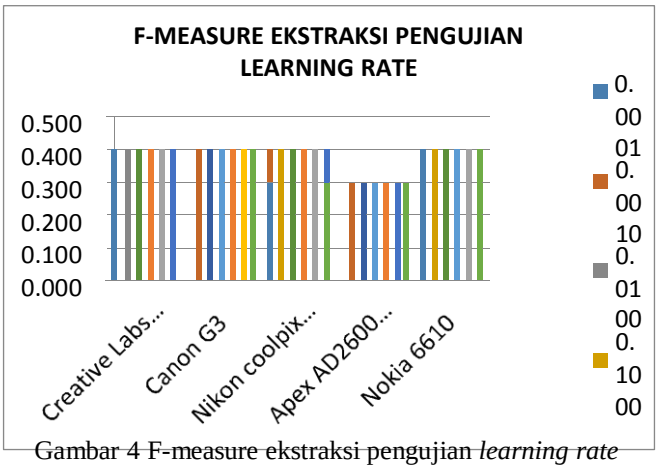
[ ]

Tahapan terakhir adalah membandingakan nilai akhir backtracking untuk label A dan NA. Jika nilai backtracking label A lebih besar, maka kata tersebut diberi label A(aspek). Begitu juga bilai nilai backtracking label NA lebih besar, maka kata tersebut diberi label NA(aspek).

Didapatkan hasil pengujian terhadap nilai standard deviasi dan *learning rate* sebagai berikut.



Gambar 3 F-measure ekstraksi pengujian standar deviasi



Gambar 4 F-measure ekstraksi pengujian *learning rate*

Dalam penggunaan metode CRF ini,berdasarkan penelitian maka didapatkan nilai optimum standar defiasi yaitu 1.75 dengan f-measure maksimum ekstraksi sebesar 0.438 pada data Creative Labs Nomad Jukebox Zen Xtra 40GB dan nilai optimum learning rate adalah 0.01 dengan f-measur maksismum ekstraksi sebesar 0.458 pada data Canon G3

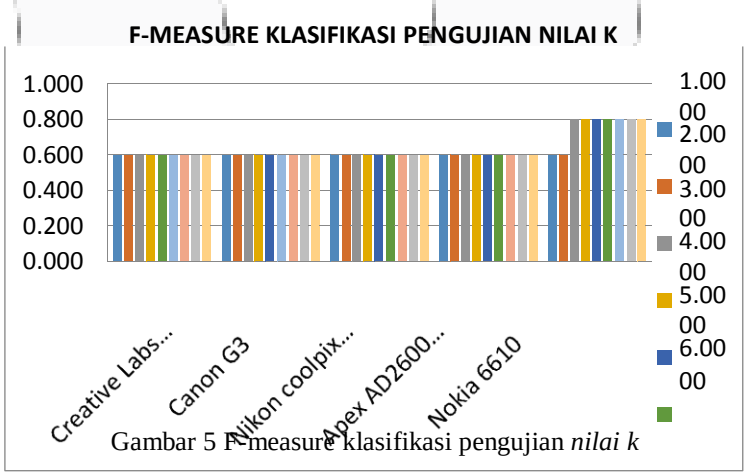
3.3 Klasifikasi menggunakan KNN

Tahapan K-NN dilakukan menggunakan hasil dari ekstraksi opini yang sudah dilakukan sebelumnya. Proses K-NN dimulai dengan membuat tabel tf dan idf berdasarkan data yang dimiliki. Setelah selesai membuat tf-idf, dilakukan penghitungan nilai bobot dimana merupakan hasil dari perkalian antara tf dan idf yang merupakan model KNN. Semua angka tersebut akan membantu dalam perhitungan *cosine similarity* dengan rumus sebagai berikut:

$$\frac{\Sigma}{\sqrt{\Sigma}\quad\sqrt{\Sigma}}$$

Dengan rumus *cosine similarity* tersebut, kita dapat melihat kemiripan pada dokumen testing dengan dokumen *training*. Hal selanjutnya yang dilakukan adalah diambil pengkategorian data positif dan negatif. Yang diambil adalah nilai terbesar/terkecil sebanyak K. Setelah itu nilai K tersebut dijumlahkan, dan dibandingkan nilai keduanya. Yang terbesar nilai K-nya menunjukan polaritas yg menjadi nilai akhir.

Didapatkan hasil pengujian k sebagai berikut.



Gambar 5 F-measure klasifikasi pengujian *nilai k*

Dalam penggunaan metode KNN ini,berdasarkan penelitian maka didapatkan nilai optimum k yaitu >7 dengan f-measure maksimum klasifikasi sebesar 0.862 pada data Nokia 6610.

4. Kesimpulan

Berdasarkan pengujian yang dilakukan dalam penelitian ini, dapat disimpulkan bahwa:

1. Parameter standar deviasi yang efektif pada penelitian ini sebesar 1.75 dan nilai learning rate yang paling efektif pada penelitian ini adalah 0.01.
2. Parameter nilai K yang paling efektif pada penelitian ini adalah >7

5. Saran

Saran yang diperlukan untuk pengembangan penilitian ini adalah :

1. Penambahan penggunaan leksikal pada ekstraksi aspek
2. Melakukan pengembangan proses klasifikasi agar dapat menentukan *polarity score*
3. Mampu mengidentifikasi aspek implisit

6. Daftar Pustaka

[1] Khalid Sale, "The Importance Of Online Customer Reviews [Infographic]" 31 Agustus 2016. [Online]. Available : http://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

[2] B. Susanto, "lecturer.ukdw.ac.id," 5 Februari 2015. [Online]. Available: http://lecturer.ukdw.ac.id/budsus/pdf/textwebmining/TextMining_Kuliah.pdf.

[3] S.M Weiss,, N. Indrukhya and T.Zang, "Fundamentals of Predictive Text Mining," Springer.

[4] Bruno Trstenjak, Sasa Mikac, Dzenana Donko "KNN with TF-IDF Based Framework for Text Categorization" 2013.

[5] Universitas Sumatera Utara, 01 September 2016. [Online]. Available: http://repository.usu.ac.id/bitstream/123456789/33585/3/Chapter%20II.pdf

[6] S. Nirenburg, Ed., *Language Engineering for Lesser-studied Languages*. IOS Press, 2009, p. 31.

[7] (2016, Mar.) Standford Natural Language Processing. [Online]. http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[8] Jianxin Wu, "Some Properties of the Normal Distribution" 01 September 2016. [Online]. http://cs.nju.edu.cn/wujx/paper/Gaussian.pdf

[9] Cyril Goutte and Eric Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation", Meylan, France.

[10] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[11] K. T. Chan, "Improving Opinion Mining with Feature-Opinion Association and Human Computation," p. 31, 2009.

[12] Chang, C, Wu, Y., Hou, S. (2009) *Preparation and Characterization of Superparamagnetic Nanocomposites of Aluminosilicate/Silica/Magnetite*, Coll. Surf. A336: 159,166.

[13] Kusrini, Emha T. Luthfi, 2009, *Algoritma Data Mining*. Andi, Yogyakarta

[14] Jatisi, Vol. 1 No. 1 September 2014 : http://www.mdp.ac.id/jatisi/vol-1-no-1/JATISI_Vol_1_No_1_September_2014_1.pdf