PERINGKASAN REVIEW PRODUK BERBASIS FITUR MENGGUNAKAN SEMANTIC SIMILARITY SCORING DAN SENTENCE CLUSTERING

SUMMARIZING PRODUCT REVIEW BASED ON FEATURE USING SEMANTIC SIMILARITY SCORING AND SENTENCE CLUSTERING

Yanuar Ega Ariska[1], Warih Maharani[2], M. Syahrul Mubarok[3]

[1,2,3]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom

[1]yanuaregaariska@gmail.com,[2]wmaharani@gmail.com,[3]msyahrulmubarok@gmail.com

Abstrak

Ulasan produk dari konsumen merupakan salah satu faktor yang penting dalam penjualan suatu produk. Menganalisis klasifikasi sentimen dan peringkasan suatu *review* produk memiliki tiga tahap yang harus dilakukan. Tahap pertama yaitu ekstraksi fitur menggunakan *frequent itemset mining* dengan algoritma apriori. Kemudian, dilakukan klasifikasi opini menggunakan SentiWordnet untuk penentuan polaritas kata opini. Tahap terkahir yaitu dilakukan peringkasan menggunakan semantic similarity scoring dan sentence clustering. Hasil dari penelitian ini didapat bahwa *filtering* kata yang sesuai juga mempengaruhi performansi dari ekstraksi pada penelitian ini. *Filtering* kata digunakan yaitu *Noun*, *Noun Phrase*, irisan serta gabungan keduanya, keempat *filtering* kata yang digunakan memiliki hasil yang cukup seimbang, gabungan dan irisan yang diharapkan dapat meningkatkan performansi juga masih didapat hasil yang tidak terlalu jauh dengan hanya *Noun* dan *Noun Phrase*. Hasil performansi ekstraksi pada penelitian ini adalah sekitar 20-40% pada dataset yang digunakan. Klasifikasi menggunakan SentiWordNet menunjukkan hasil performansi yang cukup baik namun pada beberapa dataset yang memiliki kompleksitas kalimat yang cukup tinggi juga terjadi penurunan walaupun tidak terlalu berbeda jauh dan masih pada sekitaran 40-90%. Peringkasan dokumen dapat dilakukan dengan baik pada dataset yang disediakan karena dataset memiliki jumlah kalimat ulasan produk yang memadai dan peringkasan dengan metode yang digunakan memperlihatkan beberapa representasi kalimat dari *clustering* dengan baik.

Kata kunci: analisis sentimen, ulasan produk, *frequent pattern generation, association mining, semantic smimilarity scoring, sentence clustering.*

Abstract

Product reviews from consumers is one of important factor for saling product. Analyzing and summarizing sentiment classification of product reviews has three steps that must be done. First step is feature extraction using frequent itemset mining with apriori algorithm. Second, opinion classification using SentiWordnet to determine opinion word polarity of a sentence. Last, summarizing document with semantic similarity scoring and sentence clustering. The results of this study found that filtering corresponding word also affect the performance of the extraction in this study. Word filtering that used are Noun, Noun Phrase, intersection and union of both, four of that filtering words used have a fairly balanced outcome, union and intersection that are expected to improve performance still got the result which is not too far away with just Noun and Noun Phrase. Results of the extraction performance in this study was about 20-40% in the dataset used. Classification using SentiWordNet show a good results, but in some datasets that have fairly high complexity sentence also decreased, although not much different and the performance still at 40-90%. Sumarization documents can be done well on a provided dataset for the dataset has a number of sentences adequate product reviews and summary with this methods some sentences representation of clustering are done well.

Keywords: sentiment analysis, product review, frequent pattern generation, association mining, semantic smimilarity scoring, sentence clustering.

1. Pendahuluan

Ulasan produk dari konsumen merupakan salah satu faktor yang penting dalam penjualan suatu produk. Ulasan produk sering digunakan produsen suatu produk menunjang pemasaran suatu produk, terlihat pada penelitian oleh *Dimensional Search* pada 1046 orang koresponden menyatakan dalam menentukan keputusan dalam pembelian suatu produk 90% calon konsumen terpengaruh oleh ulasan produk positif, sedangkan di sisi lain 86% keputusan dipengaruhi oleh ulasan negatif suatu produk [1]. Kebebasan konsumen dalam memberikan ulasan kepada suatu produk pastinya memberikan dampak kepada banyaknya jumlah ulasan yang dapat menyulitkan konsumen dalam membaca serta memilih produk. Dibuktikan pada penelitiaan oleh *BrightLocal* yang menyatakan bahwa 85%

konsumen yang membaca hingga 10 ulasan dan 7% saja yang membaca hingga 20 lebih ulasan mengenai suatu produk [2], hal ini memperlihatkan ketidak efisiannya ulasan yang menumpuk dan membuat bingung konsumen dalam mengambil kesimpulan suatu produk. Oleh karena itu dibutuhkan suatu ringkasan ulasan produk, ringkasan ini diharapkan dapat mempermudah konsumen dalam membaca dan melakukan penarikan kesimpulan dari ulasan tersebut.

Menganalisis klasifikasi sentimen dan peringkasan suatu ulasan produk memiliki tiga tahap yang harus dilakukan, tahap pertama yaitu ekstraksi fitur suatu produk. Ekstraksi fitur yang digunakan pada penelitian ini adalah *frequent pattern generation* pada *association mining* [3]. Fitur dalam ulasan produk pada umumnya merupakan kata benda dalam ulasan sehingga metode ini cocok untuk digunakan untuk melakukan ulasan produk.

Proses selanjutnya setelah mendapatkan fitur yang diekstraksi dari ulasan produk, dilanjutkan pada tahap klasifikasi fitur produk untuk menentukan orientasi positif dan negatif pada fitur produk dalam ulasan produk. Memperhatikan keberadaan daftar kata ataupun kamus kata positif dan negatif telah cukup lengkap seiring perkembangan penelitian analisis sentimen [4] maka klasifikasi yang digunakan untuk penelitian kali ini dengan memanfaatkan kamus opini adjektif pada WordNet atau disebut dengan SentiWordNet [5].

Proses setelah dilakukan tahapan *preprosesing*, ekstraksi dan juga klasifikasi adalah tahap yang terakhir yaitu peringkasan ulasan produk. Peringkasan dilakukan guna memudahkan pembaca untuk melihat, membaca serta memahami ulasan suatu produk sehingga diperlukan peringkasan yang mudah dibaca serta dipahami. Untuk melakukan peringkasan dilakukan dengan *semantic similarity scoring* dan *sentence clustering* [6]. Peringkasan ini merupakan peringkasan ulasan produk pengembangan dari peringkasan ekstraktif dengan menerapkan *clustering* kalimat maka didapat peringkasan yang lebih baik dibandingkan dengan peringkasan ekstraktif biasa.

2. Landasan Teori

2.1 Analisis Sentimen

Analisis sentimen yang biasa dikenal juga dengan *Opinion Mining* merupakan ilmu yang bertujuan untuk menganalisis opini, sentimen, perilaku, penilaian dan emosi seseorang terhadap suatu entitas, seperti produk, jasa pelayanan, organisasi, atau individu. *Opinion extraction, sentiment mining, subjectivity analysis, emotion analysis, review mining,* merupakan contoh lain dari *opinion mining* yang memiliki tugas yang sedikit berbeda[8]. Analisis sentimen secara umum dibagi menjadi tiga level utama, yaitu level dokumen, level kalimat, dan level aspek dan entitas.

2.2 Lemmatization

Lemmatization merupakan proses pembentukan sebuah kata kedalam bentuk *lemma*-nya[9]. *Lemma* adalah bentuk kata dasar dari sebuah kata yang memiliki arti pada kamus. Contoh perubahan kata kerja "*to be*" pada bahasa Inggris akan di rubah dari "*am, is, are, was, were*" menjadi kata dasarnya yaitu "*be*".

Perbedaan antara *stemming* dan *lemmatization* adalah pada *stemming* akan langsung melakukan pemotongan pada awalan atau akhiran pada kata berimbuhan untuk mencapai tujuannya, sedangkan pada *Lemmatization* prosesnya menggunakan kamus dan analisis morfologi sebuah kata untuk mencapai tujuannya[10].

2.3 Association Mining

Association mining diperlukan untuk menemukan hubungan dari banyak data. Data masukan yang dibutuhkan untuk *association mining* adalah sekumpulan transaksi yang terdiri dari itemset pada setiap transaksi. Kasus yang sering dijumpai menggunakan *association mining* adalah menemukan hubungan antar barang yang dibeli oleh pelanggan melalui serangkaian transaksi pembelian. Proses ini dilakukan dengan melalui dua tahapan utama [12], yaitu:

1. *Frequent Itemset Generation*
   Tahapan ini berfungsi untuk menemukan *itemset* yang memenuhi *minimum support threshold*. *Itemset* yang lolos threshold ini disebut frequent itemset.
2. *Rule Generation*
   Tahapan ini berfungsi untuk mengeksrak rule yang memiliki nilai *confidence* diatas nilai tertentu. Rule yang dihasilkan merupakan keluaran terahir dari association mining yang nantinya dianggap sebagai hubungan antar item.

2.4 Nearest Opinion Word

Nearest Opinion Word Merupakan salah satu cara yang simpel dalam proses *Feature-Opinion Association Problem* (*FOA*) yaitu dengan memasangkan kata opini dengan kandidat fitur terdekat [6]. Metode tersebut mencari

nilai jarak kedekatan suatu kata opini dengan kata fitur, dimana kata opini yang paling dekat dengan salah satu kata fitur produk, maka akan menjadi kata opini dari kata fitur produk tersebut. Kedekatan dinyatakan dengan nilai *rel(f,w)* terbesar. *rel(f,w)* memiliki rumus perhitungan sebagai berikut:

$$rel(f,w) = \frac{1}{dist(f,w)} \quad (1)$$

rel(f,w) merupakan nilai invers dari jarak antara kata opini(*w*) dengan kata fitur produk(*f*) tertentu yang dinyatakan dengan *dist(f,w)*. Dengan metode ini kata opini akan menjadi milik kata fitur yang jaraknya paling dekat dengan dirinya. Metode ini juga dilengkapi dengan sebuah *threshold* yang akan menjadi batas maksimum terjauh kedekatan antara suatu kata opini dengan kata fitur. Dengan kata lain, kata opini akan menjadi pasangan suatu kata fitur yang jaraknya paling dekat dan memenuhi batas maksimum terjauh jarak antar keduanya.

2.5 Wordnet

WordNet adalah suatu kamus leksikal bahasa Inggris. Kata kerja, kata benda, kata sifat, dan kata keterangan digabungkan dalam sebuah *synsets*. Synsets dihubungkan oleh *conceptual-semantic* dan hubungan leksikal. Struktur WordNet sangat membantu dalam *computational linguistic* dan *natural language processing*[13]. Selain *synsets* pada WordNet juga terdapat SentiWordNet yang merupakan daftar kata adjective yang telah memiliki polaritas atau nilai postif maupun negatif dalam suatu kalimat[13]. SentiWordNet ini yang digunakan penulis dalam melakukan klasifikasi positif atau negatif suatu kalimat.

2.6 Evaluasi

Evaluasi untuk menentukan penghitungan akurasi pada tugas akhir ini dulakukan dengan menggunakan perhitungan *Precision* dan juga *Recall*. Perhitungan *Precison* dan *Recall* adalah perhitungan untuk mengukur dokumen yang relevan terhadap suatu informasi yang diinginkan, untuk mendapatkan nilai akurasi yang tinggi. Berikut adalah 4 pengkategorian dokumen dalam suatu proses pencarian[14], yaitu:

1. *True positives* (TP) yaitu, dimana dokumen yang relevan dapat diidentifikasi dengan benar sebagai dokumen yang relevan.
2. *True negatives* (TN) yaitu, dimana dokumen tidak relevan dapat diidentifikasi dengan benar sebagai dokumen yang tidak relevan.
3. *False positives* (FP) yaitu, dimana dokumen yang tidak relevan namun salah teridentifikasi sebagai dokumen yang relevan.
4. *False negatives* (FN) yaitu, dimana dokumen yang relevan namun salah teridentifikasi sebagai dokumen yang tidak relevan.

Precision dan *recall* merupakan suatu perhitungan yang diperggunakan untuk menghitung evaluasi akurasi pada sebuah dokumen

1. *Precision*, menghitung banyaknya item yang teridentifikasi sebagai item yang relevan, dengan rumus:

$$\frac{TP}{TP + FP} \quad (2)$$

2. *Recall*, menghitung banyaknya item yang relevan yang berhasil diidentifikasi, dengan rumus

$$\frac{TP}{TP + FN} \quad (3)$$

3. F-Score, menghitung akurasi keseluruhan yang merupakan penggabungan dari precision dan recall, dengan rumus:

$$2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}} \quad (4)$$

Pada penelitian ini evaluasi dilakukan berbasis kalimat, untuk setiap kalimat akan dihitung performansinya dan selanjutnya akan dihitung rata-ratanya dalam satu dokumen. Untuk ekstraksi fitur performansi dihitung melalui nilai *precision*, *recall* dan *f-score*. Untuk proses klasifikasi atau perberian orientasi opini performansi ditunjukkan dengan melalui nilai akurasi.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Table 1-1 Evaluasi Klasifikasi

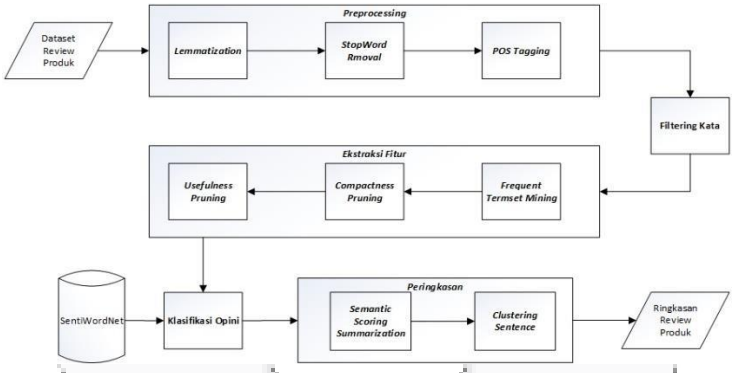
No.	Polarity Fitur Label	Polarity Fitur Prediksi	Akurasi

1	-	-	0.0
2	camera[+]	camera[+]	1.0
3	camera[-]	camera[-]	1.0
4	camera[+]	camera[-]	0.0
5	camera[+]	camera[+] camera[+]	1.0
6	camera[+]	camera[+] camera[-]	0.5
7	camera[+] size[+]	camera[+] camera[-] camera[0] size[+]	0.5
8	camera[+] camera[-]	camera[+]	0.5
Rata-Rata			4.5 / 7 = 64.3 %

3. Pembahasan

3.1 Gambaran Umum Sistem

Tahapan penentuan opini ini dibangun oleh empat tahapan utama yaitu, (1) Ekstraksi fitur berdasarkan ulasan data, (2) Identifikasi orientasi kata dan kalimat opini, (3) Pengklasifikasian orientasi opini terhadap fitur, dan (4) Pembangkitan ringkasan. Berikut gambaran sistem yang akan dibangun:



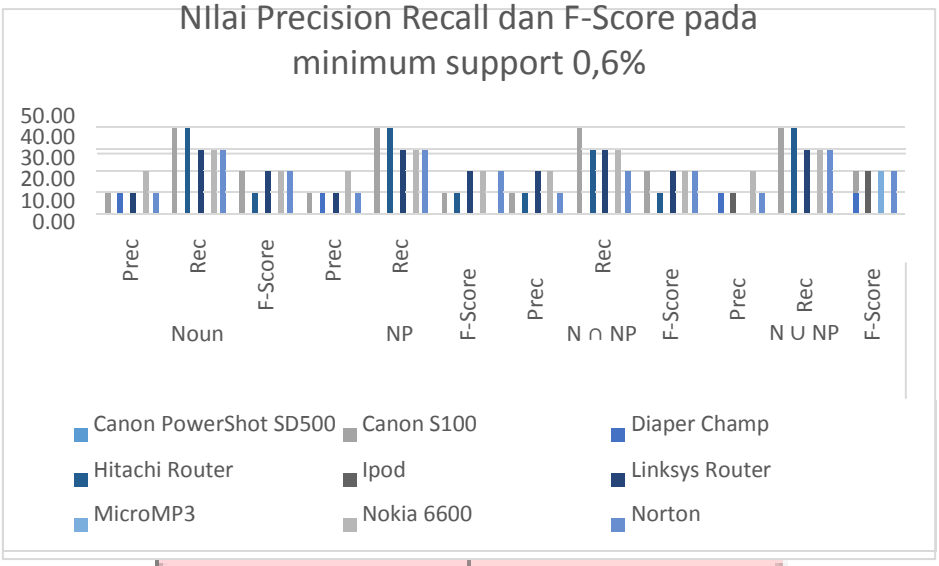
Gambar 3-1 Gambaran Umum Sistem

3.2 Ekstraksi Fitur dengan Association Mining

Tahap ektraksi fitur digunakan *frequent pattern generation* pada *association mining* [3]. Kata benda ataupun frase kata benda yang sering dibicarakan oleh konsumen dalam sebuah review produk menjadi kandidat fitur yang potensial [3]. Ekstraksi fitur tersebut dapat dilakukan menggunakan *association mining*, dengan tujuan menemukan *frequent itemset* yang dapat mewakili fitur sebuah produk. Beberapa masalah timbul pada metode ini seperti adanya gabungan beberapa kata yang tidak memiliki arti. *Compactness pruning* dilakukan untuk mengeliminasi gabungan kata tersebut. Eliminasi fitur dapat dilakukan dengan menghitung jarak antar kedua kata yang muncul, apabila jaraknya terlalu jauh dan melewati *threshold* tertentu maka fitur tersebut akan dieliminasi. Selain itu dilakukan juga *usefulness pruning.* Jika *pure support* lebih kecil dari *minimum pure support* dan fitur yang diidentifikasi merupakan subset dari frase fitur yang lain, maka fitur tersebut dieliminasi. *Pure support* merupakan jumlah kalimat yang terdapat fitur dalam bentuk kata atau frase dan tidak terdapat frase fitur yang merupakan superset dari fitur tersebut. Contohnya, fitur *life* dianggap bukan merupakan fitur yang berarti, sedangkan *battery life* merupakan frase fitur yang dianggap lebih memiliki arti untuk menjadi sebuah fitur produk.

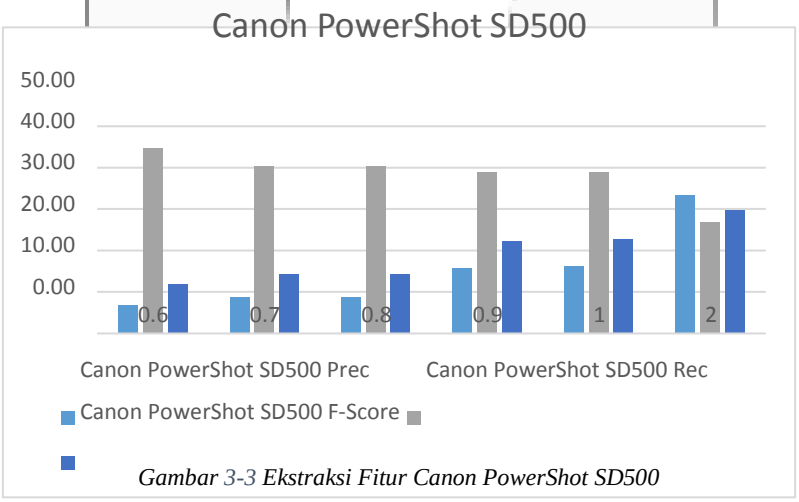
Ekstraksi Fitur pada penelitian tugas akhir ini menggunakan algoritma apriori dengan bantuan library SPMF pada Bahasa pemrograman Java. Pengujian dilakukan dengan menggunakan nilai parameter *minimum support* pada rentang 0,6%-2%. Nilai *minimum support* merupakan parameter yang paling menentukan dalam mendapatkan jumlah fitur terekstrak yang sesuai. Semakin kecil *minimum support*, maka semakin besar jumlah fitur yang terekstrak. Jumlah fitur yang terekstrak akan berpengaruh terhadap nilai *precision*. Jika jumlah fitur terekstrak

melebihi jumlah fitur yang seharusnya, maka nilai *precision* akan lebih kecil dari nilai *recall*. Berikut daftar tabel menggunakan *minimum support* terkecil pada pengujian yaitu 0,6% dapat dilihat hasil *precision* akan selalu lebih kecil dibanding *recall* karena fitur yang terekstrak banyak dengan *minimum support* rendah.

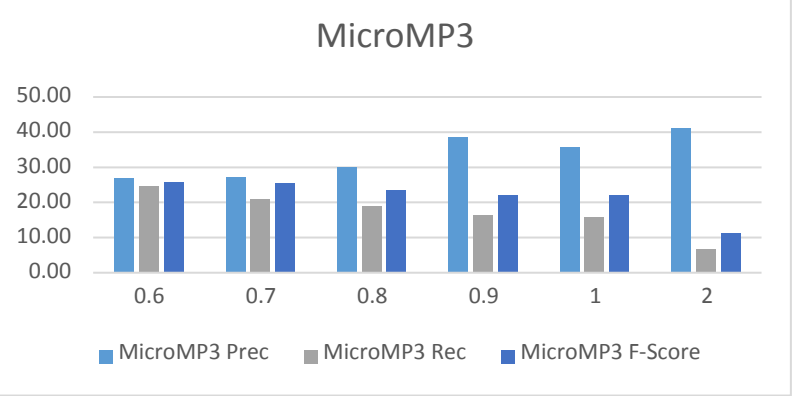


Gambar 3-2 Perbandingan Nilai Precision, Recall, dan F-Score

Jumlah baris kalimat dalam dataset juga mempengaruhi penentuan nilai *minimum support* untuk mendapatkan hasil ekstraksi yang sesuai. Semakin banyak jumlah kalimat dalam dataset, *minimum support* yang sesuai akan semakin besar, dan begitu pula sebaliknya. Contohnya adalah dataset Canon PowerShot SD500 dengan jumlah kalimat paling sedikit yaitu 229 baris kalimat selalu meningkat saat *minimum support* dinaikan. Sedangkan pada dataset dengan jumlah kalimat terbanyak akan semakin tinggi apabila *minimum support* diturunkan. Dapat dilihat pada gambar dibawah ini.



Gambar 3-3 Ekstraksi Fitur Canon PowerShot SD500



Gambar 3-4 Ekstraksi Fitur MicroMP3

Frequent termset yang dihasilkan oleh algoritma apriori merupakan kandidat fitur produk. Kandidat fitur tersebut nantinya akan memasuki tahapan *prunning* untuk menyeleksi kembali kandidat yang cocok dianggap sebagai fitur.

Proses *prunning* memiliki parameter masukan seperti *minimum distance*, *min occurence*, dan *minimum pure support*. Penentuan nilai parameter yang sesuai akan menghasilkan hasil *prunning* yang baik sehingga meningkatkan akurasi ekstraksi. Berikut pengujian *prunning* dengan menggunakan parameter *minimum distance* 2, *minimum occurence* 5, dan *minimum pure support* 3 pada dataset Canon PowerShot SD500.

Tabel 3-5 Hasil ekstraksi Sebelum Prunning

Dokumen	Noun			NP			N ∩ NP			N ∪ NP		
	Prec	Rec	F-Score	Prec	Rec	F-Score	Prec	Rec	F-Score	Prec	Rec	F-Score
Canon PowerShot SD500	15.03	34.33	**20.91**	13.23	37.31	**19.53**	16.28	31.34	**21.43**	16.15	38.81	**22.81**
Canon S100	19.38	31.31	**23.94**	17.95	35.35	**23.81**	22.56	30.30	**25.86**	21.29	33.33	**25.98**
Diaper Champ	12.20	26.67	**16.74**	11.73	25.33	**16.03**	14.66	22.67	**17.80**	13.38	25.33	**17.51**
Hitachi Router	19.13	24.18	**21.36**	23.15	27.47	**25.13**	20.00	26.37	**22.75**	22.43	26.37	**24.24**
Ipod	19.69	28.41	**23.26**	26.97	27.27	**27.12**	25.84	26.14	**25.99**	24.04	28.41	**26.04**
Linksys Router	23.53	25.00	**24.24**	23.53	25.00	**24.24**	28.95	22.92	**25.58**	26.32	26.04	**26.18**
MicroMP3	32.00	14.88	**20.32**	32.35	15.35	**20.82**	35.06	12.56	**18.49**	35.79	15.81	**21.94**
Nokia 6600	29.63	20.38	**24.15**	35.35	22.29	**27.34**	40.00	19.11	**25.86**	33.02	22.29	**26.62**
Norton	18.24	23.68	**20.61**	20.28	25.44	**22.57**	22.12	21.93	**22.03**	20.93	23.68	**22.22**

Tabel 3-6Hasil Ekstraksi Setelah Prunning

Dokumen	Noun			NP			N ∩ NP			N ∪ NP		
	Prec	Rec	F-Score	Prec	Rec	F-Score	Prec	Rec	F-Score	Prec	Rec	F-Score
Canon PowerShot Sd500	24.72	32.84	**28.21**	24.74	35.82	**29.27**	27.40	29.85	**28.57**	27.47	37.31	**31.65**
Canon S100	33.33	30.30	**31.75**	30.28	33.33	**31.73**	36.25	29.29	**32.40**	35.16	32.32	**33.68**
Diaper Champ	24.36	25.33	**24.84**	20.88	25.33	**22.89**	26.56	22.67	**24.46**	22.50	24.00	**23.23**
Hitachi Router	32.84	24.18	**27.85**	32.89	27.47	**29.94**	32.00	26.37	**28.92**	34.29	26.37	**29.81**
Ipod	30.12	28.41	**29.24**	30.26	26.14	**28.05**	30.67	26.14	**28.22**	31.25	28.41	**29.76**
Linksys Router	28.57	25.00	**26.67**	28.24	25.00	**26.52**	33.85	22.92	**27.33**	31.25	26.04	**28.41**
MicroMP3	38.10	14.88	**21.40**	36.26	15.35	**21.57**	38.03	12.56	**18.88**	39.08	15.81	**22.52**
Nokia 6600	40.51	20.38	**27.12**	46.05	22.29	**30.04**	51.72	19.11	**27.91**	43.21	22.29	**29.41**
Norton	27.59	21.05	**23.88**	29.55	22.81	**25.74**	32.39	20.18	**24.86**	31.65	21.93	**25.91**

Dapat dilihat pada tabel sebelum dan sesudah prunning terjadi peningkatan performansi pada *precision*, *recall* maupun *F-Score*.

Dari analisis penelitian ini didapat nilai F-Score dengan evaluasi dokumen sebesar 20-40%. Hal ini juga disebabkan oleh beberapa fitur implisit yang tidak secara langsung disebutkan dalam kalimat ulasan produk. Berikut contoh kalimat yang mempunyai fitur eksplisit atau tidak terlihat secara langsung atau ada kata yg mengacu ke sebuah fitur produk yang lain.

3.3 Pemberian Orientasi Opini

Penentuan orientasi terhadap fitur dalam kalimat dilakukan dengan menemukan pasangan fitur dan opini terlebih dahulu. Setelah pasangan ditemukan, orientasi fitur akan mengikuti orientasi kata opini pasangannya yang telah ditentukan pada proses sebelumnya. Sebagai contoh pada kalimat "image quality [Good_pos] ## Good image quality" memiliki fitur image quality dengan kata good sebagai pasangannya. Pasangan fitur player dengan opini loved ditemukan pada kalimat "player[+]##i took it to my father's house to play a tom jones concert dvd and he loved the player so much i gave it to him". Kata amazing dan loved berdasarkan SentiWordnet memiliki orientasi positif, hal ini menyebabkan fitur picture quality dan player memiliki orientasi opini positif.

Pasangan fitur dan opini pada penelitian tugas akhir ini dilakukan mengikuti salah satu metode dalam penelitian Kam Tong Chan [21]. Metode yang digunakan termasuk kedalam metode sederhana dalam penelitian tersebut, yaitu dengan menghitung jarak antara fitur dan opini. Nilai relasi antara fitur dan opini dilambangkan dengan dengan fungsi rel(f, w), dengan f adalah fitur produk dan w adalah kata opini. Nilai rel(f,w) yang melebihi batas threshold tertentu akan dijadikan sebagai pasangan fitur dan opini yang valid. Fungsi tersebut dihitung menggunakan persamaan(1).

3.4 Peringkasan Review Produk

Peringkasan ulasan merupakan fokus dari tugas akhir ini, peringkasan dilakukan setelah melakukan klasifikasi orientasi sebuah kalimat per fitur kemudian secara ektraktif dilakukan peringkasan yang akan menghasilkan ringkasan ektraktif seperti berikut.

a. Lens

(+): 57 sentences

1. The lens feels very solid!
2. I have taken a whole bunch of excellent pictures with this lens.

...

(-): 15 sentences

1. I do not satisfy with the included lens kit.
2. The lens cap is very loose and co,e off very easily!

...

b. Battery life

(+): 32 sentences

1. The battery lasts forever on one single charge.
2. The battery duration is amazing!

...

(-): 8 sentences

1. I experienced very short battery life from this camera.
2. It uses a heavy battery.

...

Tahap selanjutnya akan dilakukan *cluster* kalimat dengan menghitung kesamaan tiap kalimat yang kemudian apabila nilai similarity yang dihitung melebihi *treshold* yang ditentukan maka kalimat tersebut akan dilakukan cluster atau dikelompokkan menjadi satu. Adapun penghitungan nilai kesamaan dua kalimat untuk menentukan pengelompokan kalimat pada peringkasan ulasan produk dapat dilakukan dengan rumus penghitungan fungsi *cosinus* seperti pada formula (2.3)

Dimana V_1 dan V_2 merupakan vector representasi dari kalimat pertama dan kedua yang akan dihitung nilai kesamaannya.

Berikut contoh penghitungan kesamaan antar 2 kalimat [6] dapat dihitung nilai *semantic smimilaritynya.*

= The battery of this camera is very impressive.

= Canon camera always has a long battery life.

Sehingga didapatkan *join vector* antara dan adalah sebagai berikut. *Join vector* merupakan *vector* gabungan antar dua kalimat yang akan dibandingan nilai *similarity*-nya.

C = {battery, camera, impressive, has, long, life}

Dihasilkan penghitungan vector masing masing kalimat mengasilkan dua vector kalimat yang akan dihitung dengan vektor didapat dengan persamaan[7].

$$(\quad,\quad) = \frac{2 * \quad}{N_1 + \quad + 2 * \quad} \tag{6}$$

V1 = {1.0, 1.0, 1.0, 0.0, 0.3, 0.15}

V2 = {1.0, 1.0, 0.3, 1.0, 1.0, 1.0}
Sehingga didapat nilai *similarity* yaitu sim(V1, V2) = 0.69 dengan persamaan

$$sim(V_1, V_2) = \frac{V_1 . V_2}{\|V_1\| . \|V_2\|} \quad (7)$$

Berikut adalah hasil peringkasan dokumen setelah dilakukan *clustering* kalimat dengan menghitung semantic similarity score tiap kalimatnya apa bila nilai memenuhi threshold yang ditentukan maka kalimat akan di *cluster* dengan kalimat yang mempunyai nilai besar *vector* yang lebih tinggi

a. Lens

{ (+) The lens feels very solid (+10 similar)
(-) I think the lens does not worth it, it's a bit too fragile. (+2 similar)

{ (+) I have taken a lot of excellent pictures with this lens. (+7 similar)
(-) Don't buy this lens, I always get my pictures blurred. (+0 similar)

...

b. Battery life

{ (+) The battery lasts forever on one single charge. (+18 similar)
(-) I experienced very short battery life from this camera. (+4 similar)

{ (+) 0 sentence
(-) It uses a heavy battery (+3 similar)

...

4. Kesimpulan

Berdasarkan hasil pengujian dan analisa yang telah dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Performansi ekstraksi fitur menggunakan *frequent itemset mining* dengan algoritma apriori sangat dipengaruhi oleh seberapa *minimum support* yang digunakan serta jumlah kalimat yang ada pada dataset apabila *minimum support* sesuai maka performansi akan semakin baik begitu juga sebaliknya. Performansi ekstraksi fitur yang didapat adalah sekitar 20-40% dibandingkan dengan fitur yang ada pada dataset. Penambahan *pruning* pada ekstraksi fitur dapat meningkatkan nilai performansi sebesar 3-10% dengan menggunakan parameter yang sesuai.
2. Penentuan orientasi opini menggunakan SentiWordNet dan juga *Nearest Opinion Word* pada kalimat menghasilkan hasil performansi yang baik pada semua data yang diuji yaitu apabila dibandingkan dengan hasil klasifikasi pada dataset dengan hasil performansi sekitar 40-90%.
3. Peringkasan kalimat dengan metode *clustering sentence* dan *semantic similarity scoring* dapat dilakukan apabila kalimat yang akan di cluster memiliki kesamaan serta jumlah kalimat pada datasetnya tidak terlalu kecil.
4. *Filtering* kata gabungan NN dan NP lebih baik untuk melakukan ekstraksi disbanding dengan filtering dengan NN atau NP saja walaupun perbedaan keduanya tidak terlalu jauh namun gabungan NN dan NP lebih baik di semua rata-rata dataset dan *minimum support*. Namun, untuk kasus dengan dataset yang memiliki mayoritas kalimat yang kompleks seleksi kata yang dilakukan pada penelitian ini m,asih kurang karena memiliki nilai evaluasi 20%-40% saja.

5. Saran

Saran yang diperlukan dari tugas akhir ini untuk pengembangan sistem selanjutnya adalah sebagai berikut:

1. Mampu menangani singkatan kata.

2. Mampu menangani fitur implisit yang diutarakan dalam kalimat pada dataset.
3. Menggunakan penerapan *coreference resolution* guna meningkatkan performansi sistem dan menangani fitur implisit.
4. Menggunakan *negation handling* sebelum klasifikasi dengan SentiWordnet agar performansi meningkat untuk data yang memiliki banyak ulasan dendan negasi.

6. Daftar Pustaka

[1] A. Gesenhues, "Survey: 90% Of Customers Say Buying Decisions Are Influenced By Online Reviews," Marketing Land, 9 April 2013. [Online]. Available: http://marketingland.com/survey-customers-more-frustrated-by-how-long-it-takes-to-resolve-a-customer-service-issue-than-the-resolution-38756. [Diakses 21 October 2015].

[2] M. Anderson, "88% Of Consumers Trust Online Reviews As Much As Personal Recommendations," Search Engine Land, 7 July 2014. [Online]. Available: http://searchengineland.com/88-consumers-trust-online-reviews-much-personal-recommendations-195803. [Diakses 21 October 2015].

[3] S. H. Ghorashi, R. Ibrahim, S. Noekhah dan S. Dastjerdi, "A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Review," *IJCSI International Journal of Computer Science Issues,* vol. IX, no. 4, pp. 29-35, 2012.

[4] B. Liu, "Opinion Mining, Sentiment Analysis, and Opinion Span Detection," University of Illinois, [Online]. Available: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. [Diakses 3 Juni 2015].

[5] V. S. Jagtap dan K. Pawar, "Analysis of Different Approach to Sentence Level Sentiment Classification," *International Journal of Scientific Engineering and Technology,* vol. 2, no. 3, pp. 164-170, 2013.

[6] D. K. Ly, K. Sugiyama, Z. Lin dan M.-Y. Kan, "Product Review Summarization based on Facet Identification and Sentence Clustering," pp. 1-10, 2011.

[7] Z. Wu dan M. Palmer, "Verb Semantics and Lexical Selection," vol. 6, pp. 133-138, 1994.

[8] B. Liu, "Sentiment Analysis and Opinion Mining," 2012.

[9] S. Nirenburg, Penyunt., dalam *Language Engineering for Lesser-studied Languages*, IOS Press, 2009, p. 31.

[10] March 2016. [Online]. Available: http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.

[11] B. Santorini, "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)," Pennsylvania, 1990.

[12] P.-N. Tan, M. Steinbach dan V. Kumar, Introduction to Data Mining, Boston: Addison-Wesley Longman Publishing, 2005.

[13] G. A. Miller, R. Beckwith dan C. Fellbaum, "Intoduction to WordNet: An On-line Lexical Database," 1993.

[14] E. K. d. E. L. S. Bird, "Natural Language Processing with Python," 2009.