EKSTRAKSI OPINI DENGAN MENGGUNAKAN PENDEKATAN ASSOCIATION RULE MINING

OPINION EXTRACTION USING ASSOCIATION RULE MINING APPROACH

Kurniawan Adina Kusuma[1], Warih Maharani, S.T., M.T.[2], Moch. Arif Bijaksana, Ph.D.[3]

[1,2,3]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung
[1]kurniawanadinak@gmail.com, [2]wmaharani@gmail.com, [3]arifbijaksana@gmail.com

Abstrak

Transaksi secara *online* sudah berkembang secara pesat dewasa ini. Jumlah produk yang terjual secara *online* semakin bertambah banyak dan beranekaragam jenisnya. Sebuah *review* tentang produk acap kali diberikan oleh konsumen untuk mengomentari produk-produk yang sudah mereka beli. Pada jenis produk dengan *brand* yang sudah terkenal memiliki *review* yang sangat banyak. Seharusnya *review* dari konsumen bisa dimanfaatkan produsen sebagai *feedback* dan digunakan calon konsumen sebagai referensi saat akan membeli barang. *Review* produk yang jumlahnya semakin banyak akan menyulitkan pembaca jika harus membacanya satu persatu. Solusinya adalah dengan mengidentifikasi fitur produk secara spesifik dari *review* yang sudah ditulis konsumen. Tugas akhir ini dilakukan untuk mengidentifikasi masalah *featured-based opinion summarization* dari *review* konsumen. Proses identifikasi ini terdiri dari dua tahap utama, yaitu : (1) ekstraksi fitur produk yang sudah di*review* oleh konsumen (*feature extraction*); dan (2) identifikasi polaritas fitur untuk menentukan polaritas kalimat opini (*sentiment analysis*). Metode yang digunakan pada *opinion extraction* ini adalah dengan metode *association rule mining* dengan algoritma apriori. Berdasarkan hasil pengujian, penggunaan metode *association rule mining* terbukti dapat mengekstrak fitur produk. Fitur yang sudah diekstraksi tersebut kemudian dicek pada setiap kalimat untuk menemukan orientasi opini dengan bantuan *SentiWordNet* sehingga didapatkan polaritas opini fitur yang dibicarakan oleh konsumen pada *review.*

Kata kunci : ***Association rule mining, Feature extraction, Opinion summarization, Sentiment analysis***

Abstract

Online transactions has been growing rapidly these days. The number of products which sold through online transactions is increasing and many kind of types. A review of the product is often provided by the consumer to comment on the products they have purchased. On the types of branded products has a review that very much. Supposed to be a review of the consumer can be used as a feedback producers and consumers would use as a reference when buying goods product. A review which increasing number would make it difficult to read one by one. The solution is to identify the specific features of the products on the reviews that have been written by consumer. The final task is done to identify problems featured-based opinion summarization of consumer reviews. The identification process consists of two main steps: (1) extraction of product features that have been reviewed by the consumer (feature extraction); and (2) identification of polarity opinion feature to determine the polarity of opinion sentence (sentiment analysis). Author use association rule mining method with apriori algorithm to feature extraction process. Based on testing result, association rule mining method proved to be able to extract the features of the product. Features that have been extracted is used to check on each sentence to find the orientation of opinion using SentiWordNet to obtain the polarity of opinion were discussed by the consumer features on the review.

Keywords : Association rule mining, Feature extraction, Opinion summarization, Sentiment analysis

1 Pendahuluan

Berkembangnya *online shop* dan *e-commerce* dengan jumlah yang banyak berdampak pada semakin banyaknya produk yang terjual secara *online.* Produsen juga memberikan kesempatan bagi konsumen untuk memberikan *review* tentang produk yang sudah dibelinya. Hal itu merupakan sebuah *feedback* bagi perusahaan dan bisa membantu calon pembeli untuk mengetahui informasi produk yang akan dibelinya. Melalui *feedback* tersebut perusahaan memperoleh informasi tentang *segmention, targeting,* dan *positioning* produk mereka. Jumlah *review* dari konsumen semakin bertambah banyak seiring meningkatnya transaksi *online* sehingga mengakibatkan calon pembeli merasa kebingungan saat membaca *review* dan memilih produk yang sesuai dengan keinginannya. Selain itu perusahaan juga kesulitan dalam menganalisis informasi dari *feedback* tersebut.

Oleh karena itu, dibutuhkan sebuah sistem yang mampu untuk memberikan sebuah ringkasan dari banyaknya jumlah *review* dari konsumen. Salah satu solusi untuk mengatasi masalah ini adalah dengan *opinion summarization. Opinon summarization* ini bertujuan untuk mengidentifikasi dan mengekstraksi fitur produk dan

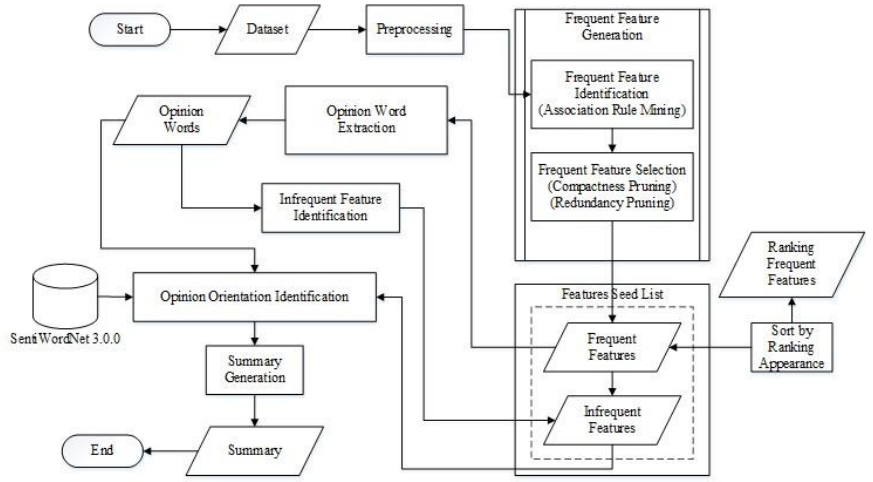
mengidentifikasi orientasi atau polaritas opini yang berhubungan dengan fitur produk tersebut. Proses *opinion summarization* ini dibagi menjadi dua tahap utama yaitu: (1) *feature extraction* dan (2) *opinion orientation identification* [1].

Isi *review* dari konsumen terdiri dari banyak hal dan tidak semua opini berhubungan langsung dengan fitur produk. Setiap konsumen mempunyai cara masing-masing dalam menyampaikan opini mereka tentang sebuah produk yang sudah dibelinya. Kata-kata yang digunakan oleh konsumen untuk mengomentari produk juga mempengaruhi sebuah *review* produk tersebut. Metode yang digunakan pada penelitian *opinion summarization* ini adalah metode *association rule mining* dengan algoritma apriori.

Alasan menggunakan metode *association rule mining* adalah metode ini mampu menemukan aturan asosiatif antara kombinasi item, sehingga didapatkan aturan asosiatif yang berupa *frequent itemset* dari kombinasi *review*. *Frequent itemset* tersebut cenderung merupakan sebuah fitur produk yang dilihat berdasarkan *minimum support* (presentase kemunculan kombinasi *item* tersebut dalam *database*) dan *minimum confidence* (kepastian yang berasal dari hubungan antar item dalam aturan tersebut) [1]. Penelitian ini hanya menggunakan *frequent itemset* dari algoritma apriori untuk memperoleh fitur produk dari *review* konsumen.

2 Teori dan Tahap Perancangan

Gambar 1 di bawah ini merupakan gambaran arsitektur *opinion summarization* pada penelitian yang dilakukan. Penelitian ini bertujuan untuk mengekstrak opini berdasarkan fitur produk yang sebelumnya sudah didapatkan dari *review* produk. Hasil akhir berupa sebuah ringkasan yang mengklasifikasikan kalimat opini berdasar fitur produk dan polaritas atau orientasinya.



Gambar 1 Flowchart perancangan sistem

2.1 Preprocessing

Sebelum masuk ke proses ekstraksi fitur, *dataset* melalui *preprocessing* terlebih dahulu dengan tujuan membersihkan *dataset* dan mengembalikannya ke benttuk *lemma* nya. Pada penelitian ini digunakan data *cleaning, lemmatization,Part-of-Speech Tagging* atau sering disingkat dengan *POS Tagging,* dan *stopword removal.*

Data *cleaning* sendiri mempunyai maksud untuk membersihkan *noise* pada *dataset*, semua karakter selain huruf dan angka akan dihapus oleh sistem.

Lemmatization adalah salah satu teknik dalam normalisasi teks. Teknik *lemmatization* ini akan menghapus bentuk kata yang mempunyai akhiran infleksi untuk kemudian dikembalikan ke bentuk *lemma* atau bentuk dasarnya [2]. Dalam pengerjaan tugas akhir ini digunakan sebuah *lemmatizer* NLP *Stanford* untuk melakukan analisis morfologi dalam mengidentifikasi *lemma* sebuah kata. Berikut adalah contoh kalimat sebelum dan setelah melalui *lemmatization* :

Troubleshooting ad-2500 no picture scrolling bw strengths are well listed by other reviewers.
Troubleshoot ad-2500 no picture scroll bw strengths be well list by other reviewer.

POS Tagging sendiri merupakan proses memberi label pada kata sesuai dengan klasifikasi *part-of-speech* nya. Sistem ini menggunakan *library Stanford Parser* untuk mengurai setiap *review* yang menghasilkan *part of speech* untuk setiap kata (*noun, verb, adjective,* dan lainnya). Berikut ini adalah contoh sebuah kalimat yang sudah melalui proses *POS Tagging* :

We_PRP really_RB enjoyed_VBD shooting_NN with_IN the_DT Canon_NNP Powershoot_NNP. *Stopword removal* merupakan sebuah teknik yang digunakan untuk menghilangkan kata-kata umum seperti *a, all, after* dari sebuah kalimat. Dalam beberapa aplikasi yang tidak memperhatikan struktur kalimat, kata-kata tersebut bisa memberikan bobot nilai kecil. Awalnya *stopword removal* dilakukan untuk menghemat ruang pada *indexing,*

tetapi pada perkembangannya tidak semua kata-kata *stopword* dihilangkan dari dokumen karena bisa membantu ataupun memberikan pengaruh pada kata-kata lainnya [3]. Frequent Feature Generation

2.2 Frequent Feature Identification

Identifikasi fitur produk dilakukan untuk menemukan fitur produk yang banyak diperbincangkan konsumen dalam mengekspresikan opininya terhadap suatu produk tertentu. Dalam tugas akhir ini, fitur produk yang diidentifikasi adalah fitur produk yang dinyatakan secara langsung di dalam sebuah kalimat opini.

Proses ini berfokus pada *frequent itemset.* Konsumen mempunyai cara yang beragam dalam menyampaikan opininya pada sebuah *review*, maka dari itu digunakan *association rule mining* untuk menemukan *frequent itemset* yang mana *frequent itemset* tersebut memiliki kemungkinan merupakan sebuah fitur produk. Selain itu, fitur produk biasanya berupa *noun* atau *noun phrase* yang terdapat pada *review* [4].

Tidak semua fitur produk yang sudah digenerasi merupakan fitur yang berguna atau relevan, terdapat beberapa fitur yang tidak menarik dan mubazir. Solusinya adalah dengan melakukan *feature selection* melalui metode *pruning* sebagai berikut [1]:

1. *Compactness Pruning*
   Metode ini dilakukan untuk mengecek fitur yang mengandung sedikitnya dua kata yang disebut *feature phrases*, dan menghilangkan fitur yang tidak berarti. Pada metode ini dilakukan pengecekan fitur yang *compact* pada suatu kalimat [1]. *Feature phrase* yang terdiri dari dua kata akan dicek jaraknya pada kalimat, apabila jaraknya melibihi parameter yang sudah ditentukan maka fitur tersebut dihapus [5].
2. *Redundancy Pruning*
   Metode ini berfokus untuk menghilangkan fitur yang bersifat redundan yang terdiri dari satu kata. Pada *pruning* ini dilakukan proses penghitungan *p-support (pure support)* dari sebuah fitur [1]. *P-support* yang dimaksud yaitu kemunculan sebuah fitur itu sendiri yang merupakan *subset* dari *feature phrase* pada kalimat tanpa *superset* nya) [5].

Proses *feature selection* selain menggunakan metode *pruning* juga digunakan perangkingan kemunculan dari *frequent itemset* kemudian diseleksi dengan menggunakan *threshold* minimum kemunculannya. Apabila *frequent itemset* tersebut kemunculannya kurang dari *threshold* yang sudah ditentukan maka akan dibuang. Namun hasil perangkingan *frequent itemset* ini hanya digunakan untuk analisis. Hasil dari *pruning frequent itemset* yang digunakan ke proses selanjutnya.

2.3 Opinion Words Extraction

Opinion words merupakan kata-kata yang digunakan untuk mengekspresikan opini yang bersifat subjektif. Pada penelitian ini, kata *adjective* dan *adverb* digunakan sebagai *opinion words* dengan syarat kalimat tersebut memiliki satu atau lebih fitur produk [4]. Konsumen mengekspresikan opininya dengan menggunakan berbagai macam kata. Proses *opinion words extraction* dapat dilakukan dengan menggunakan *frequent feature* yang sudah melalui proses *pruning* [1]. Kata opini digunakan untuk mengekspresikan orientasi/ mengidentifikasi polaritas fitur produk pada *review*. Proses ekstraksi kata opini dilakukan dengan menggunakan *n-gram.*

2.4 Infrequent Feature Identification

Fitur produk yang berupa *noun* atau *noun phrase* tidak semuanya terekstrak melalui *association rule mining* karena kemunculannya pada dokumen sangat sedikit, namun fitur tersebut memiliki kemungkinan merupakan sebuah fitur yang menarik, fitur-fitur tersebut adalah *infrequent features.* Proses menemukan *infrequent feature* tersebut dilakukan ekstraksi *noun* atau *noun phrase* yang terletak di sekitar *opinion words* yang terdapat dalam kalimat tersebut [1].

2.5 Feature Opinion Orientation Identification

Proses identifikasi orientasi opini fitur merupakan tahapan untuk mengetahui bobot nilai dari kata opini yang selanjutnya digunakan untuk menentukan orientasi atau polaritas pada sebuah fitur produk. Identifikasi dilakukan dengan menggunakan kamus *SentiWordNet* yang berisi sekumpulan kata-kata beserta dengan bobot nilai positif dan nilai negatif. Bobot nilai yang sudah didapat dari kamus *SentiWordNet* kemudian disematkan pada fitur produk sesuai dengan pasangan opininya. Apabila bobot nilai lebih dari nol maka fitur produk diberi label polaritas positif [+], jika bobot nilai kurang dari nol maka fitur produk diberi label polaritas [-], namun jika bobot nilai sama dengan nol maka fitur produk tersebut dihapus.

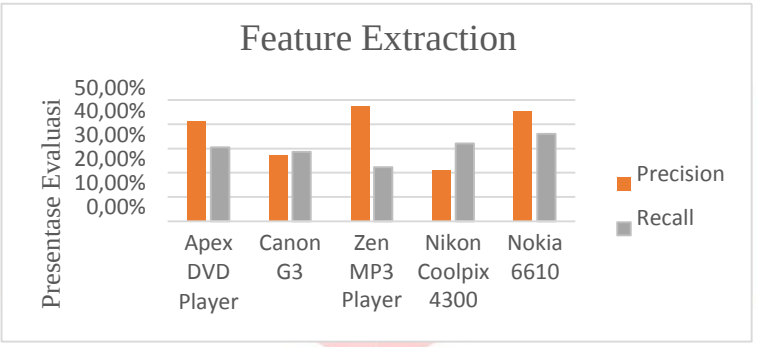
3 Pengujian dan Analisis

Penelitian pada jurnal ini menggunakan *dataset* pada *review* lima produk elektronik: 2 kamera digital, 1 DVD player, 1 MP3 player, dan 1 telepon seluler, yang diambil dari situs jual beli *online* Amazon.com. *Dataset* yang digunakan sudah dalam format **.txt.* Tabel 1 berikut ini adalah rincian *dataset* dari masing-masing produk :

Tabel 1 Rincian dataset

Nama Dataset	Rincian Data
Apex AD2600 Progressive-scan DVD Player	739 kalimat
Canon G3	597 kalimat
Creative Labs Nomad Jukebox Zen Xtra 40GB	1716 kalimat
Nikon Coolpix 4300	346 kalimat
Nokia 6610	546 kalimat

3.1 Analisis Frequent Feature Extraction



Gambar 2 Evaluasi hasil ekstraksi frequent feature

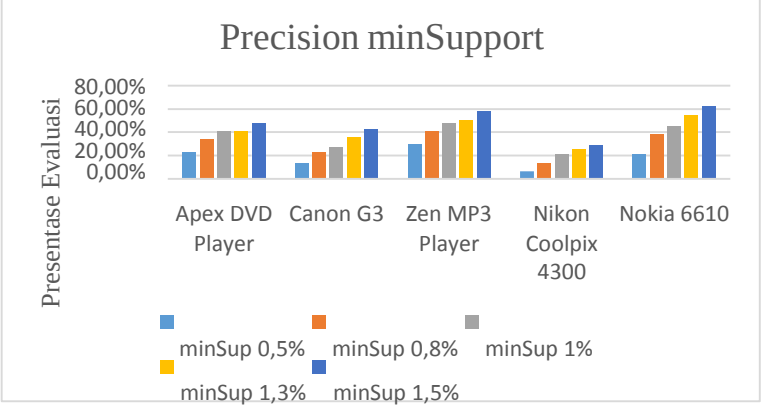
Berdasarkan Gambar 2 di atas dapat dilihat bahwa nilai *precision* untuk data *Apex DVD Player*, *Zen MP3 Player*, dan *Nokia 6610* berada pada kisaran angka 40% sampai 50%, sedangkan data *Canon G3* dan *Nikon Coolpix 4300* berada pada kisaran angka 20% sampai 30%. Nilai *recall* untuk data *Apex DVD Player*, *Canon G3*, dan *Zen MP3 Player* berada pada kisaran angka 20% sampai 30%, sedangka *recall* data *Nikon Coolpix 4300* dan *Nokia 6610* berada pada kisaran angka 30% sampai 40%.

Tinggi rendahnya nilai *precision* dan *recall* dari ekstraksi fitur dipengaruhi oleh beberapa faktor diantaranya adalah persebaran fitur, nilai *minimum support apriori* yang digunakan sebesar 1%, dan jenis fitur eksplisit yang hanya bisa ditangani oleh sistem. Sedangkan fitur produk yang tercantum pada *dataset* terdapat dua jenis fitur, yaitu fitur eksplisit dan fitur implisit. Sistem tidak bisa mengidentifikasi fitur implisit karena diperlukan penanganan kata ganti benda.

Proses identifikasi fitur yang *infrequent* dilakukan dengan menggunakan prosedur sederhana yaitu dengan cara mencari *noun* atau *noun phrase* yang terletak di sekitar *opinion word*, kemudian dilakukan pengecekan apakah *noun* atau *noun phrase* tersebut merupakan *frequent feature* atau bukan, apabila bukan merupakan *frequent feature* maka diidentifikasi sebagai *infrequent feature*. Proses identifikasi *infrequent feature* pada sistem ini menghasilkan fitur-fitur yang tidak relevan karena kalimat yang kompleks dimana terdapat lebih dari satu fitur dan *opinion word* yang posisinya beragam.

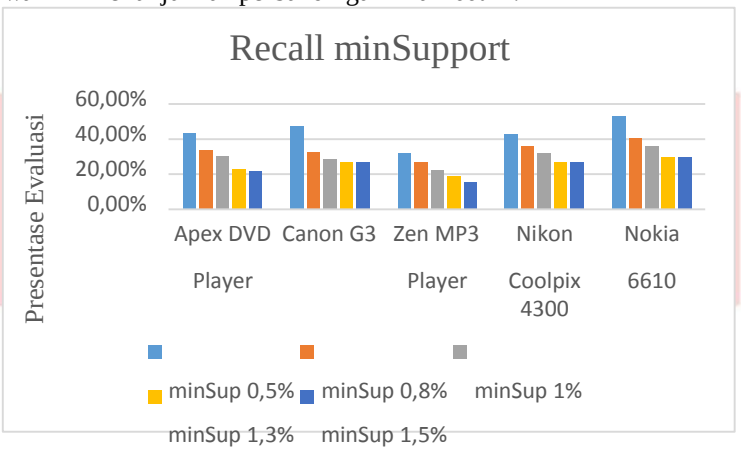
3.2 Analisis Nilai Threshold untuk Minimum Support

Proses identifikasi *frequent feature* dilakukan dengan menggunakan algoritam apriori yang terdapat parameter *minimum support* dalam proses di dalamnya. Pada tahap ini dilakukan pengujian terhadap parameter *minimum support* dengan rentang nilai mulai 0,5% sampai dengan 1,5%. Pada sistem ini digunakan *minimum support* sebesar 1% yang dipakai ke proses selanjutnya. Gambar 3 di bawah ini dapat dilihat hasil *precision* yang diperoleh :



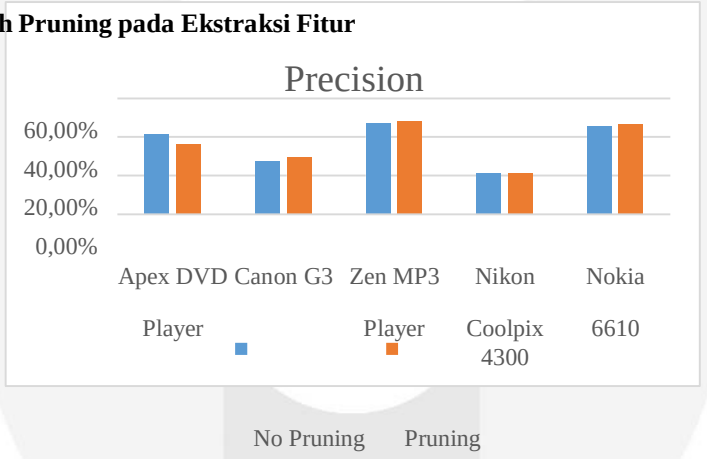
Gambar 3 Precision nilai threshold minimum support

Berdasarkan gambar di atas dapat kita lihat bahwa pada penggunaan *minimum support* 0,5% dan 0,8% dihasilkan nilai *precision* yang lebih rendah daripada penggunaan *minimum support* yang lainnya. Seharusnya semakin rendah *minimum support* yang dipakai maka hasil *frequent itemset* yang diperoleh menjadi lebih banyak, namun pada kasus ini jumlah *frequent itemset* yang dihasilkan tidak sebanding dengan jumlah fitur yang relevan dengan fitur produk yang terdapat pada *dataset* sehingga nilai *precision* yang dihasilkan menjadi rendah.
Hal tersebut akan berkebalikan dengan hasil nilai *recall* yang didapat. Pada penggunan *minimum support* yang lebih tinggi maka *frequent items* yang diidentifikasi jumlahnya semakin sedikit dan karena jumlah *frequent items* yang dihasilkan jumlahnya sedikit maka jumlah fitur yang relevan dengan fitur produk yang terdapat pada *dataset* juga semakin sedikit. Sehingga nilai *recall* cenderung menurun jika *minimum support* yang digunakan semakin tinggi. Gambar 4 di bawah ini menunjukkan perbandingan nilai *recall* :



Gambar 4 Recall nilai threshold minimum support

3.3 Analisis Pengaruh Pruning pada Ekstraksi Fitur

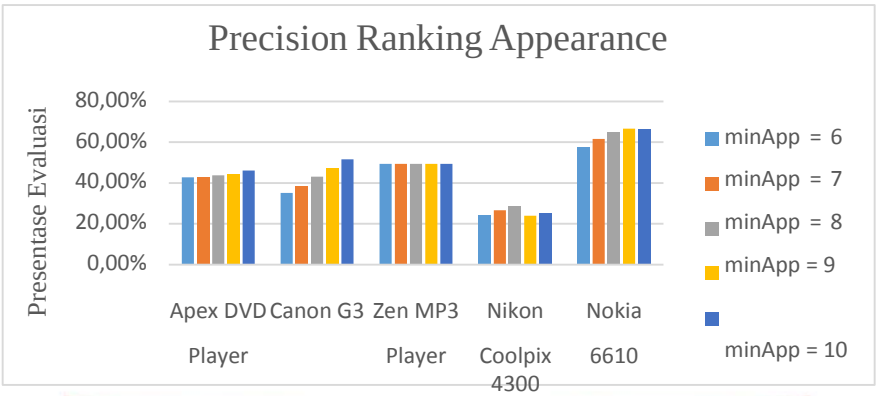


Gambar 5 Pengaruh pruning pada frequent feature

Dari Gambar 5 dapat ditarik kesimpulan bahwa penggunaan metode *pruning* dapat berpengaruh menentukan fitur produk yang relevan. Nilai *precision* setelah proses *pruning* untuk data *Canon G3, Zen MP3 Player, Nikon Coolpix 4300,* dan *Nokia 6610* mengalami kenaikan. Untuk data *Apex DVD Player* nilai *precision* nya mengalami penurunan karena fitur relevan yang sebelumnya sudah teridentifikasi menggunakan *association rule mining* menjadi terhapus oleh *pruning*.

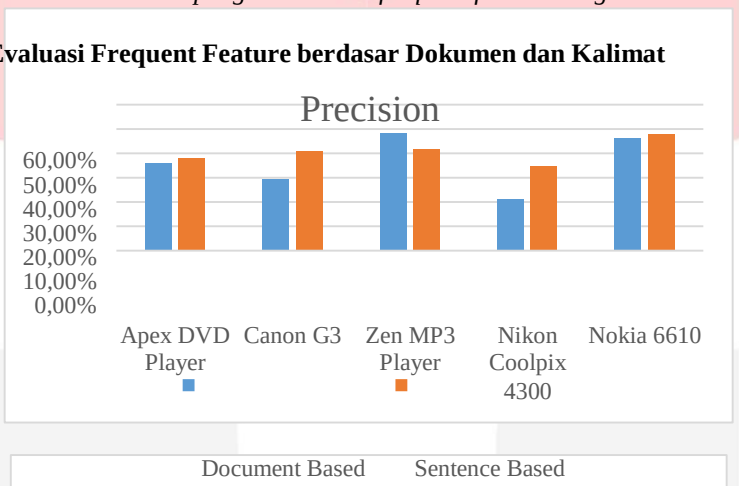
3.4 Analisis Threshold Minimum Appearance pada Frequent Fitur

Dari gambar 6 di bawah dapat dilihat bahwa nilai *precision* untuk semua data baik *Apex DVD Player, Canon G3, Zen MP3 Player, Nikon Coolpix 4300,* dan *Nokia 6610* mengalami peningkatan seiring dengan bertambahnya nilai parameter *minimum appearance.*
Meningkatnya nilai *precision* untuk seleksi berdasarkan *minimum appearance* ini dipengaruhi oleh jumlah fitur yang terekstrak semakin berkurang diiringi dengan bertambahnya jumlah fitur yang relevan. Jumlah fitur relevan yang terekstrak semakin banyak karena fitur dengan kemunculan tinggi besar kemungkinannya merupakan fitur produk yang sesuai dengan fitur produk pada *dataset*.

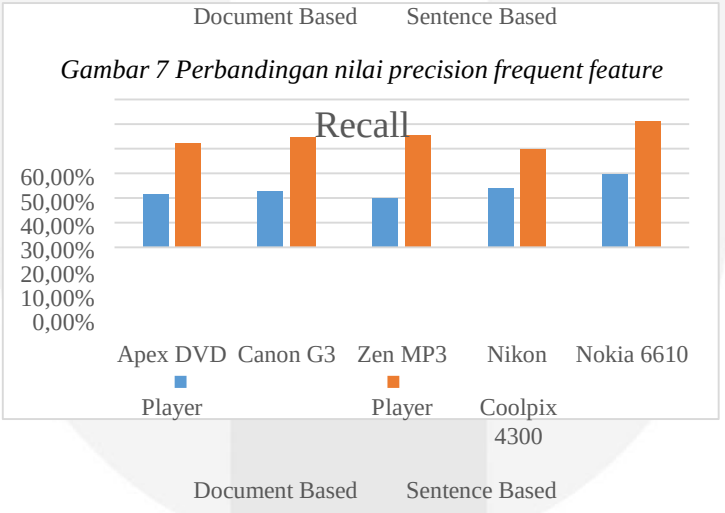


Gambar 6 Precision pengaruh seleksi frequent feature dengan kemunculan

3.5 Perbandingan Evaluasi Frequent Feature berdasar Dokumen dan Kalimat



Gambar 7 Perbandingan nilai precision frequent feature

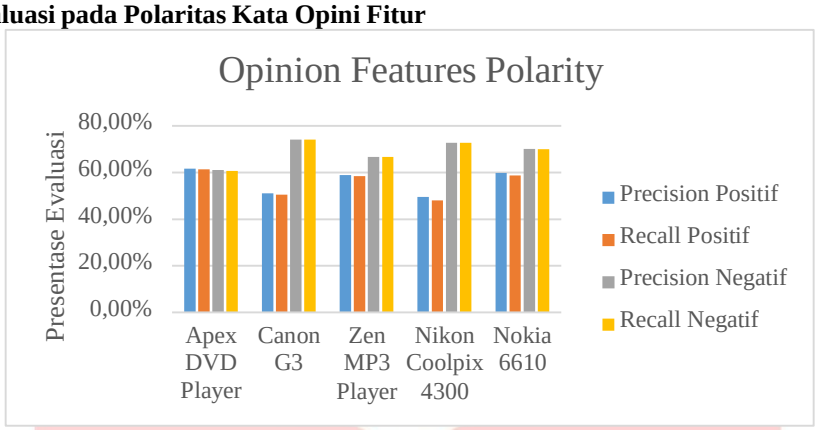


Gambar 8 Perbandingan nilai recall frequent feature

Perbandingan metode evaluasi ini menunjukkan peningkatan yang dapat dilihat dari Gambar 7 dan Gambar 8 di atas. Semua *dataset* hampir semuanya mengalami kenaikan kecuali pada *precision* data *Zen MP3 Player*.

Hasil ekstraksi *frequent feature* yang dihasilkan oleh *association rule mining* hanya berupa sekumpulan *vocab item* dievaluasi dengan cara dicocokkan ke fitur produk yang terdapat pada *dataset*. *Frequent feature* hasil *association rule mining* tersebut kemudian dikembalikan ke setiap kalimat, kemudian setiap kalimat tersebut dievaluasi per kalimat dengan acuan kalimat yang ada pada *dataset* (pada *dataset* terdapat fitur kunci atau label fitur yang terletak sebelum tanda ##). Evaluasi *precision* dan *recall* setiap kalimat kemudian dicari nilai rata-ratanya, dengan metode evaluasi per kalimat ini di dapat hasil nilai *precision* dan *recall* yang lebih bagus.

3.6 Analisis Evaluasi pada Polaritas Kata Opini Fitur



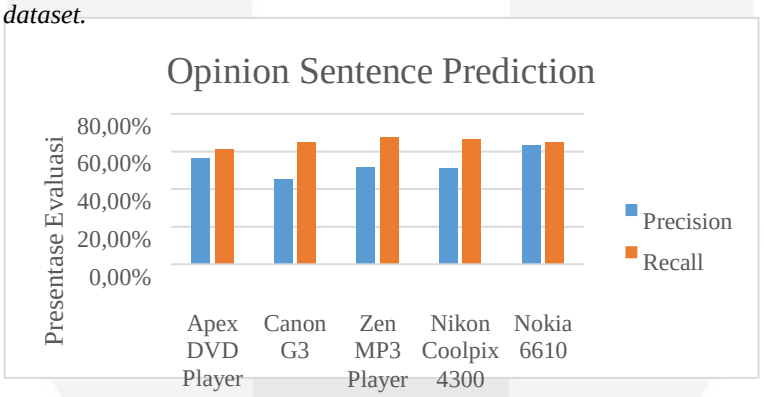
Gambar 9 Evaluasi polaritas opini fitur

Hasil pengujian *precision* dan *recall* pada polaritas opini fitur dapat dilihat pada Gambar 9. Nilai *precision* dan *recall* berada pada rentang 30% sampai dengan 45%. Data *Nokia 6610* mempunyai nilai *precision* dan *recall* paling tinggi diantara data yang lainnya yaitu sebesar 42,28% untuk *precision* dan 44,58% untuk *recall*. Sedangkan data *Nikon Coolpix 4300* mempunyai nilai *precision 33,60%* dan nilai *recall* sebesar 36,95% yang menempatkannya sebagai data dengan *precision* dan *recall* paling rendah diantara data yang lainnya.

Tinggi rendahnya nilai *precision* dan nilai *recall* pada pengujian ini dipengaruhi oleh *feature set* yang sudah didapat, persebaran kata opini yang kompleks pada kalimat, dan penggunaan *SentiWordNet*. Polaritas opini fitur didapatkan dari bobot nilai kata opini pasangan fitur tersebut yang sudah dilakukan pembobotan menggunakan *SentiWordNet*. Apabila kata opini yang bobot nilainya nol atau netral maka fitur pasangannya akan dihapus.

3.7 Analisis Evaluasi pada Prediksi Kalimat Opini

Tahapan ini dilakukan pengujian terhadap kalimat opini yang sudah diekstraksi oleh sistem. Pada sistem ini yang dimaksud kalimat opini adalah sebuah kalimat yang mengandung fitur beserta polaritasnya baik itu positif atau negatif. Sistem memprediksi sebuah kalimat opini yang benar apabila kalimat opini yang diekstraks sistem relevan dengan kalimat pada *dataset.*



Gambar 10 Evaluasi prediksi kalimat opini

Berdasarkan Gambar 10 di atas dapat dilihat hasil evaluasi prediksi kalimat opini yang dilakukan sistem. Data *Apex DVD Player*, *Zen MP3Player*, *Nikon Coolpix 4300*, dan *Nokia 6610* mempunyai nilai *precision* dan *recall* berada pada rentang 50% sampai dengan 70%. Sedangkan data *Canon G3* mempunyai nilai *precision* 45,59% dan *recall* 64,85%. Hal tersebut menandakan bahwa sistem sudah cukup baik dalam memprediksi kalimat opini.

4 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang sudah dilakukan maka dapat ditarik beberapa kesimpulan diantaranya adalah penggunaan metode *association rule mining* mampu mengidentifikasi fitur-fitur yang *frequent*, seleksi kata dengan menggunakan metode *pruning* dan dengan meningkatkan nilai *minimum appearance* pada *frequent feature* dapat membuang fitur yang tidak relevan. Karakteristik *dataset* seperti persebaran fitur, jenis fitur, dan *misspelling* sangat mempengaruhi proses ekstraksi fitur. Proses ekstraksi kata opini dipengaruhi oleh jumlah *feature set* yang sudah didapatkan pada proses ekstraksi fitur sebelumnya. Hasil polaritas fitur dipengaruhi oleh bobot nilai opini dari kamus *SentiWordNet.* Selain itu penggunaan metode evaluasi berbasis kalimat pada pengecekan ekstraksi fitur menghasilkan nilai yang lebih tinggi daripada menggunakan evaluasi berbasis dokumen.

Untuk pengembangan lebih lanjut sebaiknya diimplementasikan *coreference*, *fuzzy matching*, dan penanganan identifikasi *noun phrase* sebelum dilakukan proses ekstraksi sehingga didapatkan hasil yang lebih optimal.

Daftar Pustaka

[1] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," 2004.

[2] T. Korenius, J. Laurikkala, K. Jarvelin and Martti Juhola, "Stemming and Lemmatization in the Culustering of Finnish Text Documents".

[3] G. Ingersoll, T. Morton and A. Farris, Taming Text, Shelter Island, NY: Manning Publications Co, 2008.

[4] S. S. Htay and K. T. Lyn, "Extracting Product Featrues and Opinion Words Using Pattern Knowledge in Customer Reviews," *The Scientific World Journal*, 2013.

[5] S. H. Ghorashi, R. Ibrahim, S. Noekhah and N. Dstjerdi, "A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews," *IJCSI International Journal of Computer Science Issues*, vol. 9, 2012.

[6] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," 2004.

[7] Stanford NLP Group, "The Stanford Parser: A statistical parser," Stanford NLP Group, 27 08 2014. [Online]. Available: http://nlp.stanford.edu/software/lex-parser.shtml. [Accessed May 2015].

[8] S. Baccianella, A. Esuli and F. Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentimen Analysis and Opinion Mining".

[9] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules in Large Databases," 1994.

[10] S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, 1st Ed., O'Reilly, 2009.

[11] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd Ed., The Morgan Kaufmann Series in Data Management Systems, 2006.

[12] B. Liu, W. Hsu and Y. Ma, "Integrating Classification and Association Rule Mining," 1998.

[13] B. Liu, "Opinion Mining," Chicago.

[14] D. Olson and D. Delen, Advanced Data Mining Techniques, 1st Ed., Berlin: Springer, 2008.