KLASIFIKASI OPINI PADA FITUR PRODUK BERBASIS *GRAPH*

OPINION CLASSIFICATION FOR PRODUCT FEATURE BASED ON GRAPH

I Kadek Bayu Arys Wisnu Kencana[1], Warih Maharani S.T., M.T.[2]

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

[1]aryswisnu@student.telkomuniversity.ac.id, [2]wmaharani@gmail.com

Abstrak

Produk merupakan sesuatu yang ditawarkan oleh produsen kepada konsumen, di mana setiap produk memiliki opini yang berbeda-beda bagi setiap konsumen. Opini produk yang berbeda-beda tersebut dapat berupa opini positif maupun negatif. Untuk menganalisis opini yang berupa opini positif atau negatif secara otomatis, diperlukan adanya suatu sistem yang dapat digunakan oleh produsen maupun konsumen. Pada tugas akhir ini, dilakukan penelitian terhadap klasifikasi opini dengan menggunakan nilai kemiripan *(similarity)* antar kata dengan menggunakan dua pendekatan berbasis *graph* yaitu Word2Vec dan WordNet. Word2Vec merupakan representasi kata dalam bentuk vektor yang digunakan untuk menghasilkan *word embeddings* [1]. Sedangkan WordNet merupakan sebuah *database* kamus bahasa Inggris yang memiliki hirarki keterhubungan antar kata melalui jalur yang dimilikinya [2]. Penelitian tugas akhir ini menunjukkan bahwa hasil klasifikasi opini fitur produk menggunakan Word2Vec memiliki persentase akurasi yang lebih tinggi jika dibandingkan dengan klasifikasi opini fitur produk dengan menggunakan WordNet dengan rata-rata selisih persentase akurasi dari 6 *dataset* yaitu 2.07%. Hal tersebut disebabkan karena pada pendekatan Word2Vec, kosakata *(vocabulary)* yang dimiliki dapat dikembangkan sendiri berdasarkan data *training* yang digunakan. Sedangkan pada WordNet, kumpulan kata yang terdapat pada *corpus* merupakan data dari WordNet itu sendiri, jadi tidak dapat dikembangkan sendiri seperti pada Word2Vec.

Kata kunci : ***sentiment analysis, word2vec, wordnet, graph-based classification, word embeddings.***

Abstract

Products are something that producers offer to consumers, which each product has a different opinion for each consumer. Different product opinions can be either positive or negative. To analyze opinions in the form of positive or negative opinions automatically, it needs a system that can be used by producers and consumers. In this final project, conducted a research on the classification of opinion by using similarity values between words by using two graphical approach i.e. Word2Vec and WordNet. Word2Vec is a word representation in vector form that is used to generate word embeddings [1]. While WordNet is an English dictionary database that has a hierarchy of connectivity between words through the path it has [2]. This research shows that the opinion classification for product feature using Word2Vec has a higher percentage of accuracy when compared with the opinion classification for product feature using WordNet with average percentage difference of accuracy in 6 datasets is 2.07%. That is because in the Word2Vec approach, owned vocabulary can be developed alone based on used training data. While in the WordNet, the collection of words contained in the corpus are data from WordNet itself, so it can't be developed on its own like in Word2Vec.

Keywords: sentiment analysis, word2vec, wordnet, graph-based classification, word embeddings.

1. Pendahuluan

Produk merupakan sesuatu yang ditawarkan oleh produsen terhadap konsumen. Setiap produk yang dikeluarkan oleh produsen, konsumen berhak memberikan opini atau pendapat mengenai suatu produk tersebut. Pendapat atau opini oleh konsumen tersebut dapat dijadikan tolak ukur bagi konsumen maupun produsen. Sebagai contoh, calon konsumen akan melihat terlebih dahulu opini-opini dari orang lain mengenai suatu produk sebelum membelinya. Berbeda dengan konsumen, produsen dapat menjadikan opini sebagai tolak ukur untuk evaluasi produk yang telah diproduksi [3]. Akan tetapi banyaknya opini yang ada pada suatu produk membuat konsumen maupun produsen kesulitan dalam menemukan opini mana yang termasuk opini positif dan opini mana yang termasuk opini negatif. Selain itu, konsumen maupun produsen kesulitan untuk menemukan fitur dari suatu produk mana yang paling banyak mendapat opini positif maupun negatif sebagai bahan evaluasi mereka. Maka dari itu diperlukan adanya sistem yang dapat menangani hal tersebut.

Pada penelitian tugas akhir kali ini, akan dibangun sistem yang dapat menentukan polaritas opini dari suatu fitur produk berdasarkan nilai kemiripan atau *similarity* suatu kata. Sistem yang akan dibangun menggunakan pendekatan Word2Vec untuk proses klasifikasi opini dimana Word2Vec merupakan model representasi kata dalam

bentuk vektor, dan pada penelitian tugas akhir ini Word2Vec digunakan untuk melakukan klasifikasi opini pada fitur suatu produk [1]. Pendekatan Word2Vec dipilih pada penelitian tugas akhir ini sebab Word2Vec dapat mempelajari hubungan antar kata-kata secara otomatis dan dapat mempelajari berbagai macam *text corpus* sebagai data latih [4]. Word2Vec juga telah digunakan dalam beberapa penelitian terkait *sentiment analysis* seperti pada riset yang melakukan penelitian terkait *sentiment* pada Twitter [5], serta pada penelitian sentiment *movie review* berikut [6].

Penggunaan model Word2Vec pada penelitian tugas akhir ini juga diterapkan pada pengelompokkan fitur produk atau *aspect grouping*, sehingga fitur produk tertentu dapat dikelompokkan dan dapat diketahui jumlah opini *positive* maupun opini *negative* dari suatu fitur produk tersebut. Selain menggunakan pendekatan Word2Vec, penelitian tugas akhir ini juga menggunakan pendekatan WordNet pada klasifikasi opini. Hal ini dilakukan karena kedua pendekatan tersebut sama-sama memiliki nilai kemiripan atau *similarity* antar kata yang berbasis *graph* yang akan digunakan pada proses klasifikasi opini pada penelitian tugas akhir ini..

2. Kajian Pustaka

2.1 Text Mining

Text mining atau penambangan teks adalah sebuah teknologi yang dapat digunakan untuk menambah data yang ada pada basis data yang *corporate* dengan membuat data teks yang tidak terstruktur agar dapat dianalisis. Dalam perihal menganalisis respon *free-form survey*, *text mining* digunakan untuk mengelompokan respon yang unik menjadi kategori-kategori dari sejumlah tanggapan, sebagai solusi dalam mengurangi banyaknya tanggapan yang bersifat individu agar menjadi kumpulan yang lebih kecil tetapi masih tetap mewakili tanggapan yang diberikan [8].

2.2 Sentiment Analysis

Sentiment analysis atau analisis sentimen adalah ilmu komputasi yang mempelajari mengenai pendapat, sentimen, dan emosi yang diekspresikan ke dalam teks [9]. Sistem *sentiment analysis* sedang diterapkan di hampir setiap bisnis dan wilayah sosial karena opini merupakan pusat dari setiap aktifitas manusia dan dapat sangat memengaruhi perilaku manusia. Kepercayaan manusia dan persepsi tentang sesuatu, serta pilihan yang dibuat, sebagian besar berasal dari bagaimana orang lain melihat dan menilai sesuatu. Semakin pentingnya *sentiment analysis* bertepatan pula dengan pertumbuhan media sosial seperti ulasan atau *review*, forum diskusi, *blog*, *micro-blog*, Twitter, dan *social network* lainnya [10].

2.3 Double Propagation

Double propagation adalah pendekatan yang digunakan untuk melakukan ekstraksi kata opini maupun kata fitur produk *(targets)* secara iteratif menggunakan kata opini dan kata fitur produk yang telah diketahui dan telah diekstrak (pada iterasi sebelumnya) melalui identifikasi hubungan *syntactic*. Identifikasi hubungan pada metode double propagation terbagi menjadi 3 hubungan yaitu hubungan antara kata opini dan kata fitur produk *(OT-Rel)*, hubungan antar kata opini itu sendiri *(OO-Rel)*, serta hubungan antar kata fitur produk yaitu *(TT-Rel). Double propagation* memiliki 4 subtugas dalam melakukan ekstraksi, diantaranya [11]: (1) ekstraksi fitur produk menggunakan kata opini; (2) ekstraksi kata fitur produk menggunakan kata fitur produk yang telah terekstrak; (3) ekstraksi kata opini menggunakan kata fitur produk yang telah terekstrak; (4) ekstraksi kata opini menggunakan kata opini yang telah diberikan dan yang telah terekstrak.. Subtugas tersebut digunakan pada hubungan-hubungan yang terdapat pada pendekatan double propagation, dimana subtugas (1) dan (3) menggunakan hubungan *OT-Rel*, dan subtugas (2) menggunakan hubungan *TT-Rel*, sedangkan subtugas (4) menggunakan *OO-Rel.*

.

2.4 Stemming

Stemming merupakan suatu teknik yang melakukan proses pengurangan akhiran *(suffix)* suatu kata yang mengubah kata tersebut menjadi kata dasar atau menjadi bentuk kamus dari suatu kata [12]. Jika fungsi pada *stemming* dipanggil oleh sistem, maka akan dilakukan pengecekan kata yang akan dilakukan *stemming* dan selanjutnya akan mengikuti sekumpulan *rules.* Pertama, fungsi *stemmer*s akan menghapus semua *stopwords* (misalnya daftar kata-kata yang akan diabaikan pada sistem). Tahap selanjutnya akan menghapus kata-kata yang memiliki akhiran seperti kata dalam bentuk *plural* atau jamak (seperti: *-s, -es*), *past tense* (seperti: *-ed*), dan *continuous tenses* (seperti: *-ing*). *Stemmer*s juga akan melakukan pengecekan dan mengubah akhiran seperti *-ic, -full, -ness, -ant, -ence.*

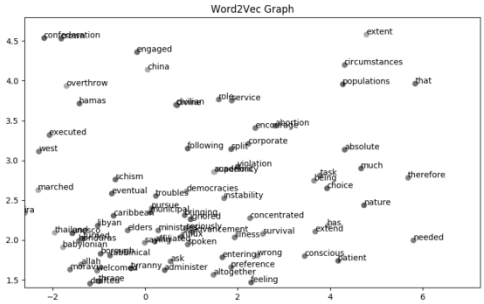
2.5 Graph-based

Proses klasifikasi yang akan dilakukan pada penelitian tugas akhir ini menggunakan nilai kemiripan atau *similarity* antar kata yang berbasis grafik atau *graph.* Proses klasifikasi berbasis *graph* yang digunakan pada penelitian tugas akhir ini menggunakan dua pendekatan, yaitu Word2Vec dan WordNet.

2.5.1. Word2Vec

Word2Vec merupakan representasi kata dalam bentuk vektor yang dibuat oleh Google [4]. Word2Vec juga merupakan sekumpulan beberapa model yang saling berkaitan yang digunakan untuk menghasilkan *word embeddings*. *Word embeddings* merupakan sebutan dari seperangkat bahasa pemodelan dan teknik pembelajaran fitur atau *feature learning* pada *Natural Language Processing (NLP)* dimana setiap kata dari kosakata (*vocabulary)* memiliki vektor yang mewakili makna dari kata tersebut dan kata-kata tersebut dipetakan ke dalam bentuk vektor bilangan riil.

Word2vec menggunakan sekumpulan teks yang besar sebagai data latih *(training)* untuk membangun *vocabulary* dan menghasilkan ruang vektor yang dapat berjumlah beberapa ratus dimensi, dengan setiap kata unik pada *corpus* berupa vektor dimana pembentukan vektor tersebut menerapkan model Skip-gram dan model CBOW *(Continous Bag-of-Words)* [1]. Vektor kata diposisikan pada ruang vektor sedemikian rupa sehingga kata-kata yang berbagi konteks umum di dalam *corpus*, terletak berdekatan satu sama lain di dalam ruang vektor [13]. Gambar 1 merupakan contoh visualisasi ruang vektor pada Word2Vec:



Gambar 1 Visualisasi ruang vektor terhadap data yang telah dilatih **[14]**

Berdasarkan visualisasi di atas, kata-kata yang mirip saling berkumpul secara berdekatan satu sama lain. Pencarian nilai kemiripan atau *similarity* atau kemiripan berdasarkan vektor dapat dilakukan dengan menggunakan perhitungan *cosine similarity*. Perhitungan *cosine distance* merupakan perhitungan yang biasa digunakan untuk menentukan kemiripan antar objek dengan mencari jarak antar kata yang telah berbentuk vektor. *Cosine* dihitung diawali dengan perhitungan *dot-product* dari dua vektor yang telah dinormalisasi. Normalisasi tersebut biasanya berbentuk Euclidean, yaitu nilai dinormalisasi ke dalam vektor satuan panjang Euclidean. Pada nilai positif, *cosine* berkisar antara 0 sampai 1 [15]. Dengan dua vektor yaitu vektor a dan verktor b sebagai contoh, perhitungan *cosine similarity* diantara kedua vektor tersebut dihitung sebagai berikut:
dot-product memiliki formula [16]

$$\vec{a} \odot \vec{b} = \sum_{i=1}^{N} \vec{a}_i \vec{b}_i, \qquad (1)$$

normalisasi jika didefinisikan pada formula akan menjadi [16]

$$||\vec{x}|| = \sqrt{\vec{x} \cdot \vec{x}}, \qquad (2)$$

dan perhitungan *cosine* yang didefinisikan pada formula [16]

$$cosine(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \times ||\vec{b}||}, \qquad (3)$$

lalu jika formula (2.1) dan (2.2) digunakan maka akan menghasilkan [16]

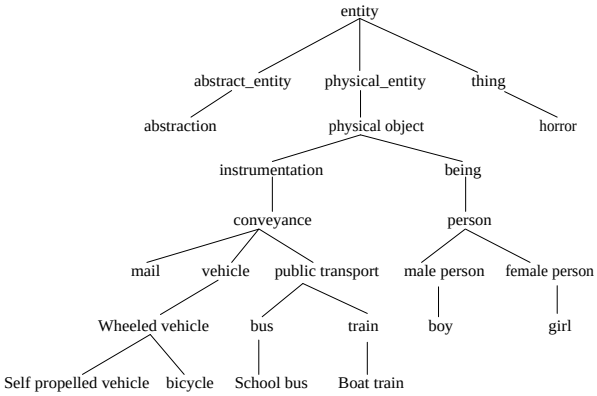
$$cosine(\vec{a}, \vec{b}) = \frac{\sum_{i=1}^{N} \vec{a}_i \vec{b}_i}{\sqrt{\sum_{i=1}^{N} \vec{a}_i^2} \sqrt{\sum_{i=1}^{N} \vec{b}_i^2}} \qquad (4)$$

maka dengan perhitungan-perhitungan tersebut, nilai kemiripan atau *similarity* untuk Word2Vec dihasilkan.

2.5.2. WordNet

WordNet merupakan sebuah *database* kamus bahasa Inggris. Berbeda dengan kamus pada umumnya, WordNet mengelompokkan kata benda *(nouns)*, kata kerja *(verbs)*, kata sifat *(adjectives)*, dan kata keterangan *(adverbs)* ke dalam kumpulan sinonim atau makna yang dimilki suatu kata yang saling memiliki keterkaitan

(synsets) secara semantik [2]. Bagian dari relasi *is-a* antar kata pada WordNet akan ditunjukan pada Gambar 2-5 berikut [17]:



Gambar 2 Potongan dari *taxonomy* relasi *is-a* pada WordNet **[17]**

Pada *taxonomy* tersebut, dapat ditunjukan bahwa jika posisi suatu kata pada *taxonomy* berada semakin di bawah maka makna kata tersebut semakin spesifik, sedangkan jika posisi suatu kata pada taxonomy berada semakin di atas maka makna kata tersebut semakin abstrak [17]. Hasil yang diperoleh dari hubungan atau relasi *hyponym/hypernym* secara umum berupa nilai kemiripan atau *similarity (similarity)* antar kata. Pada penelitian tugas akhir ini, metode perhitungan nilai *similarity* yang akan digunakan dengan WordNet adalah perhitungan berbasis *path (path based)* dengan menggunakan jalur terpendek antar kata yang saling berhubungan. Formula yang akan digunakan untuk mencari nilai *similarity* WordNet berdasarkan jalur terpendek adalah seperti ini [17]:

$$sim_{path}(a,b) = \frac{1}{len(a,b)} \qquad (5)$$

dimana $len(a,b)$ pada formula di atas menunjukkan jarak terpendek antara kata a dengan kata b pada WordNet [17].

2.6. Akurasi

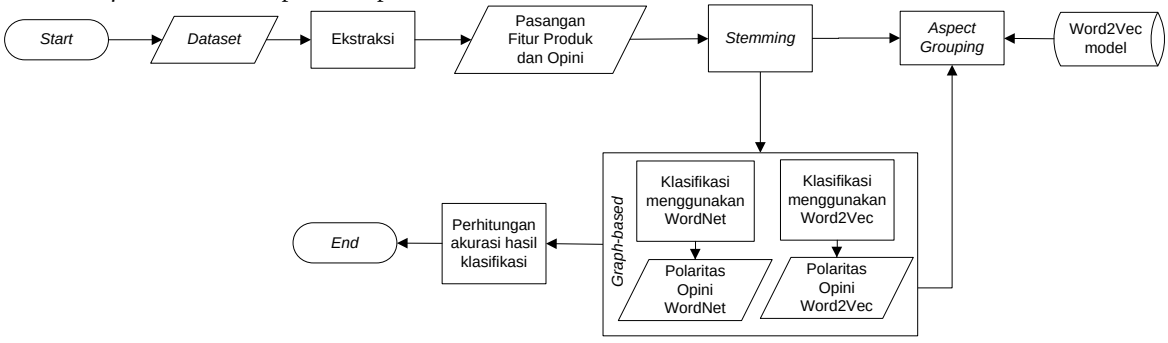
Dalam proses klasifikasi yang akan dilakukan nantinya, akan dihitung performansi dari masing-masing pendekatan yang digunakan berupa nilai akurasi. Nilai akurasi didapat dari jumlah kata yang terklasifikasi dengan benar jika dibandingkan dengan expert judgement pada dataset, lalu dibagi dengan semua jumlah klasifikasi. Berikut formula perhitungan performansi berupa nilai akurasi yang digunakan pada penelitian tugas akhir ini [18]:

$$accuracy = \frac{N\ (correct\ classifications)}{N\ (all\ classifications)} \qquad (6)$$

3. Perancangan Sistem

3.1. Gambaran Umum Sistem

Gambaran umum sistem divisualisasikan dalam bentuk *flowchart*, di mana dalam *flowchart* tersebut terdapat beberapa proses yang secara umum menggambarkan sistem yang akan dikembangkan. Gambaran umum sistem dalam bentuk *flowchart* ditampilkan seperti Gambar 3 berikut:



Gambar 3 Gambaran Umum Sistem

3.2. Perancangan Sistem

3.2.1. Dataset

Dataset yang digunakan merupakan review produk dalam bahasa inggris dengan format .txt yang pernah digunakan pada penelitian dengan judul paper "Mining and Summarizing Customers Reviews" dari Minqing Hu dan Bing Liu [7].

3.2.2. Ekstraksi

Setiap kalimat pada *dataset* akan diterapkan proses ekstraksi terlebih dahulu sebelum melalui proses klasifikasi. Proses ekstraksi yang menggunakan *double propagation* di sini dilakukan agar kalimat-kalimat pada *dataset* dapat ter-ekstrak dan menghasilkan pasangan opini dan fitur produk yang nantinya akan digunakan dalam proses klasifikasi.

3.2.3. Klasifikasi

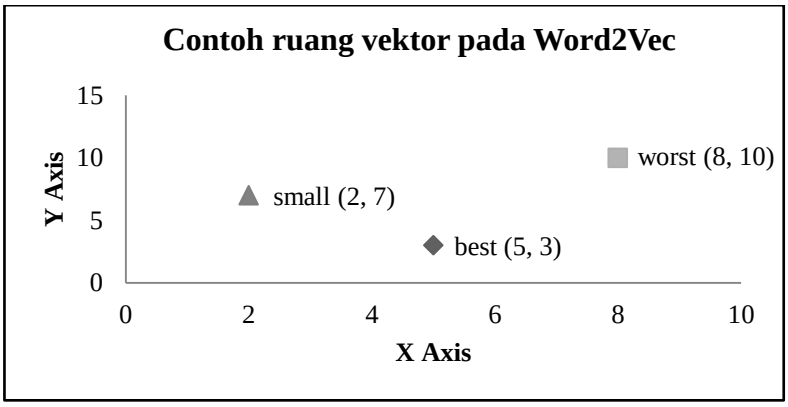
3.2.3.1. Stemming

Stemming merupakan proses yang dapat mengubah suatu kata menjadi kata dasar dengan menghilangkan semua imbuhan, dimana proses ini diterapkan pada kandidat opini dan fitur produk yang telah didapat. Penerapan proses *stemming* pada tahap ini hanya menghilangkan imbuhan yang bersifat jamak *(plural)* sebab adanya imbuhan tersebut memengaruhi akurasi dari klasifikasi nantinya.

3.2.3.2. Word2Vec

Klasifikasi yang akan dilakukan pada penelitian tugas akhir ini menggunakan pendekatan Word2Vec dengan menggunakan nilai kemiripan atau *similarity* dari Word2Vec sebagai acuan penentuan polaritas. Sebelum melakukan pencarian kemiripan menggunakan metode Word2Vec, dibutuhkan *training* data corpus menjadi *model* terlebih dahulu. Model yang dimaksud merupakan pemodelan suatu *corpus* yang akan diubah ke dalam bentuk vektor, sehingga nantinya nilai kemiripan atau *similarity* yang akan digunakan dalam klasifikasi merupakan hasil pemodelan tersebut. Pada penelitian tugas akhir ini, data *corpus* yang akan dilatih *(training)* dan dimodelkan ke dalam bentuk vektor adalah data text8 [19].

1. Pada contoh berikut, digunakan kata opini yaitu *"small"* yang akan dibandingkan dengan kedua kata pembanding yaitu *"best"* dan *"worst"*. Contoh visualisasi bentuk ruang vektor pada Word2Vec terhadap tiga sampel kata yaitu *"small"* sebagai kata opini, serta *"best"* dan *"worst"* sebagai kata pembanding:



Gambar 4 Contoh visualisasi ruang vektor pada Word2Vec terhadap tiga sampel kata opini

2. Berdasarkan contoh di atas, tiga kata sampel dalam bentuk vektor tersebut dilakukan perhitungan *dot-product* terlebih dahulu dengan menggunakan vektor masing-masing kata. Kata best dan worst merupakan kata pembanding, maka akan terdapat dua perhitungan *dot-product* yang akan dilakukan yaitu dengan membandingkan kata *("small","best")* dan *("small","worst")* dengan masing-masing vektor sebagai berikut:
   $\overrightarrow{small} = (2,7)$
   $\overrightarrow{best} = (5,3)$
   $\overrightarrow{worst} = (8,10)$
   Berikut contoh perhitungan *dot-product* antar vektor tersebut:

$\overrightarrow{small} \odot \overrightarrow{best} = 2 \cdot 5 + 7 \cdot 3 = 10 + 21 = \mathbf{31}$
$\overrightarrow{small} \odot \overrightarrow{worst} = 2 \cdot 8 + 7 \cdot 10 = 16 + 70 = \mathbf{86}$
Lalu nilai *dot-product* tersebut selanjutnya digunakan untuk perhitungan *cosine distance* sebagai nilai kemiripan atau *similarity*, berikut contohnya:

$$cosine\left(\overrightarrow{small}, \overrightarrow{best}\right) = \frac{31}{\sqrt{2^2 + 7^2}\sqrt{5^2 + 3^2}} = \frac{31}{\sqrt{1802}} = \mathbf{0.73027}$$

$$cosine\left(\overrightarrow{small}, \overrightarrow{worst}\right) = \frac{86}{\sqrt{2^2 + 7^2}\sqrt{8^2 + 10^2}} = \frac{43}{\sqrt{2173}} = \mathbf{0.92244}$$

3. Berdasarkan perhitungan tersebut dapat diketahui bahwa nilai kemiripan atau *similarity* antara kata *("small","worst")* lebih tinggi dibandingkan dengan nilai kemiripan atau *similarity* antara kata *("small","best")*, dengan begitu dapat disimpulkan bahwa kata opini *"small"* termasuk dalam polaritas *negative*. Hal ini dibuktikan dengan nilai kemiripan atau *similarity* kata *"small"* yang lebih tinggi pada kata *"worst"* jika dibandingkan dengan kata *"best"*, yang mana kata *"worst"* tersebut merupakan kata yang mengandung konotasi negatif.

3.2.3.3. WordNet

WordNet Pada penelitian tugas akhir ini digunakan pula pencarian nilai kemiripan atau *similarity* dengan pendekatan WordNet menggunakan *path based*, dimana nilai kemiripan atau *similarity* dari WordNet akan digunakan pula sebagai acuan penentuan polaritas. Pencarian nilai kemiripan atau *similarity* yang akan diterapkan menggunakan jarak terpendek antar kata dan untuk menggunakan data pada WordNet tidak dibutuhkan proses *training* data terlebih dahulu sehingga data WordNet dapat langsung digunakan untuk pencarian nilai kemiripan atau *similarity*. Berikut merupakan beberapa tahapan dalam proses pencarian nilai kemiripan atau *similarity* hingga penentuan polaritas opini dengan pendekatan WordNet yang akan diterapkan pada penelitian tugas akhir ini:

1. Jarak yang dimiliki antar kata pada WordNet akan digunakan untuk menghasilkan nilai kemiripan atau *similarity*. Berdasarkan pemaparan pada bab 2.6 mengenai WordNet, untuk mencari nilai kemiripan atau *similarity* dari kata menggunakan perhitungan jarak yang terdekat [22]. Pada contoh berikut, digunakan kata opini yaitu *"small"* yang akan dibandingkan dengan kedua kata pembanding yaitu *"best"* dan *"worst"*.

2. Kata opini yaitu *"small"* akan dicari nilai kemiripan atau *similarity*-nya berdasarkan jarak kata tersebut terhadap dua kata pembanding yang digunakan yaitu *"best"* dan *"worst"*. Pertama-tama dilakukan pencarian nilai $len(small, best)$ dengan menghitung jarak terpendek antara *"small"* dengan *"best"*, dimana terdapat dua jalur yang memungkinkan kata *"small"* untuk bertemu dengan kata *"best"* begitu juga sebaliknya, yaitu:

a. Jalur pertama *(small, best)* = *small → body_part → part → thing → physical_entity → causal_agent → person → best* = **8**

b. Jalur kedua *(small, best)* = *small → body_part → part → thing → physical_entity → object → whole → living_thing → organism → person → best* = **11**

Jalur terpendek antara *"small"* dengan *"best"* setelah dilakukan penelusuran yaitu jalur pertama dengan jarak = 8. Maka nilai $len(small, best) = \mathbf{8}$. Setelah dilakukan pencarian nilai $len(small, best)$ maka selanjutnya dilakukan pencarian $len(small, worst)$ dengan menghitung jarak terpendek antara *"small"* dengan *"worst"*, dimana hanya terdapat satu jalur yang memungkinkan untuk kata *"small"* bertemu kata *"worst"* begitu juga sebaliknya, yaitu:

a. Jalur *(small, worst)* = *small → body_part → part → thing → physical_entity → entity → abstraction → attribute → quality → immorality → evil → worst* = **12**

Jalur terpendek antara *"small"* dengan *"worst"* setelah dilakukan penelusuran merupakan jalur satu-satunya dengan jarak = 12. Maka nilai $len(small, worst) = \mathbf{12}$.

3. Berdasarkan jarak terpendek dari masing-masing sampel kata maka akan dilakukan perhitungan nilai kemiripan atau *similarity* antara kata opini *"small"* dengan kata pembanding best dan worst menggunakan formula (2.5). Berikut perhitungan nilai kemiripan atau *similarity* antara sampel kata opini dengan kedua kata pembanding:

$$sim_{path}(small, best) = \frac{1}{8} = \mathbf{0.125}$$

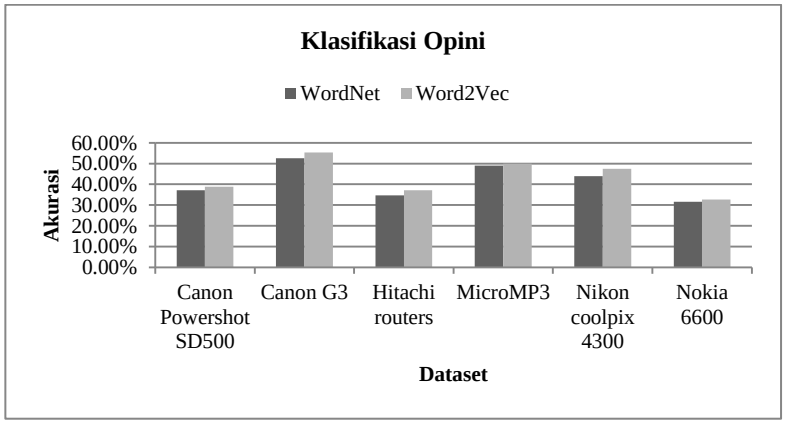
$$sim_{path}(small, worst) = \frac{1}{12} = \mathbf{0.0833}$$

4. Berdasarkan perhitungan tersebut dapat diketahui bahwa nilai kemiripan atau *similarity* antara kata *"small"*-*"best"* lebih tinggi dibandingkan dengan nilai kemiripan atau *similarity* antara kata *"small"*-*"worst"*, dengan begitu dapat disimpulkan bahwa kata opini *"small"* termasuk dalam polaritas *positive*. Hal ini dibuktikan dengan nilai kemiripan atau *similarity* kata *"small"* yang lebih tinggi pada kata *"best"* jika dibandingkan dengan kata *"worst"*, yang mana kata *"best"* tersebut merupakan kata yang mengandung konotasi positif.

4. Pengujian dan Analisis

4.1. Analisis Pengaruh dan Perbandingan Penggunaan WordNet dan Penggunaan Word2Vec terhadap Proses Klasifikasi Opini

Analisis pada pengaruh serta perbandingan penggunaan WordNet dan penggunaan Word2Vec terhadap proses klasifikasi opini dilakukan untuk mengetahui tingkat akurasi dari kedua metode tersebut terhadap hasil klasifikasi opini. Pengujian dilakukan dengan mengimplementasi kedua metode tersebut terhadap hasil ekstraksi yang telah melalui proses *stemming* dari setiap *dataset*. Berikut hasil akurasi kedua metode kaslifikasi opini yang dibandingkan, yaitu WordNet dan Word2Vec:



Gambar 5 Grafik analisis perbandingan WordNet dan Word2Vec terhadap proses klasifikasi opini

Berdasarkan Gambar 5, dapat ditunjukan perbandingan persentase akurasi antara proses klasifikasi dengan menggunakan WordNet dengan proses klasifikasi dengan menggunakan Word2Vec. Detail hasil akurasi dari grafik di atas dapat dilihat pada tabel berikut:

Tabel 1 Persentase akurasi perbandingan WordNet dan Word2Vec terhadap proses klasifikasi opini

Dataset	WordNet	Word2Vec	Selisih
Canon Powershot SD500	37,12%	38,86%	1,75%
Canon G3	52,60%	55,28%	2,68%
Hitachi routers	34,62%	37,18%	2,56%
MicroMP3	49,01%	49,90%	0,89%
Nikon coolpix 4300	43,93%	47,40%	3,47%
Nokia 6600	31,59%	32,67%	1,08%
Rata-rata	**41,48%**	**43,55%**	**2,07%**

Pada Tabel 1 di atas terdapat selisih persentase akurasi yang menunjukkan hasil klasifikasi menggunakan Word2Vec dengan hasil klasifikasi menggunakan WordNet memiliki perbedaan akurasi dengan rata-rata 2,07%. Selisih diantara kedua pendekatan pada proses klasifikasi opini tersebut setelah dilakukan analisis disebabkan oleh adanya pengaruh kata saat pencarian nilai kemiripan atau *similarity* antar kata opini, dimana terdapat kata opini yang nilai *similarity*-nya dapat dicari saat menggunakan Word2Vec tetapi tidak dapat dicari saat menggunakan WordNet sehingga menghasilkan polaritas *"null"*.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Penerapan pendekatan Word2Vec pada proses klasifikasi memiliki nilai akurasi yang lebih tinggi dibandingkan penerapan pendekatan WordNet, yaitu dengan rata-rata persentase akurasi pada Word2Vec sebesar **43,55%**

sedangkan rata-rata persentase akurasi pada WordNet sebesar **41,48%,** dengan selisih **2,07%**. Hal tersebut disebabkan karena pada pendekatan Word2Vec, kosakata *(vocabulary)* yang dimiliki dapat dikembangkan sendiri berdasarkan data *training* yang digunakan. Sedangkan pada WordNet, kumpulan kata yang terdapat pada corpus merupakan data dari WordNet itu sendiri, jadi tidak dapat dilatih seperti pada Word2Vec. Selain itu, pencarian nilai kemiripan atau *similarity* pada pendekatan WordNet dengan perhitungan *shortest path distance* hanya dapat dilakukan jika kata yang akan dibandingkan merupakan kata benda *(nouns)* atau kata kerja *(verbs).*

5.2. Saran

Saran yang dapat dijadikan bahan penelitian untuk pengembangan penelitian ini adalah sebagai berikut:

1. Menangani kata negasi dan implisit.
2. Menerapkan beberapa metode dalam proses ekstraksi.
3. Menggunakan data *training* selain text8 dalam membangun model Word2Vec.
4. Menerapkan model *word embeddings* lain seperti FastText, WordRank, dll.

Daftar Pustaka:

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ICLR Workshop,* 2013.

[2] "WordNet A Lexical Database for English," Princeton University, [Online]. Available: http://wordnet.princeton.edu. [Accessed 2016].

[3] D. K. Li, K. Sugiyama, Z. Lin and M.-Y. Kan, "Product Review Summarization based on Facet Identification and Sentence Clustering".

[4] T. Mikolov, "word2vec," Google, 30 July 2013. [Online]. Available: https://code.google.com/archive/p/word2vec/. [Accessed 23 November 2016].

[5] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu and B. Qin, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* vol. 1: Long Papers, 2014.

[6] H. Pouransari and S. Ghili, "Deep learning for sentiment analysis of movie reviews," Technical report, Standford University, 2014.

[7] M. Hu and B. Liu, "Mining and Summarizing Customers Reviews," *KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* 2004.

[8] L. Francis, "Text Mining Handbook," 2010.

[9] F. Li, M. Huang and X. Zhu, "Sentiment analysis with global topics and local dependency," *AAAI '10,* p. 1371–1376, 2010.

[10] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

[11] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," *Association for Computational Linguistics,* Vols. Volume 37, Number 1, 2011.

[12] V. B. a. E. Lloyd-Yemoh, "Stemming and Lemmatization: A Comparison of Retrieval," *Lecture Notes on Software Engineering,* 2014.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS),* 2013.

[14] TensorFlow, "Vector Representations of Words," Google, [Online]. Available: https://www.tensorflow.org. [Accessed 9 7 2017].

[15] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, David Pinto, "Soft Similarity and Soft Cosine Measure," vol. 18, 2014.

[16] G. Salton, Automatic Text Processing, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1988.

[17] Lingling Meng, Runqing Huang and Junzhong Gu, "A Review of Semantic Similarity Measures in WordNet," *International Journal of Hybrid Information Technology,* vol. 6, no. 1, 2013.

[18] P. P. Alexander Pak, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Proceedings of the International Conference on Language Resources and Evaluation,* 2010.

[19] M. Mahoney, "About the Test Data," 1 September 2011. [Online]. Available: http://mattmahoney.net. [Accessed 12 January 2017].