GRAPH-BASED OPINION CLASSIFICATION FOR PRODUCT FEATURES

Original title: Klasifikasi Opini pada Fitur Produk Berbasis Graph

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

[1,2]Undergraduate Program in Informatics Engineering, School of Informatics, Telkom University
Bandung, Indonesia

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

About this document. This is an English translation of an undergraduate final project paper originally written and published in Indonesian in 2017. The original appeared as "Klasifikasi Opini pada Fitur Produk Berbasis Graph" in e-Proceeding of Engineering, Vol. 4, No. 2, August 2017, pages 3148 to 3155, ISSN 2355-9365, Telkom University, Bandung, Indonesia. The translation reproduces the content, methods, examples, numerical results and references of the 2017 paper. No result, method or reference has been added, updated or reinterpreted. It is not a new 2026 publication.

Abstract

A product is something that a producer offers to consumers, and every product attracts a different opinion from every consumer. These differing product opinions may be either positive or negative. In order to analyse whether an opinion is positive or negative automatically, a system is required that can be used by producers as well as by consumers. In this final project, research was carried out on opinion classification using the similarity value between words by means of two graph-based approaches, namely Word2Vec and WordNet. Word2Vec is a representation of words in vector form that is used to produce word embeddings [1]. WordNet, in contrast, is an English lexical database that possesses a hierarchy of relatedness between words through the paths it contains [2]. This research shows that the classification of opinions on product features using Word2Vec attains a higher accuracy percentage than the classification of opinions on product features using WordNet, with an average accuracy difference across 6 datasets of 2.07%. This is because, in the Word2Vec approach, the vocabulary held by the model can be developed by the researcher on the basis of the training data used, whereas in WordNet the collection of words contained in the corpus is the data of WordNet itself and therefore cannot be developed independently as it can in Word2Vec.

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

1. Introduction

A product is something that a producer offers to consumers. For every product released by a producer, consumers are entitled to give an opinion or a judgement concerning that product. Such judgements or opinions expressed by consumers can serve as a point of reference both for consumers and for producers. For example, a prospective consumer will first look at the opinions of other people about a product before buying it. Producers, by contrast, can treat opinions as a point of reference for the evaluation of the products they have already manufactured [3]. However, the sheer number of opinions attached to a product makes it difficult for consumers and producers alike to determine which opinions count as positive opinions and which count as negative ones. In addition, consumers and producers find it difficult to identify which feature of a product receives the largest number of positive or negative opinions as material for their evaluation. A system that can handle this problem is therefore required.

In this final project research, a system will be built that can determine the opinion polarity of a product feature on the basis of the similarity value of a word. The system to be built uses the Word2Vec approach for the opinion classification process, where Word2Vec is a model that represents words in vector form; in this final project research, Word2Vec is used to perform opinion classification on the features of a product [1]. The Word2Vec approach was chosen for this final project research because Word2Vec can learn the relationships between words automatically and can learn from a wide variety of text corpora used as training data [4]. Word2Vec has also been used in several

studies related to sentiment analysis, such as research conducted on sentiment on Twitter [5], and the following research on sentiment in movie reviews [6].

The use of the Word2Vec model in this final project research is also applied to the grouping of product features, or aspect grouping, so that particular product features can be grouped together and the number of positive opinions and negative opinions belonging to a given product feature can be determined. Besides using the Word2Vec approach, this final project research also uses the WordNet approach for opinion classification. This is done because both approaches equally possess a graph-based similarity value between words, which will be used in the opinion classification process in this final project research.

2. Literature Review

2.1 Text Mining

Text mining is a technology that can be used to augment the data held in a corporate database by rendering unstructured textual data analysable. In the matter of analysing free-form survey responses, text mining is used to group unique responses into categories drawn from a body of responses, as a solution for reducing the large number of individual responses into a smaller collection that nevertheless still represents the responses that were given [8].

2.2 Sentiment Analysis

Sentiment analysis is the computational discipline that studies opinions, sentiments and emotions expressed in text [9]. Sentiment analysis systems are being applied in almost every business and social domain, because opinion is at the centre of every human activity and can strongly influence human behaviour. Human beliefs and perceptions about something, as well as the choices that are made, derive for the most part from how other people see and evaluate that thing. The growing importance of sentiment analysis also coincides with the growth of social media such as reviews, discussion forums, blogs, micro-blogs, Twitter and other social networks [10].

2.3 Double Propagation

Double propagation is an approach used to extract opinion words as well as product feature words (targets) iteratively, using opinion words and product feature words that are already known and have already been extracted (in a previous iteration) through the identification of syntactic relations. The identification of relations in the double propagation method is divided into 3 relations, namely the relation between an opinion word and a product feature word (OT-Rel), the relation between opinion words themselves (OO-Rel), and the relation between product feature words (TT-Rel). Double propagation has 4 subtasks for performing extraction, namely [11]: (1) extraction of product features using opinion words; (2) extraction of product feature words using product feature words that have already been extracted; (3) extraction of opinion words using product feature words that have already been extracted; (4) extraction of opinion words using opinion words that have been supplied and that have already been extracted. These subtasks are used on the relations contained in the double propagation approach, where subtasks (1) and (3) use the OT-Rel relation, subtask (2) uses the TT-Rel relation, and subtask (4) uses OO-Rel.

2.4 Stemming

Stemming is a technique that carries out the process of removing the suffix of a word, changing that word into its root form or into the dictionary form of a word [12]. If the stemming function is invoked by the system, a check is performed on the word to be stemmed and a set of rules is then followed. First, the stemmer function will remove all stopwords (for example, the list of words that the system is to ignore). The next stage removes words that carry endings such as the plural form (for example: -s, -es), the past tense (for example: -ed), and the continuous tenses (for example: -ing). Stemmers will also check and modify endings such as -ic, -full, -ness, -ant, -ence.

2.5 Graph-Based Approach

The classification process to be carried out in this final project research uses the graph-based similarity value between words. The graph-based classification process used in this final project research employs two approaches, namely Word2Vec and WordNet.

2.5.1 Word2Vec

Word2Vec is a representation of words in vector form created by Google [4]. Word2Vec is also a collection of interrelated models used to produce word embeddings. Word embeddings is the name given to a set of language modelling and feature learning techniques in Natural Language Processing (NLP) in which every word of the vocabulary has a vector that represents the meaning of that word, and those words are mapped into vectors of real numbers.

Word2Vec uses a large collection of text as training data in order to build a vocabulary and to produce a vector space that may amount to several hundred dimensions, with every unique word in the corpus taking the form of a vector, where the construction of that vector applies the Skip-gram model and the CBOW (Continuous Bag-of-Words) model [1]. Word vectors are positioned in the vector space in such a way that words which share a common context within the corpus are located close to one another within the vector space [13]. Figure 1 is an example of a visualisation of the vector space in Word2Vec:

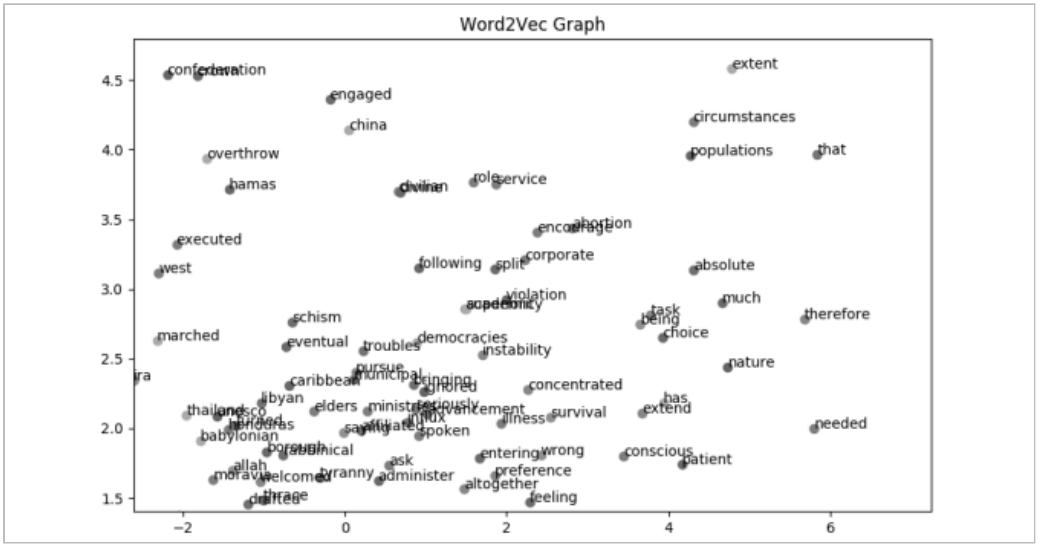

Figure 1 Visualisation of the vector space for data that has been trained **[14]**

On the basis of the visualisation above, words that are similar cluster close to one another. The search for the similarity value on the basis of vectors can be performed using the cosine similarity computation. The cosine distance computation is the computation ordinarily used to determine the similarity between objects by finding the distance between words that have been put into vector form. The cosine is computed beginning with the computation of the dot-product of two vectors that have been normalised. That normalisation ordinarily takes the Euclidean form, that is, the value is normalised into a unit vector of Euclidean length. For positive values, the cosine ranges between 0 and 1 [15]. Taking two vectors, vector a and vector b, as an example, the cosine similarity between those two vectors is computed as follows:

the dot-product has the formula [16]

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

normalisation, when defined as a formula, becomes [16]

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

and the cosine computation that is defined as a formula [16]

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

then, if formulas (1) and (2) are used, the result will be [16]

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

and so, with those computations, the similarity value for Word2Vec is produced.

2.5.2 WordNet

WordNet is an English lexical dictionary database. Unlike dictionaries in general, WordNet groups nouns, verbs, adjectives and adverbs into sets of synonyms, or senses possessed by a word, that are semantically related to one another (synsets) [2]. A part of the is-a relation between words in WordNet is shown in Figure 2 below [17]:

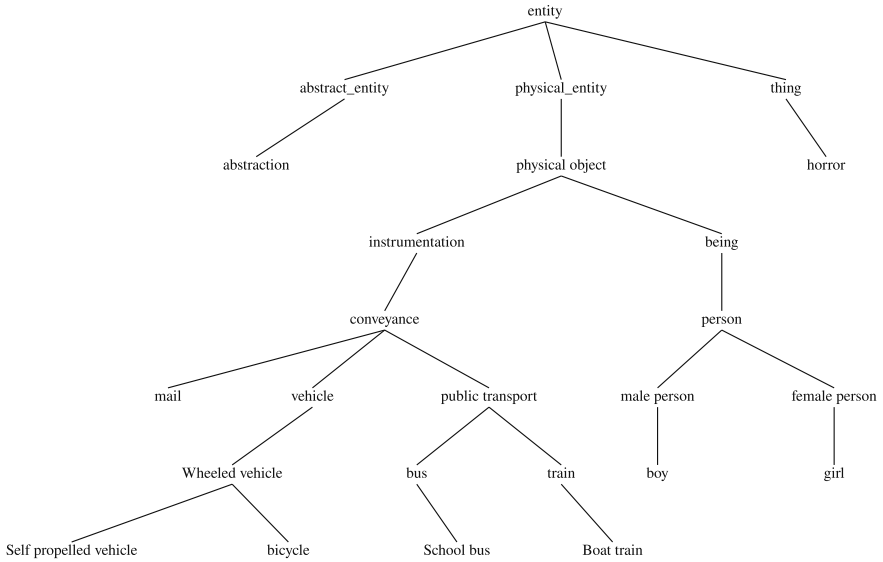

Figure 2 An excerpt from the taxonomy of the is-a relation in WordNet **[17]**

In that taxonomy it can be shown that the further down the taxonomy the position of a word lies, the more specific the meaning of that word becomes, whereas the further up the taxonomy the position of a word lies, the more abstract the meaning of that word becomes [17]. The result obtained from the hyponym and hypernym relation is in general a similarity value between words. In this final project research, the method for computing the similarity value that will be used with WordNet is the path based computation, using the shortest path between words that are related to one another. The formula that will be used to find the WordNet similarity value on the basis of the shortest path is as follows [17]:

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

where *len(a, b)* in the formula above indicates the shortest distance between word *a* and word *b* in WordNet [17].

2.6 Accuracy

In the classification process that will be carried out subsequently, the performance of each approach used will be computed in the form of an accuracy value. The accuracy value is obtained from the number of words that are correctly classified when compared with the expert judgement in the dataset, and is then divided by the total number of classifications. The following is the formula for computing performance in the form of the accuracy value that is used in this final project research [18]:

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

3. System Design

3.1 General Overview of the System

The general overview of the system is visualised in the form of a flowchart, in which there are several processes that describe in general terms the system that is to be developed. The general overview of the system in flowchart form is presented as in Figure 3 below:

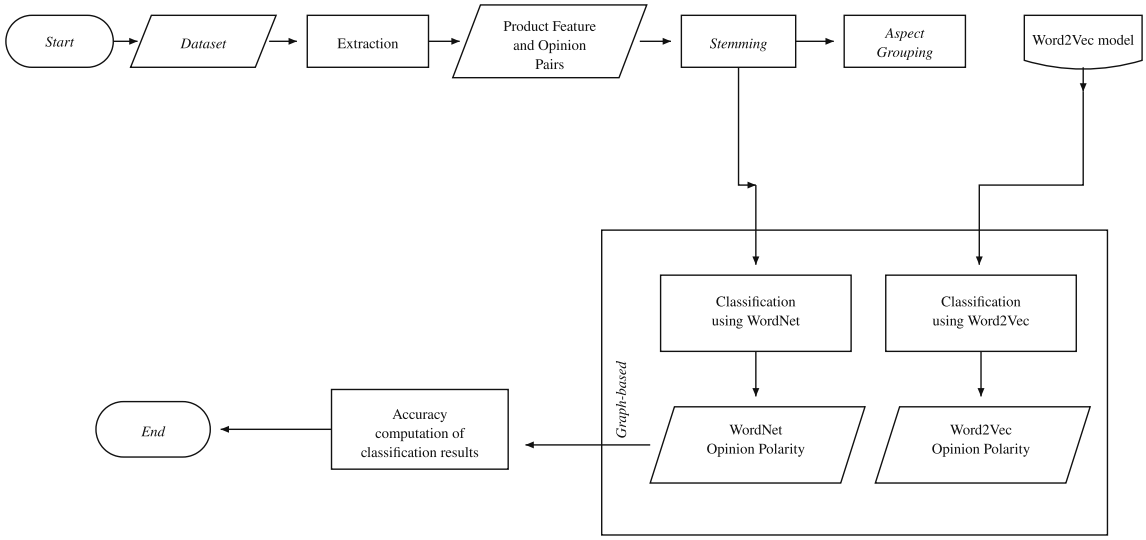

Figure 3 General Overview of the System

3.2 System Design

3.2.1 Dataset

The dataset used consists of product reviews in English in .txt format that were previously used in the research published under the paper title "Mining and Summarizing Customers Reviews" by Minqing Hu and Bing Liu [7].

3.2.2 Extraction

Every sentence in the dataset has the extraction process applied to it first, before it passes through the classification process. The extraction process, which here uses double propagation, is carried out so that the sentences in the dataset can be extracted and can yield pairs of opinions and product features that will subsequently be used in the classification process.

3.2.3 Classification

3.2.3.1 Stemming

Stemming is a process that can change a word into its root form by removing all affixes, and this process is applied to the candidate opinions and product features that have been obtained. The application of the stemming process at this stage removes only affixes of the plural type, since the presence of such affixes affects the accuracy of the subsequent classification.

3.2.3.2 Word2Vec

The classification that will be carried out in this final project research uses the Word2Vec approach, employing the similarity value from Word2Vec as the reference for determining polarity. Before searching for similarity using the Word2Vec method, the corpus data must first be trained into a model. The model in question is the modelling of a corpus that will be converted into vector form, so that the similarity value that will subsequently be used in classification is the result of that modelling. In this final project research, the corpus data that will be trained and modelled into vector form is the text8 data [19].

1. In the following example, the opinion word "small" is used and will be compared with the two comparison words "best" and "worst". An example visualisation of the form of the vector space in Word2Vec for three sample words, namely "small" as the opinion word and "best" and "worst" as the comparison words, is as follows:

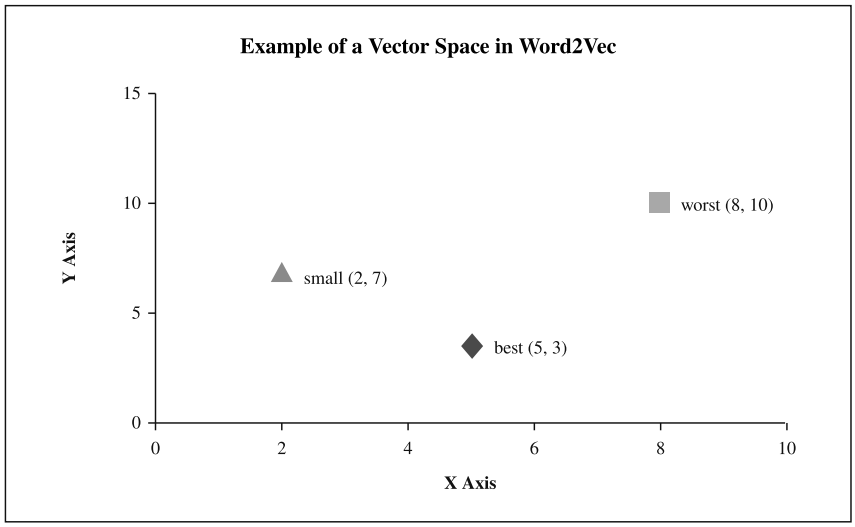

Figure 4 Example visualisation of the vector space in Word2Vec for three sample opinion words

2. On the basis of the example above, the three sample words in vector form first have the dot-product computed using the vectors of each word. The words best and worst are the comparison words, and so there will be two dot-product computations to be carried out, namely by comparing the words ("small", "best") and ("small", "worst"), with the respective vectors as follows:

   $\overline{small} = (2,7)$

   $\overline{best} = (5,3)$

   $\overline{worst} = (8,10)$

   The following is an example of the dot-product computation between those vectors:

   $\overline{small} \odot \overline{best} = 2 \cdot 5 + 7 \cdot 3 = 10 + 21 = \mathbf{31}$

   $\overline{small} \odot \overline{worst} = 2 \cdot 8 + 7 \cdot 10 = 16 + 70 = \mathbf{86}$

The dot-product values are then used for the computation of the cosine distance as the similarity value; the following is an example:

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

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

3. On the basis of that computation it can be seen that the similarity value between the words ("small", "worst") is higher than the similarity value between the words ("small", "best"), and it may therefore be concluded that the opinion word "small" falls under negative polarity. This is demonstrated by the fact that the similarity value of the word "small" is higher for the word "worst" than for the word "best", where the word "worst" is a word that carries a negative connotation.

3.2.3.3 WordNet

In this final project research, the search for the similarity value using the WordNet approach with path based computation is also used, where the similarity value from WordNet will likewise be used as the reference for determining polarity. The search for the similarity value that will be applied uses the shortest distance between words, and in order to use the data in WordNet no data training process is required beforehand, so that the WordNet data can be used directly for the search for the similarity value. The following are the stages in the process of searching for the similarity value up to the determination of opinion polarity with the WordNet approach that will be applied in this final project research:

1. The distance held between words in WordNet will be used to produce the similarity value. On the basis of the exposition in Section 2.6 concerning WordNet, in order to find the similarity value of a word the computation of the nearest distance is used [22]. In the following example, the opinion word "small" is used and will be compared with the two comparison words "best" and "worst".
2. The opinion word "small" will have its similarity value determined on the basis of the distance of that word from the two comparison words used, namely "best" and "worst". First, a search is made for the value *len(small, best)* by computing the shortest distance between "small" and "best", where there are two possible paths by which the word "small" may meet the word "best" and vice versa, namely:
   a. *First path (small, best) = small → body_part → part → thing → physical_entity → causal_agent → person → best = **8***
   b. *Second path (small, best) = small → body_part → part → thing → physical_entity → object → whole → living_thing → organism → person → best = **11***

   The shortest path between "small" and "best" after tracing is the first path, with distance = 8. Thus the value *len(small, best)* = **8**. After the search for the value *len(small, best)* has been carried out, a search is then made for *len(small, worst)* by computing the shortest distance between "small" and "worst", where there is only one possible path by which the word "small" may meet the word "worst" and vice versa, namely:
   a. *Path (small, worst) = small → body_part → part → thing → physical_entity → entity → abstraction → attribute → quality → immorality → evil → worst = **12***

   The shortest path between "small" and "worst" after tracing is the sole path, with distance = 12. Thus the value *len(small, worst)* = **12**.
3. On the basis of the shortest distance for each sample word, the similarity value between the opinion word "small" and the comparison words best and worst will be computed using formula (5). The following is the computation of the similarity value between the sample opinion word and the two comparison words:

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

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

4. On the basis of that computation it can be seen that the similarity value between the words "small" and "best" is higher than the similarity value between the words "small" and "worst", and it may therefore be concluded that the opinion word "small" falls under positive polarity. This is demonstrated by the fact that the similarity value of the word "small" is higher for the word "best" than for the word "worst", where the word "best" is a word that carries a positive connotation.

4. Testing and Analysis

4.1 Analysis of the Effect and Comparison of the Use of WordNet and the Use of Word2Vec on the Opinion Classification Process

The analysis of the effect and the comparison of the use of WordNet and the use of Word2Vec on the opinion classification process was carried out in order to determine the accuracy level of the two methods with respect to the opinion classification results. Testing was carried out by implementing the two methods on the extraction results that had passed through the stemming process for each dataset. The following are the accuracy results of the two opinion classification methods that are compared, namely WordNet and Word2Vec:

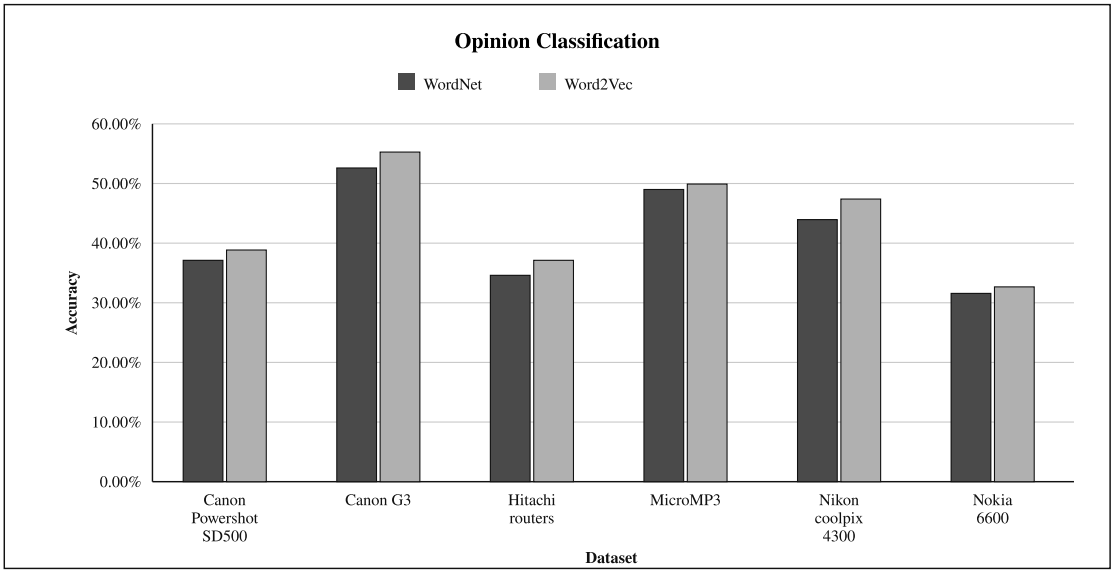

Figure 5 Graph of the comparative analysis of WordNet and Word2Vec with respect to the opinion classification process

On the basis of Figure 5, the comparison of the accuracy percentage between the classification process using WordNet and the classification process using Word2Vec can be shown. The detailed accuracy results from the graph above can be seen in the following table:

Table 1 Accuracy percentage of the comparison of WordNet and Word2Vec with respect to the opinion classification process

Dataset	WordNet	Word2Vec	Difference
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%
Average	**41.48%**	**43.55%**	**2.07%**

In Table 1 above there is a difference in accuracy percentage which shows that the classification results using Word2Vec and the classification results using WordNet differ in accuracy by an average of 2.07%. After analysis, the difference between the two approaches in the opinion classification process is caused by the influence of words during the search for the similarity value between opinion words, where there are opinion words whose similarity value can be found when using Word2Vec but cannot be found when using WordNet, thereby producing a "null" polarity.

5. Conclusion and Recommendations

5.1 Conclusion

The application of the Word2Vec approach in the classification process attains a higher accuracy value than the application of the WordNet approach, namely with an average accuracy percentage for Word2Vec of **43.55%** while the average accuracy percentage for WordNet is **41.48%**, with a difference of **2.07%**. This is because, in the Word2Vec approach, the vocabulary held by the model can be developed by the researcher on the basis of the training data used. In WordNet, by contrast, the collection of words contained in the corpus is the data of WordNet itself, and therefore cannot be trained as it can in Word2Vec. In addition, the search for the similarity value in the WordNet approach with the shortest path distance computation can only be carried out if the words to be compared are nouns or verbs.

5.2 Recommendations

Recommendations that may serve as material for research to develop this study further are as follows:

1. Handling negation words and implicit words.
2. Applying several methods in the extraction process.
3. Using training data other than text8 in building the Word2Vec model.
4. Applying other word embeddings models such as FastText, WordRank, and so on.

References

[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].