Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study

by Adi Wibowo

Submission date: 23-Jan-2023 04:59PM (UTC+0700)
Submission ID: 1997629181
File name: 1-s2.0-S1877050919310968-main.pdf (552.97K)
Word count: 2934
Character count: 16384

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 157 (2019) 360–366

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019

Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study

Rizka Putri Nawangsari[a], Retno Kusumaningrum[a*], Adi Wibowo[a]

[a]Department of Informatics Universitas Diponegoro, Jl. Prof. Soedarto SH Tembalang, Semarang 50275, Indonesia

Abstract

Hand-crafted features engineering is a labor-intensive and highly-cost task. In this paper, we implement Word2Vec as an alternative solution of hand-crafted features for sentiment analysis of hotel reviews in the Indonesian language. To obtain the highest performance of sentiment analysis, we evaluate three parameters of Word2Vec include Word2Vec model architecture, evaluation method, and vector dimension. This evaluation process was implemented towards our proposed corpus for a specific domain, i.e. hotel reviews, consists of 2500 hotel reviews in the Indonesian language (1250 positive reviews and 1250 negative reviews). The result shows that the highest accuracy values are obtained under the combination of the following parameters, namely the architecture of Word2Vec Model is Skip-gram model, the evaluation method is Hierarchical Softmax, as well as the vector dimension is 100. The Skip-gram model results highest accuracy for words that rarely appear, such as in sentiment analysis task, whereas the Hierarchical Softmax provides better results since during the training process using a binary tree model to represent all of the words in the vocabulary and leaf nodes representing rare words so that rarely appearing words will inherit vector representations in it. Furthermore, to obtain the optimal value of accuracy, then we should increase the vector dimensions and amount of data simultaneously.

© 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the 4th International Conference on Computer Science and Computational Intelligence 2019.

Keywords: Word2Vec; Sentiment Analysis; Hotel Reviews

* Corresponding author. Tel.:+62-24-7474754; fax: +0-000-000-0000 .
E-mail address: retno@live.undip.ac.id

1877-0509 © 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the 4th International Conference on Computer Science and Computational Intelligence 2019.
10.1016/j.procs.2019.08.178

1. Introduction

The rapid growth of internet technology increases the internet users and the high duration per day. This behavior is a big opportunity for the e-commerce industry, especially the Online Travel Agent (OTA). One of the services of OTA is an online hotel reservation where the service is also equipped with facilities to review the hotels that have been booked by its users. The reviews commonly used as a benchmark of customers satisfaction and information source of features and services that need to be improved for a service provider. A large number of reviews causing difficulty in manually analyzing and concluding the entire review. Therefore, there is a need to analyze the existing reviews automatically, or it is commonly called a sentiment analysis task.

Sentiment analysis is a task to collect and analyze opinions, sentiments, attitudes, emotions, evaluations, and appraisals about products, services, organizations, individuals, events, issues, topics, and their attributes made which are commonly present in blog posts, comments, reviews, or social media [1]. English has been the target language in most sentiment analysis research. However, Indonesian and English have different writing rules and syntax, so it is necessary to know the implementation of sentiment analysis towards hotel reviews in the specific domain language, i.e. Indonesian Language.

There are several conducted studies of sentiment analysis tasks for the Indonesian language, such as sentiment analysis towards movie reviews, tweets of social media twitter, sales reviews of the marketplace, hotel reviews, etc. Most of those studies implement the baseline shallow learning methods such as MaxEnt (Maximum Entropy) [2,3], Naive Bayes [2,3,4,5], and Support Vector Machine (SVM) [2,6]. Furthermore, those studies generally employ hand-crafted features such as word occurrence, word presence, TF-IDF, sentiment lexicon, etc. The performance of many studies which implement combination between shallow learning methods and hand-crafted features commonly depends on the selected of data representation or features [7]. In other words, the performance of shallow learning methods depends on the success of hand-crafted features. It is caused by an inability of shallow learning methods to extract and organize discriminative information from data [7]. Therefore, the implementation of shallow learning methods always requires feature engineering process. Unfortunately, feature engineering process has several drawbacks, i.e. (i) It is labor-intensive [7,8], (ii) It highly-cost to obtain the preferred accuracy since it needs manual pre-processing, e.g. POS tagging and stemming, that is time-consuming and challenging task [9].

Therefore, this study implements Word2Vec as an alternative way of hand-crafted features for this study, i.e. sentiment analysis towards hotel reviews in the Indonesian Language. The Word2Vec model can process unstructured text data by taking word corpus as input and producing word vectors as output. One of the main advantages the Word2Vec model is this model represents features as dense vectors instead of the conventional sparse representations in which it is generally able to overcome the problems of synonyms and homonyms that are often found in the NLP task.

The main contributions of this study are as follows:

- A labelled corpus (positive or negative) of hotel reviews in the Indonesian language.
- A pre-trained Indonesian Word2Vec for domain specific, i.e. sentiment analysis of hotel reviews.
- The best Word2Vec parameters combination, including Word2Vec model architecture, evaluation method, and vector dimension.

The remainder of this paper is organized as follows. Section 2 presents Word2Vec model in detail, Section 3 outlines the experiment results and its analysis, and, finally Section 4 concludes the paper.

2. Word2Vec Model in Detail

The Word2Vec model was proposed by Mikolov with the advantage of this vector representation is able to capture the syntax and semantic meaning of natural language [10]. In addition, the Word2Vec model is a word vector representation algorithm that is able to achieve the best performance in NLP by similar grouping words, i.e. similar words have the same vector.

The Word2Vec model architecture is calculated using a Neural Network with the text body as its input and the vector space as its output. The resulting word vector is a low dimensional space vector that captures the semantic meaning of the word. There are two types of Word2Vec architectural models, namely Skip-Gram models and

Continuous Bag of Words (CBOW) models. The Skip-gram model was introduced as an efficient method for studying large numbers of word vector representations in unstructured text. The Skip-gram model architecture tries to make predictions in the range before or after the current word whose input comes from the current word, whereas CBOW models predict current words based on context words. Word2Vec model architecture, CBOW model and Skip-Gram model can be seen in Fig. 1. and Fig. 2, respectively.

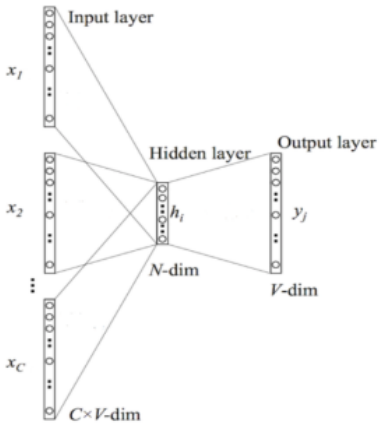

Fig. 1. CBOW Model Architecture

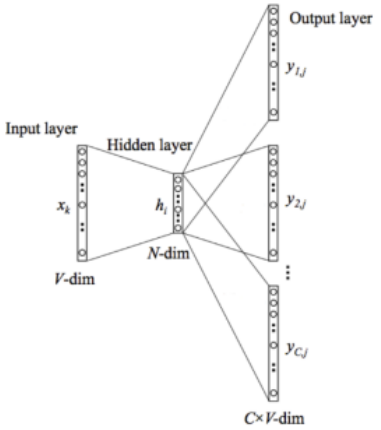

Fig. 2. Skip-Gram Model Architecture

There are two evaluation methods in the Word2Vec model, namely Hierarchical Softmax and Negative Sampling. Hierarchical Softmax was first introduced Morin and Bengio [11]. Hierarchical Softmax uses binary tree representations of the output layer with words as leaves and each node, explicitly representing the relative probabilities of their child nodes while negative sampling is simpler than Hierarchical Softmax because it only updates samples of several output words as negative samples [12]. The following equations show the formula to compute probability of word output O given word input I for Hierarchical Softmax [13].

$$p(w|w_I) = \prod_{j=1}^{L(w)-1} \sigma\left([\![n(w,j+1) = ch(n(w,j))]\!] . v'^{\top}_{n(w,j)} v_{w_I}\right)$$

where $\sigma(x) = 1/(1 + \exp(-x))$. While the objective of Negative Sampling is as follow [13].

$$\log(v'_{wO} \top v_{wI}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)}[\log \sigma(-v'_{w_I} \top v_{w_I})]$$

3. Results and Analysis

As mentioned before, this study aims to evaluate the combination among word2vec model architecture, evaluation method, and the vector dimension. Word2vec model architecture consists of two types, which are CBOW model and Skip-gram model, while the evaluation methods are hierarchical softmax and negative sampling. Three sizes of vector dimension are 100, 200, and 300. Therefore, this study employs 12 parameter combinations of word2vec model as depicted in Table 1.

Table 1. Parameter combinations of word2vec model

ID combination	Architecture	Evaluation method	Vector dimension
W2V-01	CBOW Model	Hierarchical Softmax	100
W2V-02	CBOW Model	Hierarchical Softmax	200
W2V-03	CBOW Model	Hierarchical Softmax	300
W2V-04	CBOW Model	Negative Sampling	100
W2V-05	CBOW Model	Negative Sampling	200
W2V-06	CBOW Model	Negative Sampling	300
W2V-07	Skip-Gram Model	Hierarchical Softmax	100
W2V-08	Skip-Gram Model	Hierarchical Softmax	200
W2V-09	Skip-Gram Model	Hierarchical Softmax	300
W2V-10	Skip-Gram Model	Negative Sampling	100
W2V-11	Skip-Gram Model	Negative Sampling	200
W2V-12	Skip-Gram Model	Negative Sampling	300

Subsequently, all of the word2vec generated from those combinations used as input for CNN-based classification model with the following parameters, i.e. (i) Activation of convolution (ReLU and TanH), (ii). Dropout values (0.2, 0.5, and 0.7), (iii). Activation of output is sigmoid, (iv) Optimizer is SGD, and (v). One output node. Therefore, both of word2vec and CNN parameters result in 72 processes. In addition, the employed validation process in this study is 10-folds cross validation. As a result of this study, we compute the average values of accuracy for each parameter combination, as mentioned in Table 1.

This experimental study employs hotel review in the Indonesian language as data in the amount of 2500 reviews. The details of the data distribution were 1250 positive labelled review data and 1250 negative labelled review data. This amount is felt to be sufficient to perform word2vec training in sentiment analysis for hotel reviews because it has covered various aspects that are usually assessed by hotel guests, such as service, location, food, room, cleanliness, bathroom, neatness, and comfort.

We perform the crawling process from the Traveloka website, i.e. one of unicorn startups in Indonesia with the leading business in Online Travel Agent, to obtain the dataset. The collected reviews consist of several reviews of several hotels in several cities/regions. Those data was crawled by using two libraries, that are, Scrapy (https://scrapy.org) and Selenium (https://selenium-python.readthedocs.io). The crawling process starts from the link to find hotel location or city that are available in Traveloka. Subsequently, the reviews are crawled for each hotel that are appeared in the first result page. All the extracted data are manually labelled by two labels as mentioned

before (positive and negative). The following details explain the effect of changing those word2vec parameters against the accuracy values.

3.1. Effect of Word2Vec Model Architecture on Accuracy Values

Fig. 3 shows the Skip-gram model produces an average value of accuracy better than the CBOW model, that is, the average accuracy value is 92,377% while the CBOW model obtains an average accuracy value of 85,222%. It is because the Skip-gram model works by predicting the given context from one word, while the CBOW model is given the context word to predict the target word. In general, sentiment analysis, many unique words appear so that more rare words appear, and it can be concluded that the Skip-gram model is better at predicting words that rarely appear than CBOW models.

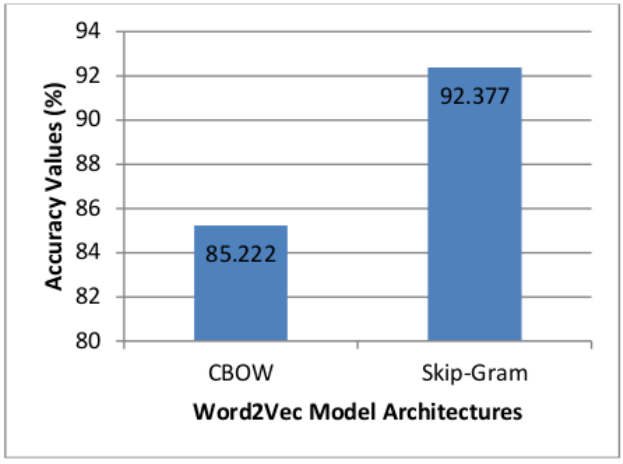

Fig. 3. The Effects of Word2Vec Model on Accuracy Values

3.2. Effect of Evaluation Method on Accuracy Values

The following figure presents the effect of evaluation methods, which are Hierarchical Softmax or Negative Sampling, on accuracy values.

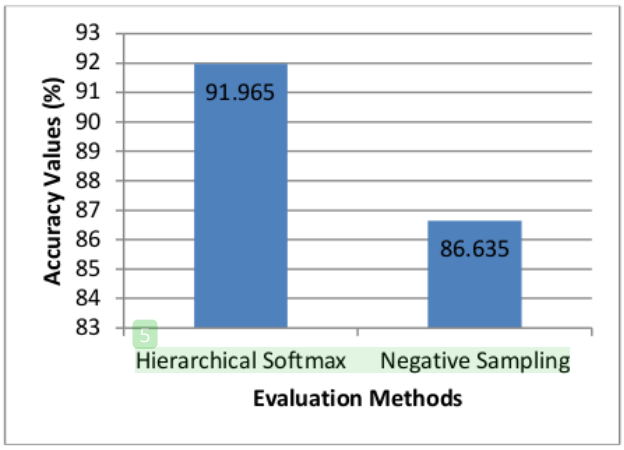

Fig. 4. Effect of Evaluation Method on Accuracy Values

According to Fig. 4, it presents the highest value of accuracy is obtained when the employed evaluation method is Hierarchical Softmax. The accuracy value of Hierarchical Softmax is 91.965%, whereas Negative Sampling is 86.635%. It is because the Hierarchical Softmax evaluation method during the training process uses a binary tree model to represent all the words in the vocabulary and leaf nodes representing rare words so that words that rarely appear will inherit vector representations above [10]. On the contrary, Negative Sampling is simpler than Hierarchical Softmax because it only updates samples of several output words as negative samples.

3.3. Effect of Evaluation Method on Accuracy Values

Fig. 5 depicts that the vector dimensions are directly proportional to the average value of accuracy, which means that the higher vector dimensions give a smaller amount of the average accuracy. The highest value of accuracy is 93.939% for the vector dimension is 100. The accuracy value decreases for vector dimensions 200 and 300, i.e. 93.852 and 93.78, respectively. Increasing the dimensions and amount of data can reduce accuracy, so it is necessary to increase the vector dimensions and amount of data simultaneously [14]. Therefore, the exact vector dimensions for this study are 100 dimensions in harmony with the number of datasets, which only 2500 reviews, hence with a higher size dimension will decrease the accuracy value.

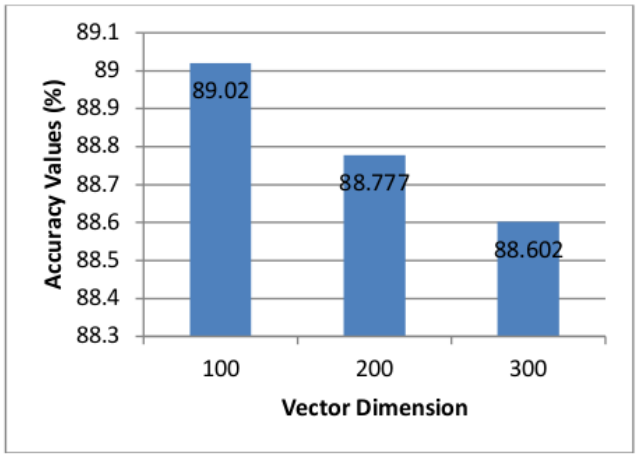

Fig. 5. Effect of Vector Dimension on Accuracy Values

4. Conclusion

The conclusion of this study is parameters of Word2Vec model that give the highest accuracy values are the architecture of Word2Vec Model is Skip-gram model, the evaluation method is Hierarchical Softmax, as well as the vector dimension is 100. That is because the Skip-gram model provides excellent accuracy for words that rarely appear, such as sentiment analysis that the represent words are generally unique. As a result, the Hierarchical Softmax for the evaluation method provides better results because during the training process using a binary tree model to represent all of the words in the vocabulary and leaf nodes representing rare words so that rarely appearing words will inherit vector representations in it. Furthermore, the vector dimensions also affect the average value of accuracy. To obtain the optimal value of accuracy, then we should increase the vector dimensions and amount of data simultaneously.

Acknowledgements

This research was supported by Directorat Research and Development, Ministry of Research, Technology and Higher Education, Indonesia under the grant of Basic Research, fiscal year 2019 Number 257-20/UN7.P4.3/PP/2019.

References

1. Liu B. Sentiment Analysis and Opinion Mining: Morgan & Claypool Publishers; 2012.
2. Satriaji W, Kusumaningrum R. Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis. In Proceedings of the 2nd International Conference on Informatics and Computational Sciences (ICICoS); 2018; Semarang, Indonesia. p. 99-103.
3. Wicaksono AF, Vania C, Distiawan BT, Adriani M. Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation; 2014; Phuket, Thailand. p. 185-194.
4. Nurdiansyah Y, Bukhori S, Hidayat R. Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method. IOP Journal of Physics: Conference Series. 2018; 1008(1).
5. Kurniawan S, Kusumaningrum R, Timu ME. Hierarchical Sentence Sentiment Analysis Of Hotel Reviews Using The Naïve Bayes Classifier. In Proceeding of the 2nd International Conference on Informatics and Computational Sciences (ICICoS); 2018; Semarang, Indonesia. p. 104-108.
6. Lutfi AA, Permanasari AE, Fauziati S. Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine. Journal of Information Systems Engineering and Business Intelligence. 2018 April; 4(1).
7. Bengio Y. Deep Learning of Representations: Looking Forward. In Proceedings of International Conference on Statistical Language and Speech Processing; 2013; Tarragona, Spain. p. 1-37.
8. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014; Maryland, USA. p. 1555–1565.
9. Altowayan AA, Tao L. Word Embedding for Arabic Sentiment Analysis. In Proceedings of IEEE International Conference on Big Data (Big Data); 2016; Washington DC, USA. p. 3820 - 3825.
10. Rong X. word2vec Parameter Learning Explained. CoRR. 2014.
11. Morin F, Bengio Y. Hierarchical Probabilistic Neural Network Language Model. In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics; 2005. p. 246-252.
12. Mikolov T, Yih Wt, Zweig G. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2013; Georgia, USA. p. 746–751.
13. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. In NIPS'13; 2013; Nevada. p. 3111-3119.
14. Mikolov T, Corrado GS, Chen K, Dean J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the International Conference on Learning Representations; 2013; Arizona, USA. p. 1301–3781.

Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study

ORIGINALITY REPORT

17%	10%	13%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to West Coast University Student Paper	2%
2	Submitted to University of Florida Student Paper	1%
3	toc.proceedings.com Internet Source	1%
4	Hankz Hankui Zhuo, Yantian Zha, Subbarao Kambhampati, Xin Tian. "Discovering Underlying Plans Based on Shallow Models", ACM Transactions on Intelligent Systems and Technology, 2020 Publication	1%
5	Dwi Intan Af'idah, Retno Kusumaningrum, Bayu Surarso. "Long Short Term Memory Convolutional Neural Network for Indonesian Sentiment Analysis towards Touristic Destination Reviews", 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), 2020 Publication	1%

6	Serdar Koçak, Yusuf Tansel Ic, Mustafa Sert, Kumru Didem Atalay, Berna Dengiz. "Development of a Decision Support System for Selection of Reviewers to Evaluate Research and Development Projects", International Journal of Information Technology & Decision Making, 2022 Publication	1 %
7	Sicong Kuang, Brian D. Davison. "Class-Specific Word Embedding through Linear Compositionality", 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), 2018 Publication	1 %
8	Rizka Vio Octriany Inggit Sudiro, Sri Suryani Prasetiyowati, Yuliant Sibaroni. "Aspect Based Sentiment Analysis With Combination Feature Extraction LDA and Word2vec", 2021 9th International Conference on Information and Communication Technology (ICoICT), 2021 Publication	1 %
9	Qian Wang, Minxin Du, Xiuying Chen, Yanjiao Chen, Pan Zhou, Xiaofeng Chen, Xinyi Huang. "Privacy-Preserving Collaborative Model Learning: The Case of Word Vector Training", IEEE Transactions on Knowledge and Data Engineering, 2018 Publication	1 %

10	Internet Source	1 %
11	Submitted to School of Business and Management ITB Student Paper	1 %
12	link.springer.com Internet Source	1 %
13	Rahmat Jayanto, Retno Kusumaningrum, Adi Wibowo. "Aspect-based sentiment analysis for hotel reviews using an improved model of long short-term memory", International Journal of Advances in Intelligent Informatics, 2022 Publication	<1 %
14	Submitted to University of Birmingham Student Paper	<1 %
15	courses.ece.ubc.ca Internet Source	<1 %
16	www.ncbi.nlm.nih.gov Internet Source	<1 %
17	Chao Zhang, Hui Song, Zhenyu Liu. "MiSAS", Proceedings of the 3rd International Conference on Communication and Information Processing, 2017 Publication	<1 %
18	J. Shobana, M. Murali. "Improving feature engineering by fine tuning the parameters	<1 %

	of Skip gram model", Materials Today: Proceedings, 2021 Publication	
19	dokumen.pub Internet Source	<1 %
20	downloads.hindawi.com Internet Source	<1 %
21	eprints.kfupm.edu.sa Internet Source	<1 %
22	"Digital Libraries: Data, Information, and Knowledge for Digital Lives", Springer Science and Business Media LLC, 2017 Publication	<1 %
23	Borko Furht, Flavio Villanustre. "Big Data Technologies and Applications", Springer Science and Business Media LLC, 2016 Publication	<1 %
24	Lecture Notes in Computer Science, 2013. Publication	<1 %
25	Shulin Niu. "Emotion research on education public opinion based on text analysis and deep learning", Frontiers in Psychology, 2022 Publication	<1 %
26	iopscience.iop.org Internet Source	<1 %
27	www.science.gov Internet Source	<1 %

28	Yiran Gu, Jiajia Shen. "Short Text Classification Based on Keywords Extension", 2019 Chinese Automation Congress (CAC), 2019 Publication	<1 %
29	"Recent Advances in Information and Communication Technology 2017", Springer Science and Business Media LLC, 2018 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

Word2Vec for Indonesian Sentiment Analysis towards Hotel Reviews: An Evaluation Study

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7