ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter1218102020.pdf
https://doi.org/10.30534/ijeter/2020/1218102020

Word2vec Feature Extraction in Traveler Comments Using Machine Learningin Imbalance Data

[1]Abba Suganda Girsang, [2]Isra Nurul Habibi
[1]Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480,Email: agirsang@binus.edu
[2]Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, Email: isra.habibi@binus.ac.id

ABSTRACT
The tourism industry is one industry that utilizes promotion of product and services through Internet and web technology. One of the uses is on the tripadvisor website which provides a place for users to give their opinions on attractions, accommodations and hospitality. The opinion given is in the form of comments and ratings on a topic. This research was conducted to classify user comments on tourist objects to be in the form of rating on a scale of 1 to 5. The dataset used is user opinion data on the tripadvisor application with a total data of 17675. Word2vec is used to extract semantic features from words from the data they have. The data classification algorithms used in this study are random forest and SVM, and use the SMOTE and NearMiss handling imbalance techniques. The results showed that the utilization of random forest algorithms and SMOTE provided the best accuracy results with an accuracy of 82.11%, an average precision of 81.2% and an average F Score of 81.58%.

Key words :imbalance, machine learning, tourism, word2vec

1. INTRODUCTION

Social media has been used by all ages, be it children, eenagers to adults. Various types of social media that are popularly used include Facebook, Twitter and Instagram. According to data from the Smartinstigth site[1], the number of social media usage is Facebook with 2,167 million users, Instagram with 800 million users and Twitter with 330 million users. Social media is used on a daily basis, whether it's just to spend time or to get the desired information.

The popularity of social media with the large number of users who use it regularly and the large amount of content provided by social media applications has made the use of social media widespread. Not only used to share content and as a place to communicate, social media is also used as a place to discuss and exchange ideas among users without any spatial boundaries[2]. Users from various places and at any time can communicate with each other and exchange ideas.

The use of social media as a place to promote a product creates a term social commerce. Social commerce combines Web 2.0 technology, e-business and community - based online communities. Social Commerce makes use of promotions through comments and feedback from a product. A product or service can be promoted through an e-commerce site Comments - comments in the form of testimonials that are given and go viral are also called Viral Marketing, meaning that promotions are carried out by mouth - by mouth and the ones doing the promotion are the customers themselves.

The tourism industry is an industry that uses social media and the internet to promote its products, both through e-commerce and social media. User Generated Content type social media applications, such as tripadvisor, make it easy for users to find information about tourist attractions to go to[3]. In addition to users, the UGC application also has a very significant influence in promoting a product, both goods and services[3]. This application also makes it easier for users to comment on values in the form of numbers. The value given can be in the form of points, stars, or other methods that can be easily aggregated.

One of application tourism industry is tripadvisor. Many researchers take attention to do the research in this system[4][5][6]. Tripadvisor application provides a scoring in the form of numbers 1 to 5. A score of 1 for bad opinion and 5 for very good opinion. Other applications such as YouTube provide an assessment of a video in the form of likes or dislikes. The process of ranking by users is done by providing comments and followed by giving ratings of the comments given. This process can be done by the user after the user registers on the site. Registration is used so that data provided by users can be accounted for.

In addition to using a User Generated Content-based application with ease in providing comments and providing ratings, users also frequently comment through microblog social media applications such as Twitter and the Facebook application in providing comments. If on the tripadvisor application the user provides a review on the provided place,

on social media, Facebook or Twitter, there is no special place to provide a review[7]. Users only provide comments and there are no facilities that provide value and points for comments made.

Text classification can be used to score a review. On several sites that provide review of tourist attractions such as tripadvisor.com, users are required to provide an assessment point in the form of a score on the given review. Text classification using word2vec can be used to classify a text review and categorize the score of a text review[8][9]The use of text classification to determine the sentiment of a sentence has been carried out by previous researchers[10] using word2vec and SVM to calculate sentiment from user comment data on clothing products in Chinese. The research was conducted to take advantage of the semantic relationship between words in a sentence. The results of the study prove that the word2vec and SVM methods provide satisfactory results for determining the classification of sentiments of a text.

Based on observations of the distribution of data to be collected, there is a considerable discrepancy in the amount of data with different ratings. For example, the rating data on the Mount Batur tourist attraction with a total of 874, has an unbalanced rating ratio between the 5 types of ratings provided, ranging from "Very Bad" to "Extraordinary". This imbalance can cause problems in the classification process to be carried out, where the prediction results obtained will be more likely to the data that has the largest number of classes.

There are several methods that can be used to overcome this, namely undersampling and oversampling. Oversampling is used to balance the amount of data distribution by increasing the amount of data with a few classes[11], and undesampling is used to balance data by decreasing the amount of data with a large number of classes[12]. In this research, the test data will be applied undersampling and oversampling methods and it will be seen which method results in more data accuracy.

Reviews provided by travelers are not limited to the tripadvisor application which provides a rating for each comment, but also from other social media or other platforms that can provide comments but do not provide a rating for the comment.

The expected result of this research is a model. The resulting model is expected to be used to classify tourist comments related to a tourist attraction into 5 classes based on the existing rating in tripadvisor data. So that other tourists who want to collect information related to a tourist attraction can easily draw conclusions about a tourist attraction without having to read every review sentence on the attraction.

This studywill discuss the use of using word2vec to represent words into vectors and classification using the random forest algorithm and SVM to classify points on a user's opinion on the tripadvisor application.This study also compares the result of that method using balance and imbalance data.

2. RELATED STUDY

Previous research[13]conducted research in the field of text classification on a dataset of 18,000 news items with 20 different types of topics. The research was conducted to test the use of word2vec, tf idf with the help of the SVM algorithm to categorize the types of news from the dataset. The research was conducted with the vector size parameter in word2vec with a value of 100 and the Linear SVM algorithm. The research step was carried out by preprocessing the data, then converting the data into a vector representation using word2vec and predicting news topics on the data using the Linear SVM algorithm. The research results show that the use of word2vec with the addition of tf-idf weighting and the Linear SVM algorithm produces the highest classification accuracy with a value of 0.89.

Research using word2vec and the SVM machine learning algorithm was also carried out by other researchers[14]. This study uses data sources from Twitter with the topics of sports, religion, culture, food and clothing. The study was conducted using word2vec to represent words from twitter into vectors and feature extraction was carried out using CNN. Then the sentiment classification is carried out using the Linear SVM algorithm. The results of the study resulted in the highest level of data accuracy obtained using the above methods and algorithms, which was 97.3%.

Unlike the previous researchers, this study[15]uses the Random Forest algorithm in its implementation. The study was conducted to classify twitter based on emotions in data using word embedding and random forest. The types of emotions used in this study were happy, sad, angry and surprised emotions. The data obtained from Twitter amounted to more than 100,000 datasets with data that had certain hashtags according to the category of each emotion to facilitate classification. The results of this study are the use of the Random Forest algorithm with the parameter number of estimators 200 and word embeding fast text with 300 vector dimensions which have the highest accuracy level of 0.91.

The use of the Random Forest algorithm is also carried out by Yang et al [16] on EMR data. EMR (Electronic Medical Record) is a medical record that contains a lot of information about patients. This data is very useful for analysis and can be used as material for decision making and further action by the hospital. The obstacle that this study found was the amount of missing information and data, such as the absence of labels on some data. The study was conducted to determine the label of the missing EMR data by making use of the EMR records. The study was conducted by utilizing a total of 282740 medical records from Shenzen hospitals. However, from these data, only 15466 total data have complete data labels. This means that more than 90% of the data is unlabeled. The research was conducted using word2vec and the Random Forest algorithm. The results of the study provide an average accuracy rate of 91%.

3. PROPOSED METHOD

The research was conducted by utilizing data available on the tripadvisor.co.id website. This website does not have a public API with which to interact with raw data. However, every web application that is opened through a browser has a data source in html format. This research will use a web scrapper to get data from tripadvisor and the data will be stored in an SQL database for further processing.This research focuses on building a model using word2vec to represent words into vectors and applying machine learning algorithms for the comment classification process according to the data labels. Research on the classification of a text has been carried out by previous research and is discussed in Section 2 with the topic of data and the use of different algorithms. This study uses data with Indonesian text obtained from user opinions on the tripadvisor site.Figure 1 describes the research framework to be carried out. The research that will be carried out starts from collecting data from the tripadvisor.co.id site.

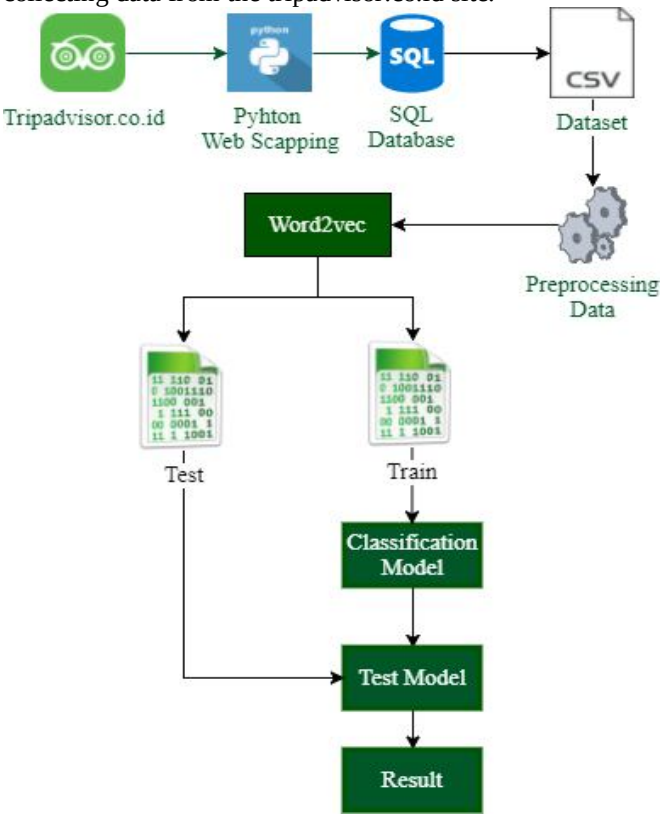

Figure 1 Methodology Framework

The data was collected by creating a scrapping web application based on the python programming language, the data collected was data from comments and opinions of application users that had weight on their comments. The data obtained is first stored in the SQL database, the goal is to create a data storage place and facilitate the integration and addition of data and the descriptive data analysis process.

3.1 Gathering Data

The data used in this research is data taken from the tripadvisor.co.id site. On the application page, information will be found about a tourist spot according to the entered keywords. Figure 2 is an example of a review page display with Borobudur Temple attractions.The data that appears on the web page can be retrieved using a web scrapping application made in python language.On the tripadvisor application web page, there is some information that can be used for analysis, as for the required data, among others

Title: Is the name of a tourist spot,
Review Id: The id for each existing review,
Rating Date: The date the rating was given
Comment Title: The title of the rating given
Rating: The rating provided by the user with the number 10, 20, 30, 40 or 50
Comment: Comments from the rating given
Insert Date: This is the time when the web scrapping process is done

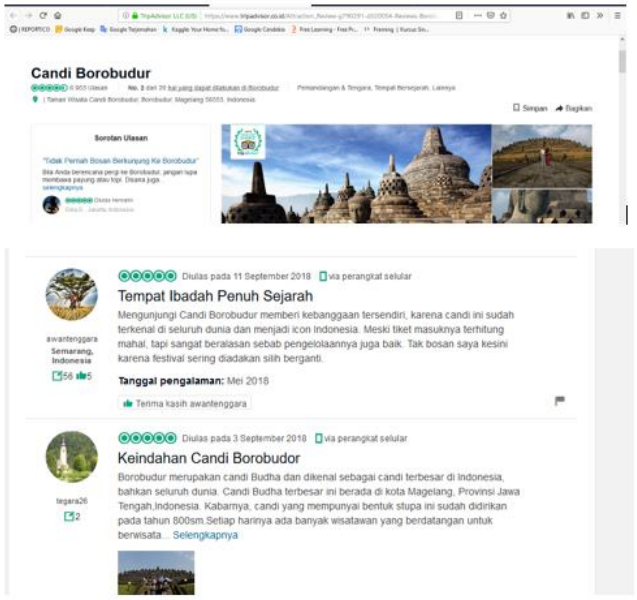

Figure 2 Tripadvisor Web Page Views

The data collection process was carried out on February 21, 2019. The data obtained in the data collection process were 17675 data. The data selection is done by selecting the types of tourist objects of the temples, lakes, mountains and beaches. Determination of data collection is done by looking at popular tourist objects on the islands of Java, Bali and Lombok. The data collected does not only take tourist attractions randomly, but also takes into account the types of tourist attractions. Of the 20 types of tourist attractions, they can be grouped into 3 types, namely mountain tours, beach tours and temple tours. Table 1 shows the data for these objects.

Table1 Number Data based on Tourist Type

No	Tourist Type	Number Data
1	Temple	5874
2	Mountain	5587
3	Beach	6214

Table 2 Percentage Data

Rating	Number Data	Percentage
10	246	1.39%
20	326	1.84%
30	1783	10.09%
40	6351	35.93%
50	8969	50.74%

The distribution of rating data on the obtained dataset is distributed as shown in Table 2. The rating data is obtained in raw data with a value of 10 to 50. With the lowest rating of 10 and the highest rating of 50. The distribution of the data is not balanced. The most data is data with a rating of 50 is 50.74%. And the least data is data with a rating of 10 with a percentage of 1.39%.Class normalization is done by changing the name of the existing class. In the data collection process, the column rating is 10, 20, 30, 40 and 50. The normalization process is carried out by making the data in the rating column become 1, 2, 3, 4 and 5.

3.2 Vectorization Data

The vector representation of a sentence is obtained after the prepocessing process is complete. The next process is combining all sentences into 1 large part. The vector formation of each word in a sentence is carried out using the word2vec gensim library. The vector that is formed is a representation of one word in a sentence in the form of a matrix.

The process of vector formation training is carried out for all existing words. So every word in each sentence is combined into 1 container variable of type array to be processed and used as a vector. The data training process is carried out using the generator library. The value of the vector dimension used in this study is 300.This means that every word contained in the data will be represented as a vector with dimensions of 300.

Further data testing is done by making a model using the Random Forest algorithm and SVM. The parameter used in the implementation of the Random Forest algorithm is the number of estimators 100. The type of SVM algorithm used is Linear SVM.

3.3 Balancing Data

To solve the problem of data imbalance, 2 approaches are used, namely oversampling and undersampling. Oversampling uses the SMOTE algorithm and undersampling uses the Near Miss library.The use of SMOTE as an oversampling technique on the dataset increases the amount of data. SMOTE creates new synthetic data in classes that have a smaller amount of data, so that the amount of data in each class is the same following the largest number of data, where in this experiment the number of data becomes 8969. The implementation of undersampling using Near Miss makes the data reduced to as much data with the least class, in this experiment the data amounted to 246. Table 3 shows the all data with oversampling (SMOTE) and undersampling (Near Miss)

Table 3 Data with Balancing

Rating		1	2	3	4	5	Total
ALL DATA	Training	197	261	1426	5081	7175	14140
	Test	49	65	357	1270	1794	3535
	Total	246	326	1783	6351	8969	17675
SMOTE	Training	7175	7175	7175	7175	7175	35875
	Test	1794	1794	1794	1794	1794	8970
	Total	8969	8969	8969	8969	8969	8969
Near Miss	Training	197	197	197	197	197	985
	Test	49	49	49	49	49	245
	Total	246	246	246	246	246	1230

4. ANALYSIS RESULTS

4.1 Performance Analysis

This study was carried out with variations in the use of the SVM and Random Forest algorithms, as well as the use of algorithms to solve data imbalance problems with SMOTE and Near Miss. Table 4 and Table 5 showthe result in confusing matrix using random forest, and SVM, respectively. Table 6 and Table 7 show the result in confusing matrix using oversampling (SMOTE) with random forest, and SVM, respectively. Table 8 and Table 9 show the result in confusing matrix using undersampling (Near Mess with random forest, and SVM, respectively.

Table 4 Confusion Matrix Using Random Forest

Rating		Prediction				
		1	2	3	4	5
Target	1	2	0	7	16	24
	2	2	0	6	32	25
	3	0	0	11	156	186
	4	0	0	12	471	787
	5	0	0	3	406	1385
Accuracy		4%	0%	3%	37%	77%

Table 5 Confusion Matrix Using SVM

Rating		Prediction				
		1	2	3	4	5
Target	1	0	0	0	12	37
	2	0	0	0	23	42
	3	0	0	0	154	203
	4	0	0	1	389	880
	5	0	0	0	257	1537
Accuracy		0 %	0 %	0%	31%	86%

Table 6 Confusion Matrix Using Random Forest –SMOTE

Rating		Prediction				
		1	**2**	**3**	**4**	**5**
Target	**1**	1789	0	0	4	1
	2	0	1784	1	2	7
	3	12	12	1669	30	70
	4	7	15	133	1138	501
	5	14	25	122	535	1098
Accuracy		99%	99%	93%	63%	61%

Table 7 Confusion Matrix Using SVM -SMOTE

Rating		Prediction				
		1	**2**	**3**	**4**	**5**
Target	**1**	1368	220	95	23	88
	2	790	593	175	63	173
	3	343	307	465	310	368
	4	196	142	275	518	663
	5	194	101	131	328	1040
Accuracy		76%	33%	26%	29%	58%

Table 8 Confusion Matrix Using Random Forest - Near Miss

Rating		Prediction				
		1	**2**	**3**	**4**	**5**
Target	**1**	21	9	10	6	3
	2	20	12	7	8	2
	3	9	7	12	7	14
	4	3	3	8	15	21
	5	2	2	12	9	24
Accuracy		43%	24%	24%	30%	49%

Table 9 Confusion Matrix Using SVM -NearMiss

Rating		Prediction				
		1	**2**	**3**	**4**	**5**
Target	**1**	23	8	5	6	7
	2	19	13	4	5	8
	3	11	9	7	5	17
	4	4	4	3	12	27
	5	6	5	2	7	29
Accuracy		47%	27%	14%	24%	59%

The result of accuracy, precision and FScore can be seen in Figure 3.

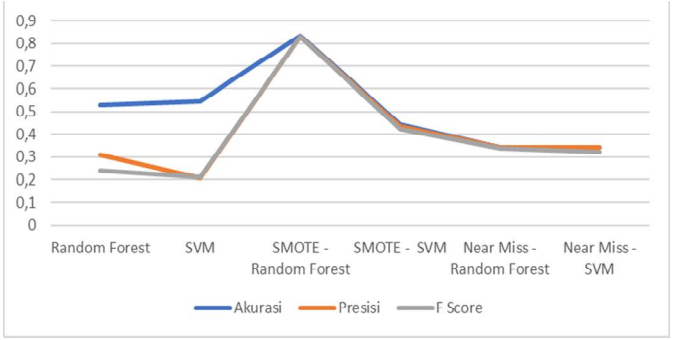

Figure 3. Evaluation of Each Algorithm

To evaluate detail, the same experiments are done by same methods, but separating of each type data. The previous scenario, with all data, is called the first scenario. The second scenario is only data with the type of mountain tourism with the number of data is 5587, the third scenario is only data with the type of beach tourism with the number of data 6214, and the fourth scenario is only data with the type of temple tourism with the number of data is 5874.

To get the algorithm and experiment that has the highest accuracy, the value of accuracy, precision and F Score of the four experimental results is averaged, the average results of the experiment can be seen in Figure 4.

It can be seen that the experiment using the Random Forest and SMOTE algorithms has the highest accuracy, precision and F score, namely with an average accuracy of 82.11%, an average precision of 81.2% and an average F Score of 81.58%. . The results of using the Random Forest algorithm with SMOTE oversampling are striking with the use of other algorithms.

The experimental scenario and data variation that have the highest accuracy value is the scenario with the Random Forest algorithm with SMOTE oversampling, from the 4 scenarios carried out the accuracy, precision and F score for each experiment are shown:

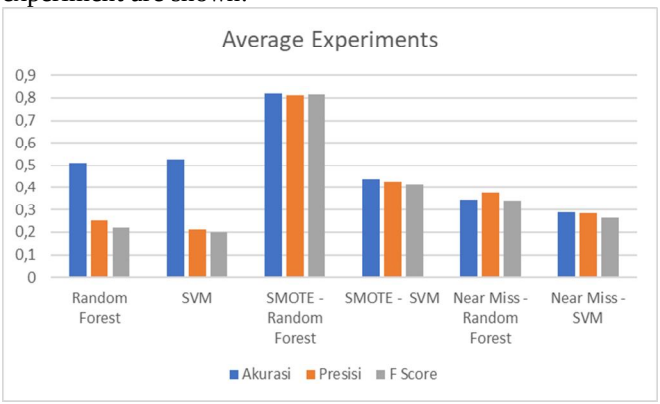

Figure 4 Average of Four Scenarios

We also found that the results of the experiment with the Random Forest Algorithm and SMOTE of 4 times the variation of the data. Accuracy, precision and F Score in the all four scenarios have almost the same value with first scenarios .This means that the scenario by separating the types of tourist objects at the time the data training process is carried out does not really affect the results of accuracy and precision. The smallest accuracy value is in experiment 3 with beach tourism object data with an accuracy value of = 76%, precision = 74% and an FScore of 0.75. While the greatest accuracy value is in experiment 4 with the type of temple tourism object with an accuracy value of = 85%, precision = 84% and an F score of 0.84. The difference between the highest and lowest scores is accuracy = 9%, precision = 10% and F score 0.9.

4.2 Analysis Using Balancing Data

The imbalance of the data in this study resulted in an unbalanced accuracy, precision and F score. In the first

scenario, each data variation is data with an imbalance condition. The accuracy and precision values are very different, this is because in the imbalance data, the prediction of the majority only piles up on the majority data. In Table 10, the data with a rating of 1, totaling 49, are predicted to be exactly 1 as 2 pieces of data, and it is predicted rating 5 with 24 data. This happens because the data used as training data in this experiment is in an imbalance condition. Other experiments that apply oversampling to the data produce accuracy, precision and F Score values that are not much different, for example in Table 11, the results of the first experiment with SMOTE to solve data imbalance problems and algorithms for random forest classification. However, if the accuracy value is taken for each class in the experimental data, the average results for each scenarios are obtained as in Figure 5.

Table 10 Distribution prediction in *Imbalance Data*

Rating		Prediction					Total
		1	2	3	4	5	
Target	1	2	0	7	16	24	49
	2	2	0	6	32	25	65
	3	0	0	11	156	186	353
	4	0	0	12	471	787	1270
	5	0	0	3	406	1385	1794
Total		4	0	39	1081	2407	

Table 11 Distribution Prediction in Data *Oversampling*Using SMOTE

Rating		Prediksi					Total
		1	2	3	4	5	
Target	1	1789	0	0	4	1	1794
	2	0	1784	1	2	7	1794
	3	12	12	1669	30	70	1793
	4	7	15	133	1138	501	1794
	5	14	25	122	535	1098	1794
Total		1822	1836	1925	1709	1677	

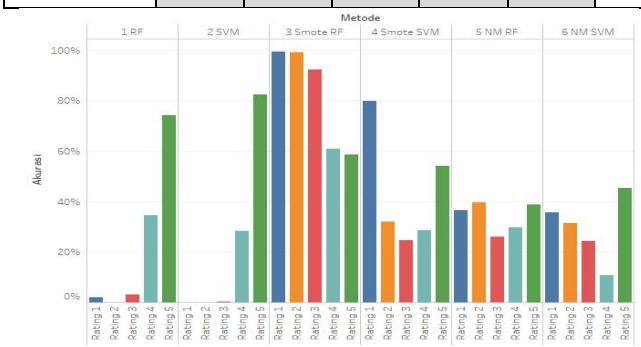

Figure 5 Graph of Average Accuracy of Each Class

Figure 5 shows that the experiment without using imbalance data handling SMOTE and NearMiss, the highest accuracy is in the data with a rating of 5, with an average accuracy of the Random Forest method of 74% and SVM of 82%. But the accuracy value in the other classes is very low. In the experiment using SMOTE and the Random Forest Algorithm, the accuracy value at rating 5 fell to 59%, while other accuracy values became high. This shows that the use of SMOTE with an unbalanced amount of data in each class (imbalance) makes overall accuracy better, but the accuracy in many classes decreases.

5. CONCLUSION

Classification of comment data on the tripadvisor site based on research results can be done automatically by using word2vec to convert words into vectors and random forest machine learning algorithms for data classification.The classification of comment data in the tripadvisor application uses word2vec to get semantic features and the random forest algorithm has a better value than the SVM algorithm. The average accuracy value in the scenario with the random forest algorithm with handling imbalance data using SMOTE increase average accuracy, precision and average F score. Variation of tourism data in this experiment does not have a significant effect on the accuracy value.

On future research, the classification of tourist opinion texts can use more datasets with more diverse types of tourist objects. Also, considering the use of other data preprocessing processes such as making a slank word data dictionary and using stemming

REFERENCES

[1] D. Chaffey, "**Global social media research summary July 2020,**" 2020.

[2] Z. N. Gastelum and K. M. Whattam, "S**tate-of-the-art of social media analytics research**," Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2013.

[3] Q. Ye, R. Law, B. Gu, and W. Chen, "**The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings**," *Comput. Human Behav.*, vol. 27, no. 2, pp. 634–639, 2011.

[4] A. Valdivia, M. V. Luzón, and F. Herrera, "**Sentiment analysis on tripadvisor: Are there inconsistencies in user reviews?,**" in International Conference on Hybrid Artificial Intelligence Systems, 2017, pp. 15–25.

[5] P. Bhardwaj, S. Gautam, and P. Pahwa, "**A novel approach to analyze the sentiments of tweets related to TripAdvisor**," *J. Inf. Optim. Sci.*, vol. 39, no. 2, pp.

591–605, 2018.

[6] Y.-C. Chang, C.-H. Ku, and C.-H. Chen, "**Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor**," *Int. J. Inf. Manage.*, vol. 48, pp. 263–279, 2019.

[7] J. Miguéns, R. Baggio, and C. Costa, "**Social media and tourism destinations: TripAdvisor case study**," *Adv. Tour. Res.*, vol. 26, no. 28, pp. 1–6, 2008.

[8] R. Kurnia, Y. D. Tangkuman, and A. S. Girsang, "**Classification of User Comment Using Word2vec and SVM Classifier**," *Int. J. Adv. Trends Comput. Sci. Eng*, vol. 9, pp. 643–648, 2020.

[9] I. N. Habibi and A. S. Girsang, "**Classification of tourist comment using word2vec and random forest algorithm**," *J. Environ. Manag. Tour.*, vol. 9, no. 8, pp. 1725–1732, 2018.

[10] D. Zhang, H. Xu, Z. Su, and Y. Xu, "**Chinese comments sentiment classification based on word2vec and SVMperf**," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, 2015.

[11] R. Y. Rumagit and A. S. Girsang, "**Predicting personality traits of facebook users using text mining**," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 20, pp. 6877–6888, 2018.

[12] S.-J. Yen and Y.-S. Lee, "**Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset**," in Intelligent Control and Automation, Springer, 2006, pp. 731–740.

[13] J. Lilleberg, Y. Zhu, and Y. Zhang, "**Support vector machines and word2vec for text classification with semantic features**," in 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015, pp. 136–140.

[14] A. H. Ombabi, O. Lazzez, W. Ouarda, and A. M. Alimi, "**Deep learning framework based on Word2Vec and CNNfor users interests classification**," in 2017 Sudan Conference on Computer Science and Information Technology (SCCSIT), 2017, pp. 1–7.

[15] P. Vora, M. Khara, and K. Kelkar, "**Classification of tweets based on emotions using word embedding and random forest classifiers**," *Int. J. Comput. Appl.*, vol. 178, no. 3, pp. 1–7, 2017.

[16] B. Yang et al., "**Automatic text classification for label imputation of medical diagnosis notes based on random forest**," in International Conference on Health Information Science, 2018, pp. 87–97.