Classification of User Comment Using Word2vec and Deep Learning

Rafly Indra Kurnia¹, Abba Suganda Girsang²

^{1,2}Computer Science Department, BINUS Graduate Program-Master of Computer Science,
Bina Nusantara University, Jakarta, Indonesia 11480

Abstract --This study will classify the text based on the rating of the provider application on the Google Play Store. This research is classification of user comments using Word2vec and the deep learning algorithm in this case is Long Short Term Memory (LSTM) based on the rating given with a rating scale of 1-5 with a detailed rating 1 is the lowest and rating 5 is the highest data and a rating scale of 1-3 with a detailed rating, 1 as a negative is a combination of ratings 1 and 2, rating 2 as a neutral is rating 3, and rating 3 as a positive is a combination of ratings 4 and 5 to get sentiment from users using SMOTE oversampling to handle the imbalance data. The data used are 16369 data. The training data and the testing data will be taken from user comments MyTelkomsel's application from the play.google.com site where each comment has a rating in Indonesian Language. This review data will be very useful for companies to make business decisions. This data can be obtained from social media, but social media does not provide a rating feature for every user comment. This research goal is that data from social media such as Twitter or Facebook can also quickly find out the total of the user satisfaction based from the rating from the comment given. The best f1 scores and precisions obtained using 5 classes with LSTM and SMOTE were 0.62 and 0.70 and the best f1 scores and precisions obtained using 3 classes with LSTM and SMOTE were 0.86 and 0.87.

Keyword -- deep learning, LSTM, multiclass, text classification, word2vec

Users can review in the review column to provide input or others with the mobile network application. Users can give a rating of 1-5 to the review, rating 1 is the lowest rating and rating 5 is the highest rating and these reviews have their respective rating scores so that they can be sorted by rating. Users can do this review if they are registered with a Google account so that the review can be accounted for by the user who did the review.

Social commerce is part of e-commerce and uses SNS for social interaction and user contributions to facilitate the buying and selling of products and services online [1]. The last few years have seen rapid growth in social commerce but this growth has involved a number of problems, in particular, consumer trust has become an important factor in the success of companies in the modern era, which requires these companies to put more effort into earning this trust [2]. The benefits of social commerce from media are by utilizing interaction between users and feedback to get input and also promote products and services, so that user contributions through feedback and comments can influence the desire of other users to buy or sell products or services [3]. The increase in the number of internet users in a country will also lead to an increase in the phenomenon of this use through online trading. User satisfaction and loyalty from customers who buy and sell products online can also be used as one of the factors for assessing the success of a business online [4].

Text classification used here is using the word2vec extraction feature which can be used to classify opinion comments and categorize the score of a comment. The results of the text that has been classified based on the rating can be used to predict the level of desire of the user to use a provider application. Text classification using word2vec can be used to classify reviews and categorize. This research was conducted by utilizing the semantic relationship between words in a sentence. The use of text classification has been done before [5] using word2vec and deep learning, word2vec is used to obtain pre-trained data for neural networks in order to get a better initial weight value in the network, in this study the results obtained show good results when use pre-trained data using word2vec than not using word2vec.

I. INTRODUCTION

Mobile networks are now widely used for various activities such as learning, business, and so on. For now, almost all mobile network providers have their own applications in the google play store. This application can be used as a mobile network provider to get feedback directly from its users, feedback can be obtained from the features provided by the Google Play Store, specifically the comments and ratings, which are part of the application reviews. User reviews can be about the UI / UX of the application, the network speed of the mobile network, the signal stability of the mobile network, and other factors in the review column of the Google Play Store feature.

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 11, Issue 05, May 2021)

Based on the literature review, research on text classification using Indonesian language is still minimal compared to English.

In this research will be conducted with 2 datasets using data from comments on the myTelkomsel's application using 5 classes data and 3 classes with details of rating 1 as positive which is a combination of ratings 1 and 2, rating 2 as neutral is rating 3, and rating 3 as positive which is a combination of ratings 4 and 5 to get sentiment from users who do reviews. In the research the test data will be applied using 5 classes and 3 classes. Based on the observations, the distribution of data from the data collected, there is a fairly large imbalance in the amount of data available at each rating, this imbalance can lead to miss classification where the prediction results will be more likely to majority classes. In the training data research, the oversampling method will be applied to obtain a more balanced data.

This research with the title "Classification of User Comment Using Word2vec and Deep Learning" will discuss the use of word2vec to get a vector representation and text classification using deep learning in this case the LSTM (Long-Short Term Memory) method to categorize each comment so that it can be used to know the user's desire to use the provider on social media. The data used in this study are user comments in Indonesian language.

II. RELATED WORKS

This section overviews the existing works related to sentiment analysis, particularly emphasizing on the use of word embedding and deep learning algorithms for text classification.

In [5] the authors using a dataset from Yahoo in Japanese with a total of 218295 data. This study used 6 categories to categorize news with the categories used are entertainment, sports, economy, science, domestic and overseas. in this study, the authors used word2vec to get a representation of words in vector form and for text processing with 200 dimensions. After that, to categorize the news using a comparison using naive bayes and deep learning with the parameters of 2 hidden layers, 50,100,150 and 200 nodes, 0.1 to 0.9 the probability of occurrence of noise, 100 epochs and 100 batch. The best accuracy results with 6 categories using Naive Bayes is 68% and the best accuracy results using LSTM and pre-training is 78%.

In [6] the authors use data from a short text description of railroad accidents in the US. The text descriptions used in this dataset are difficult for non-experts to understand, so the main objective of this study is to find an effective method to fill in the root causes of accidents that will help to give label more accurately.

In this study, the authors used word2vec to get a representation of words in vector form and Glove to classify the cause of the accidents, by using their respective texts. After the vector representation is obtained, deep learning is used to classify words, in this case the deep learning methods used are Recurrent Neural Networks and Convolutional Neural Networks. The classification is carried out in 4 categories, namely electrical failure, human factor, miscellaneous, signal communication, and track. The results obtained using the word2vec method and LSTM methods authors applied were able to correctly classify the cause of a reported accident with overall 75 % accuracy.

In [7] uses word embedding and neural networks with LSTM (Long-Short Term Memory) to classify text where the criteria of classification are determined by the context of the application. The first data is preprocessed in order to obtain the appropriate data. The preprocessing process that is carried out is like converting data into tokens, where the frequency sequence of each token is used to be used as an index representation of a token and other processes. In this case, LSTM (Long-Short Term Memory) is used because traditional RNNs (Recurrent Neural Networks) cannot learn long-term dependencies in data due to exploding problems and gradient vanishing. LSTM (Long-Short Term Memory) is a modification of RNNs (Recurrent Neural Networks), vanishing gradient can be overcome by this method. LSTM (Long-Short Term Memory) trains its model using the back-propagation method. This study classifies different applications, namely classifying the service provider customer texts from social media as actionable or non-actionable. The social media used are Twitter, Facebook, Google+ as well as brand communities and online forums. Training data and test data are divided by the ratio 80:20. The data used in more than 30 different languages, the accuracy obtained is 86%.

After reviewing the existing works, the main steps for text classification from these works can be taken into consideration to be used in this study. The first main step would be to collect data that will be used in the study. The next step is to determine the word vectorization method and classifier. According to the results of the previous studies, using Word2vec and LSTM gives the best results when compared to other methods & therefore be used in this study.

III. METHODOLOGY

This study aims to build a classification model that is able to classify comment texts in Bahasa Indonesia into 5 different classes and 3 different classes to get sentiment from the users. As mentioned before, the 5 classes are based on a rating scale from 1-5 and a rating scale from 1-3. The following stages show the steps taken in this study to achieve its goal.

A. Data Collection

The data used was collected programmatically with python, by using automated web scraping techniques. Web mining is a technique for information retrieval, information extraction and data mining, information retrieval and information extraction have an important role in tracking and mining information that is useful for tracking and mining unstructured information from unstructured data, before it is further processed or through the pre-processing process [8]. Web scraping is a method of collecting information from the world wide web, with advances in this era, web scraping is used for various purposes with a semantic web vision that is based on a tool that crosses website constituent abstracts and stores it in a database [9]. The programming language used for the stated purposes is python, python is used because Python is rich in a set of modules that makes it easy for users and provides an interface between programs and Google web pages [10]. The data collected includes:

- Title: It is the name of the application provider.
- Username: the name of the user who made the comment.
- Rating: weight is given by the user after making a comment.
- Review_date: the time when the user makes a comment.
- Review_text: comments made by users.
- Reply_text: comment replies made by the admin.

The data that has been collected is then stored in excel or csv form. It then goes through preprocessing, after which only information that is useful for text classification should remain. The data is then divided into 2, training data and test data, after going through the preprocessing process in order to get the information needed, the data is trained using the word2vec extraction feature to get the semantics of each word and then classified using LSTM.

B. Preprocessing

In this process, the data that has been collected is cleaned so that it becomes the only data needed for text classification. There are several steps in preprocessing such as comment cleansing which is useful for removing unnecessary words and removing upper case letters in text and converting them to lower case letters for text classification, stopwords removal which is useful for removing conjunction words and other special characters such as emoticons that are not used for text classification, stemming which is useful for converting all words into basic words, and tokenization which is useful for making. Preprocessing of text usually uses the help of the Natural Language Toolkit (NLTK) library.

Preprocessing in other Indonesian language libraries is also used in this study. The library used to process Indonesian language is the Sastrawi library to carry out the stemming process and the stopword removal process [11]. The difficulty in preprocessing text using Indonesian is due to the large number of slang words in Indonesian language so it is necessary to add words to existing slang words to remove the words.

The goal for stopword removal, namely making text look heavier and reducing words that are less important for analysis, eliminating stopwords means reducing the dimensions of the term space, the most common words in the text are articles, prepositions, pronouns and others that don't give meaning to the document [12]. In text classification dimensions is one of the problems if it has high dimensions even after stopword removal and stemming filtering, high dimensions can mean incorrect text classification time, one way to reduce dimensions is preprocessing the text by using the selection feature or feature transformation [13].

TABLE I
AN EXAMPLE OF PREPROCESSING COMMENT

Original Comment	cukup membantu mengambil keputusan memilih paket2 yg tersedia :)
Comment Cleansing	cukup membantu mengambil keputusan memilih yg tersedia
Stopwords Removal	membantu mengambil keputusan memilih tersedia
Stemming	bantu ambil putus pilih sedia
Tokenization	['bantu', 'ambil', 'putus', 'pilih', 'sedia']

C. Imbalance Dataset

In the data retrieval process in data mining, the data obtained is not evenly distributed, so the amount of data availability is very important. In the supervised learning method, the data collected requires labels and classes on each data, so that the training data process can run properly, the same amount of data is needed for each data label class. This causes the data used to be imbalance. In this study, SMOTE (Synthetic Minority Oversampling Technique) were used. The SMOTE algorithm selects samples randomly from a whole set of samples from the minority category. The minority class categories that are located together are identified before the use of the new minority class, so that this technique is able to generate synthetic data rather than replicating the minority categories. Based on [14] steps on SMOTE are in below.

TABLE II
PARAMETERS USED IN TRAINING WORD2VEC MODEL

Parameter	Description	Value
size	Number of vector dimensions	100
window	Context word window size	5
workers	Number of parallel threads	8
min_count	Minimum word count	1
iter	Number of training iterations	50

Algorithm 1: SMOTE's Pseudocode

```

O is original data
P is instance from minority class
For each instance x in P
    Find k-nearest neighbors (minority classes) to x in P
    Get y by randomizing one from instance k
    Difference = x-y
    Gap = random number from 0 to 1
    n = x + difference * gap
    Plus n to O
End For

```

D. Word embedding using Word2vec

In this process, data that has passed the preprocessing process is used for the word embedding process. Implementing word2vec is used for the word embedding process using the library Gensim. This process uses preprocessing data in the form of tokenized text as input and produces a model as output. This model is a vector that represents words in vector form. The number of dimensions of the vector depends on the parameters that are set when training the model, the parameters used in this research can be seen in table 3. By using the Gensim's "most_similar" function with 1 target word as an input and the function returns the 10 most similar words from the target word. In this study the word used is "bagus" which means "good quality" and the model shows relatively similar results to the word "bagus" because it uses the "most_similar" function. The result of this testing can be seen in Figure 1. The results obtained are helpful, very, informative, not bad, really, beneficial, help, easy, useful and good

Selection for dimensions in word2vec is very important because it relates to two aspects, model accuracy and computational resources [15]. The large number of parameters causes memory resources to be wasted, so it is necessary to design a method that can decide which dimensions meet the ability to capture the required information from training data and use efficient computing resources. There are 2 methods in word2vec, namely the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model, CBOW tries to predict the target word from the context, Skip-Gram actually aims to determine the context for a particular target word, 2 approaches only modify input and output but it doesn't change anything, the network doesn't change and training always takes place between single word pairs (placed as one-hot input and output) [16].

```

In [30]: model.wv.most_similar('bagus')

Out[30]: [('membantu', 0.936974287033081),
('sangat', 0.9364749193191528),
('informatif', 0.9290339946746826),
('lumayan', 0.9251346588134766),
('banget', 0.9239262938499451),
('bermanfaat', 0.9215588569641113),
('bantu', 0.9184792041778564),
('kemudahannya', 0.9179884195327759),
('berguna', 0.9132888317108154),
('baik', 0.9112176895141602)]

```

Figure 1 Word2vec most similar result

E. Long Short Term Memory

Long-Short Term Memory (LSTM) is a method that is often used for many natural language processing (NLP) tasks. Long-Short Term Memory (LSTM) is a modification of the RNN [17]. LSTM is very suitable for solving classification problems, processing and also prediction of time for unknown duration. LSTM is widely used in various studies because it has the ability to remember data that existed in the past and eliminate data that is no longer considered important. LSTM (Long-Short Term Memory) is a modification of RNNs (Recurrent Neural Networks), vanishing gradient and exploding gradient can be solved by this method which previously became a limitation on RNN. LSTM (Long-Short Term Memory) trains its model using the back-propagation method. The problem of vanishing gradient is when the network tries to spread all information through a long sequence, in which the gradient will decrease during back propagation through time, so that LSTM can be used for this first problem where LSTM can maintain long-term information by only spreading the parts of it which are relevant [18]. LSTM can avoid this problem by managing the information available in the cell state by using 3 different gates, namely input gate, output gate and forget gate.

Figure 2 is the architecture of LSTM process.

Figure 2: LSTM Process Overview

Each term T_i is converted into a corresponding input vector x_i using the word2vec model and fed into the LSTM individually, every time to j , the output W of the hidden layer H_j will be propagated to the hidden layer along with the next input x_{j+1} at the next time the final output W_n will be fed to the output layer. Each term in T_1 is converted into the appropriate input vector form x_i using word2vec and entered into the LSTM one by one [19]. To fulfill the sequential input of LSTM, first the word is converted into 3 dimensions A, B, C where A is the word embedding dimension of word2vec, B is the number of words, C is the max number of words. To avoid a long training time process, this study uses a single hidden layer with 100 neurons in the input layer. The number of neurons in the output layer is the output of the number of classes, in this case is 5 and 3. Figure 3 is the idea of LSTM.

Figure 3: LSTM Idea

TABLE III
DATA COUNT FOR 5 CLASSES

Rating	Training	Testing
1	2197	567
2	438	108
3	746	178
4	2023	475
5	7700	1927

In this paper also using 3 classes with detailed new rating 1 as negative is the combination of rating 1 and 2, new rating 2 as neutral is the rating 3, and new rating 3 as positive is the combination of rating 4 and 5 to get sentiment from user comment. The Training data and testing data distribution is shown in table 4. In this paper, there are 4 experiments, training data with 5 classes using LSTM, training data with 5 classes using LSTM and SMOTE, training data with 3 classes using LSTM and training data with 3 classes using LSTM and SMOTE.

B. Parameters

In deep learning, we have to define several parameters to improve classification accuracy. In addition, the parameters must be determined experimentally for each task because the optimal value is different for each task. In this experiment, we determine the value of these parameters. Table 5 shows the topology of the networks used in this study and table 6 shows the training parameters used in this study.

IV. RESULT ANALYSIS

A. Datasets

The first dataset used in this study is data from play.google.com from the myTelkomsel application using web scraping techniques, the data used is 16,359 data. The word vector model used is 100 dimensions and is trained to use word2vec with 100 iterations. The results shown by the trained data can be seen using the gensim library such as "most_similar" and "distance". The data is divided into 80:20 for training and testing data. Dataset used in this study is data from play.google.com from the myTelkomsel application using web scraping techniques, the data used is 16,369 data. Training data and testing data distribution is shown in table 3

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 11, Issue 05, May 2021)

TABLE IV
DATA COUNT FOR 3 CLASSES

Rating	Training	Testing
1	2644	628
2	654	177
3	8702	2214

TABLE V
TOPOLOGY OF THE NETWORKS

Parameter	Description	Value
Embedding Layer	Used for artificial input neuron and carry initial data into the system for processing by the next layer of artificial neurons. The fixed number of dimensions in the word embedding model can facilitate more efficient computation [20]. Because using word2vec, the weight used is the result of pretrained from word2vec.	$Input_dim = 6579(vocab_size)$ $Output_dim = 100(embedding_size)$ $Weights = w2v_weights$ $Input_length = 200$
LSTM	a sequential processing model consisting of two LSTMs.	100
Dense Layer	Deeply connected layers of neural networks to perform operations from input and return output. The dense layer contains the output in this case the number of the ratings.	5 (n_categories) 3 (n_categories)

C. Results

The data obtained was split into training data and test data with a ratio of 80:20 respectively for training and testing. In this paper, there are 4 experiments, training data with 5 classes using LSTM, training data with 5 classes using LSTM and SMOTE, training data with 3 classes using LSTM and training data with 3 classes using LSTM and SMOTE.

TABLE VI
PARAMETERS USED IN TRAINING DEEP LEARNING MODEL

Parameter	Description	Value
Epochs	The number of passes of the entire training dataset.	50
Activation	Responsible for converting the summed weighted input from the node to the node activation or output for that input. Softmax is used as an activation function for multi-class classifications where class membership is more than two class labels.	Softmax
Optimizer	Methods used to change neural network attributes such as weight and speed of learning to reduce losses. Adam's method is very efficient when dealing with large problems involving a lot of data or parameters, it also takes up less memory and is efficient.	Adam
Loss	Loss is the penalty for a bad prediction, in this study using sparse_categorical_crossentropy because in this study using multiclass.	sparse_categorical_crossentropy

The first experiment using a 5 classes dataset and LSTM, the best f1 score result is 0.62 and the best precisions result is 0.60. Table 7 shows the results for the confusion matrix in this first experiment. Based on the table on confusion matrix, the predicted data tends to the majority of data, this is because the amount of data in training is imbalance. For example, the model correctly predicts 6 instances of rating 2 and confuses rating 2 with rating 5 in another 31 instances. Another example, the model correctly predicts 7 instances of rating 3 and confuses it with rating 5 in 115 instances. In the second experiment using SMOTE to balance the dataset. SMOTE is a method for generating new instances using existing ones from a minority class. SMOTE works by adding existing points around its neighbors [21].

The second experiment using a 5 classes dataset and LSTM+SMOTE, the best f1 score result is 0.62 and the best precisions result is 0.70. Table 8 shows the results for the confusion matrix in this second experiment. Based on the confusion matrix, the second experiment results, it shows the results is more distributed, because it uses a balanced dataset.

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 11, Issue 05, May 2021)

TABLE VII
FIRST CONFUSION MATRIX

	Predicted					
		1	2	3	4	5
Actual	1	384	36	33	16	98
	2	59	6	5	7	31
	3	42	4	7	10	115
	4	47	4	5	14	405
	5	44	5	23	20	1835

There was an increase in precision in the second experiment. In the second experiment at rating 2, the model correctly predicts 38 instances of rating 2 and confuses rating 2 with rating 5 in another 14 instances. Another example, the model correctly predicts 47 instances of rating 3 and confuses it with rating 5 in 75 instances. The prediction on the data becomes distributed because the training data uses balanced data using oversampling. The amount of data and data that is not so clean can affect the performance of the model during classification.

TABLE VIII
SECOND CONFUSION MATRIX

	Predicted					
		1	2	3	4	5
Actual	1	450	49	39	1	28
	2	43	38	13	0	14
	3	44	12	47	0	75
	4	55	33	70	6	311
	5	81	53	240	1	1552

The third experiment using 3 classes dataset and LSTM, the best f1 score result is 0.83 and the best precisions result is 0.82. Table 9 shows the results for the confusion matrix in this third experiment. Based on the table on confusion matrix, the predicted data tends to the majority of data, this is because the amount of data in training is imbalance. For the example in rating 2, the model correctly predicts as rating 2 is 5 and predicts as rating 3 is 106.

The fourth experiment using 3 classes dataset and LSTM+SMOTE, the best f1 score result is 0.86 and the best precisions result is 0.87. Table 10 shows the results for the confusion matrix in this fourth experiment.

TABLE IX
THIRD CONFUSION MATRIX

		Predicted		
		1	2	3
Actual	1	459	24	145
	2	66	5	106
	3	84	14	2116

Based on the confusion matrix, the fourth experiment results, it shows the results is more distributed, because it uses a balanced dataset. For the example in rating 2, the model correctly predicts as rating 2 is 66 and predicts as rating 3 is 66. The prediction on the data becomes distributed because training data uses balanced data using oversampling.

TABLE X
FOURTH CONFUSION MATRIX

		Predicted		
		1	2	3
Actual	1	501	71	56
	2	45	66	66
	3	88	138	1988

V. CONCLUSION

In this study, 4 experiments were conducted to obtain a comparison of the results, first experiment with 5 classes using LSTM, second experiment with 5 classes using LSTM+SMOTE, third experiment with 3 classes using LSTM, fourth experiment with 3 classes using LSTM+SMOTE. Best f1 score result for 5 classes is 0.62 and the best precisions result is 0.70 using LSTM and SMOTE. Best f1 score result for 3 classes is 0.86 and the best precisions result is 0.87 using LSTM and SMOTE. The unbalanced dataset produces accuracy results that are more likely to the majority class. The handling imbalance method can improve the distribution of the predicted results more distributed than the imbalance data. With the results, it is expected that the model can be used on social media that does not require users to score a review, so that the model can be used to automatically classify comments according to the prediction of the rating. So that in the future it is expected that user comment in provider application in Indonesian language text classification can be carried out in a comment even though the user does not include a rating and it is expected that this model can simplify the user comment text classification process for business actors in order to evaluate the products produced.

International Journal of Emerging Technology and Advanced Engineering

Website: www.ijetae.com (E-ISSN 2250-2459, Scopus Indexed, ISO 9001:2008 Certified Journal, Volume 11, Issue 05, May 2021)

For future works, use other word embedding methods such as fasttext or glove to extract the word embedding. Use other deep learning algorithms can also be used such as CNN, LSTM, etc. Use or modify other hyperparameters to improve the performance of the model. Making a dictionary for slang words in Indonesian is also recommended to help improve the accuracy of the classification carried out.

References

- [1] L. Chen, R. Wang, "Trust Development and Transfer from Electronic Commerce to Social Commerce: An Empirical Investigation," *American Journal of Industrial and Business Management*, 06(05), 568–576, 2016, doi:10.4236/ajibm.2016.65053.
- [2] S. Kim, H. Park, "Effects of various characteristics of social commerce (s-commerce) on consumers' trust and trust performance," *International Journal of Information Management*, 33(2), 318–332, 2013.
- [3] Y. Zhong, "Association for Information Systems AIS Electronic Library (AISeL) Social Commerce: A New Electronic Commerce Recommended Citation," 2012.
- [4] F. Taheri, A. Shourmasti, "International Academic Journal of Business Management Effects of various characteristics of social commerce on consumers' trust and trust performance," *International Academic Journal of Business Management*, 3(3), 20–26, 2016.
- [5] R. Kato, H. Goto, "Categorization of web news documents using Word2Vec and deep learning," *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 8–10 March, 476–481, 2016.
- [6] M. Heidarysafa, K. Kowsari, L. Barnes, D. Brown, "Analysis of Railway Accidents' Narratives Using Deep Learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE: 1446–1453, 2018.
- [7] A. Rao, N. Spasojevic, "Actionable and political text classification using word embeddings and lstm," *ArXiv Preprint ArXiv:1607.02501*, 2016.
- [8] A. V. Saurkar, S.A. Gode, "An Overview On Web Scrapping Techniques And Tools," *International Journal on Future Revolution in Computer Science & Communication Engineering*, 363–367, 2018.
- [9] V. Singrodia, A. Mitra, S. Paul, "A Review on Web Scrapping and its Applications," *2019 International Conference on Computer Communication and Informatics, ICCCI 2019*, 1–6, 2019, doi:10.1109/ICCCI.2019.8821809.
- [10] D. Pratiba, M.S. Abhay, A. Dua, G.K. Shanbhag, N. Bhandari, U. Singh, "Web Scrapping and Data Acquisition Using Google Scholar," *Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018*, 277–281, 2018, doi:10.1109/CSITSS.2018.8768777.
- [11] E. Setiani, W. Ce, "Text Classification Services Using Naïve Bayes for Bahasa Indonesia," *Proceedings of 2018 International Conference on Information Management and Technology, ICIMTech 2018*, (September), 361–366, 2018, doi:10.1109/ICIMTech.2018.8528258.
- [12] S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, M. Nithya, S. Kannan, V. Gurusamy, "Preprocessing Techniques for Text Mining," *International Journal of Computer Science & Communication Networks*, 5(1), 7–16, 2015.
- [13] R. Sergienko, M. Shan, W. Minker, "A comparative study of text preprocessing approaches for topic detection of user utterances," *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 1826–1831, 2016.
- [14] P. Jeatrakul, K.W. Wong, C.C. Fung, "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm," in *International Conference on Neural Information Processing*, Springer: 152–159, 2010.
- [15] P.T. Hung, K. Yamanishi, "Word2vec Skip-gram Dimensionality Selection via Sequential Normalized Maximum Likelihood," 2020.
- [16] G. Di Gennaro, A. Buonanno, A. Di Girolamo, A. Ospedale, F.A.N. Palmieri, G. Fedele, "An Analysis of Word2Vec for the Italian Language," *Smart Innovation, Systems and Technologies*, 184, 137–146, 2021, doi:10.1007/978-981-15-5093-5_13.
- [17] K. Lu, P. Mardziel, K. Leino, M. Fedrikson, A. Datta, "Influence Paths for Characterizing Subject-Verb Number Agreement in LSTM Language Models," *ArXiv Preprint ArXiv:2005.01190*, 2020.
- [18] S. Santhanam, "Context based Text-generation using LSTM networks," *ArXiv Preprint ArXiv:2005.00048*, 2020.
- [19] J.-H. Wang, T.-W. Liu, X. Luo, L. Wang, "An LSTM approach to short text sentiment classification with word embeddings," in *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018)*, 214–223, 2018.
- [20] B. Jang, M. Kim, G. Harerimana, S.U. Kang, J.W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism," *Applied Sciences (Switzerland)*, 10(17), 2020, doi:10.3390/app10175841.
- [21] A. Glazkova, "A Comparison of Synthetic Oversampling Methods for Multi-class Text Classification," *ArXiv Preprint ArXiv:2008.04636*, 2020.