UNIVERSITAS
GADJAH MADA

Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding
LUCAS S. AJI DHARMA, Drs. Edi Winarko, M.Sc., PhD
Universitas Gadjah Mada, 2021 | Diunduh dari http://etd.repository.ugm.ac.id/

APPROVAL PAGE i
PREFACE ii
TABLE OF CONTENTS iii
TABLE OF TABLES v
TABLE OF FIGURES vi
ABSTRACT 10
CHAPTER I INTRODUCTION 1
1.1 Research Background 1
1.2 Research Problem 4
1.3 Research Objective 5
1.4 Research Scope 5
1.5 Research Benefits 6
CHAPTER II LITERATURE REVIEW 7
CHAPTER III BASIC THEORY 17
3.1 Text Mining 17
3.2 Preprocessing 17
3.2.1 Case folding 18
3.2.2 Removal of special symbols and numbers 18
3.2.3 Stemming 18
3.2.4 Stopword removal 18
3.2.5 Tokenization 18
3.3 Word Embedding 19
3.4 Convolutional Neural Network 22
3.4.1 Input layer 23
3.4.2 Convolution layer 23
3.4.3 Pooling layer 23
3.4.4 Dropout layer 24
3.4.5 Fully-Connected layer 24
3.5 Performance Evaluation 25
CHAPTER IV RESEARCH METHODOLOGY 28
4.1 Research Description 28
4.2 Data Collection and Labeling 29
4.3 Data Preprocessing 31
4.4 Feature Extraction 33
4.4.2 BERT Embedding 33
4.4.3 Word2Vec 35
4.5 Modelling 36
4.5.1 Convolutional Neural Network 36
4.5.2 Keras Tuner 38
4.6 Model Evaluation 39

UNIVERSITAS
GADJAH MADA

Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding
LUCAS S. AJI DHARMA, Drs. Edi Winarko, M.Sc., PhD
Universitas Gadjah Mada, 2021 | Diunduh dari http://etd.repository.ugm.ac.id/

CHAPTER V IMPLEMENTATION 40
5.1 Software Tools 40
5.2 Data Collection 40
5.3 Preprocessing 41
5.3.1 Case folding 42
5.3.2 Removal of URLs, Mention, Hashtags, Emojis, Whitespace, and Numbers 42
5.3.3 Removal of duplicates 43
5.3.4 Stemming & Stopwords Removal 43
5.4 Initializing to a TPU runtime (Tensor Processing Unit) 44
5.5 Feature Extraction 45
5.5.1 Data Splitting 45
5.5.2 BERT Model 45
5.5.3 BERT Tokenizer 46
5.5.4 Word2Vec 47
5.6 Modeling 49
5.6.1 Implementation of BERT Embedding with CNN Model 49
5.6.2 Implementation of CNN with Word2Vec embedding 52
5.6.3 Keras Tuner 53
5.8 Evaluation 55
5.8.1 Evaluation of the Deep Learning model 56

CHAPTER VI RESULTS AND DISCUSSION 58
6.1 Results of Preprocessing 58
6.2 Results of Data Exploration 59
6.3 Training and Validation Results 62
6.3.1 CNN with BERT Embedding 62
6.3.2 CNN with Word2Vec 65
6.3.3 Keras tuner results 68
6.4 Evaluation Results 68

CHAPTER VII CONCLUSION 77
7.1 Conclusions 77
7.2 Suggestions 77

REFERENCES 78

UNIVERSITAS
GADJAH MADA

Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding
LUCAS S. AJI DHARMA, Drs. Edi Winarko, M.Sc., PhD
Universitas Gadjah Mada, 2021 | Diunduh dari http://etd.repository.ugm.ac.id/

LIST OF TABLES

Table 2.1 Comparison table between classification of disaster tweet ……………………….11

Table 4.1 Sample of raw tweets collected …………………………………………………..30

Table 4.2 Sample of the labelled data ………………………………………………………30

Table 4.3 Example of case folding ……………………………………………………..........31

Table 4.4 Example of removal special symbols ……………………………………………..32

Table 4.5 Example of tokenization …………………………………………………….........32

Table 4.6 Example of stopwords removal …………………………………………………...32

Table 4.7 Example of stemming ……………………………………………………………..32

Table 6.1 sample of text that has been preprocessed with stemming and stopwords removal ……………………………………………………………………………………………..58

Table 6.2 sample of texts that has been preprocessed without applying stemming and stopwords removal ………………………………………………………………………..59

Table 6.3 The amount of different types of natural disaster in the data ……………………..59

Table 6.4 Example classified text of the nine different disaster classes ……………………..60

Table 6.5 Result of the two dataset using CNN and BERT Embedding ……………………64

Table 6.6 shows the result of CBOW and Skip-gram CNN model tested using the two dataset ……………………………………………………………………………………………..67

Table 6.7 Hyperparameters used for the CNN model ………………………………………68

Table 6.8 Model evaluation without stemming and stopwords removal ……………………………………………………………………………………………69

Table 6.9 Model evaluation with stemming and stopword removal ………………………..69

Table 6.10 Text from gempa, banjir, and other class as False Positive by the BERT-CNN model ………………………………………………………………………………………...73

Table 6.11 False Negative text from ‘tsunami’ class predicted as ‘gempa’ by the BERT-CNN model ………………………………………………………………………………………...74

Table 6.12 Example texts from awan panas, longsor, kekeringan, puting beliung, and erupsi class predicted by the BERT-CNN model ………………………………………………...75

UNIVERSITAS
GADJAH MADA

Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding
LUCAS S. AJI DHARMA, Drs. Edi Winarko, M.Sc., PhD
Universitas Gadjah Mada, 2021 | Diunduh dari http://etd.repository.ugm.ac.id/

LIST OF FIGURES

Figure 1.1 Example of natural disaster tweet from @BNPB_Indonesia twitter page ………..3

Figure 1.2 Example of natural disaster tweet from @infoBMKG twitter page……………….3

Figure 1.3 Example of a non-disaster tweet from @kompascom twitter page ……………….3

Figure 3.1 Example of Text mining process ……………………………………...…....17

Figure 3.2 Example of CBOW and Skip-Gram architecture ……………………….......20

Figure 3.3 Example of BERT architecture ……………………………………...….21

Figure 3.4 Example of BERT input representation …………………………………....21

Figure 3.5 Example of a CNN architecture …………………………………………….....22

Figure 3.6 Applying dropout layer in an Neural Network model …………………………...24

Figure 3.7 Proposed system architecture of the deep learning CNN model ………………...25

Figure 3.8 Example of a confusion matrix ……………………………………………….26

Figure 4.1 Summary of Research flow diagram ………………………………………….29

Figure 4.2 Simple computation of BERT embedding …………………………………….34

Figure 4.3 Proposed CNN model architecture using BERT embedding

……………………………………………………………………………………….35

Figure 4.4 Simple computation of Word2Vec embedding ………………………………..35

Figure 4.5 Proposed CNN model architecture using Word2vec embedding

………………………………………………...36

Figure 4.6 Proposed general CNN model architecture created using LetNet style

……………………………………………………………………………………….38

Figure 5.1 Implementation of TWINT in python ……………………………………...41

Figure 5.2 Library used for preprocessing ………………………………………………..42

Figure 5.3 Implementation of Case folding ……………………………………………….42

Figure 5.4 Implementation of Preprocessing ……………………………………………..43

Figure 5.5 Implementation of removal of duplicates ……………………………………..43

Figure 5.6 Importing Sastrawi library and applying Stemming …………………………….43

Figure 5.7 Importing NLTK library for Indonesian Stopwords and Implementation of Stemming and Stopwords removal ………………………………………………………44

Figure 5.8 Initialization of TPU runtime ………………………………………………….45

Figure 5.9 Implementing data splitting using train-test split from sklearn library …………45

Figure 5.10 Implementation of indobert model ……………………………………………..46

UNIVERSITAS
GADJAH MADA

Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding
LUCAS S. AJI DHARMA, Drs. Edi Winarko, M.Sc., PhD
Universitas Gadjah Mada, 2021 | Diunduh dari http://etd.repository.ugm.ac.id/

Figure 5.11 Implementation of applying the BERT tokenizer ………………………………....46

Figure 5.12 Importing the Gensim library to create the Word2Vec Embedding…………......47

Figure 5.13 Shows the similarity between words from the Skip-gram model ……………....48

Figure 5.14 Implementing Tokenization and Padding to the dataset ………………………48

Figure 5.15 Implementation of the embedding matrix ……………………………………....48

Figure 5.16 Creating the CNN with BERT Embedding model ……………………………...49

Figure 5.17 Fully build CNN with BERT Embedding ……………………………………....51

Figure 5.18 Implementation of training and testing of the CNN model …………………....51

Figure 5.19 Creating the CNN with Word2Vec Embedding model……………………….....52

Figure 5.20 Creating the embedding layer from the embedding matrix ……………………52

Figure 5.21 Fully build CNN with Word2Vec Embedding…………………………………..53

Figure 5.23 Installing and importing the Keras-tuner library ……………………………....53

Figure 5.24 Implementing keras tuner into the CNN BERT embedding model …………….54

Figure 5.25 Building the Keras tuner from the CNN-BERT model…………………………55

Figure 5.26 Running the keras-tuner and its result ……………………………………………55

Figure 5.27 Shows the code for plotting the Model Accuracy graph ………………………56

Figure 5.28 Shows the code for plotting the Model Loss graph ……………………………56

Figure 5.29 Prediction evaluation of the CNN model ………………………………………57

Figure 6.1 Distribution of data label ………………………………………………………....61

Figure 6.2 Word Cloud of most frequent keyword within the Dataset ……………………..…62

Figure 6.3 Depicts the Model Loss graph for CNN with BERT Embedding………………..63

Figure 6.4 Depicts the Model Accuracy graph For CNN with BERT Embedding …………64

Figure 6.5 Depicts the Model Loss graph For CNN with Word2Vec Embedding…………...65

Figure 6.6 Depicts the Model Accuracy graph For CNN with Word2Vec Embedding……...66

Figure 6.7 Confusion matrix of the BERT-CNN model ……………………………………..72