ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

# Multilingual Short Text Categorization Using Convolutional Neural Network

Liriam Enamoto and Li Weigang*

Dept of Computer Science - University of Brasilia
Asa Norte, Brasília - DF - Brazil

**Abstract.** One of the most meaningful use of online social media is to communicate quickly during emergency. In case of global emergency, the threat might cross countries borders, affect different cultures and languages. This article aims to explore Convolutional Neural Network (CNN) for multilingual short text categorization in English, Japanese and Portuguese to identify useful information in social media. A CNN is constructed for this special purpose. The experiment results show that CNN model performs better than SVM even in small dataset. And more interestingly, the cross languages test suggests that English, Japanese and Portuguese text can use the same model with few hyperparameters changes.

## 1 Introduction

The use of social media has been extensively studied in public health emergency such as 2009 H1N1 pandemic [1], influenza virus [2] and more recently, in 2014/2015 Ebola outbreak [3]. According to the study of Hu et al. [4], contagion of Ebola Virus Disease (EVD) can be reduced if safe burial procedures are adopted in less than 34 hours after death. Failure to provide correct information about prevention and virus transmission in the Ebola outbreak early phase contributed to the virus spread in West Africa increasing people concern and anxiety around the world.

In case of risk and emergency like epidemics, communication is one of the fundamental tools for emergency management [5]. Furthermore, crisis and emergency events that occurs in one country might cross borders and continents affecting people who uses different languages.

Text categorization is the activity of labeling natural language text with thematic categories from a predefined set, and can be applied in document indexing, document filtering and in any application requiring document organization [6]. Traditional machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes, and Neural Network have made great advance in extracting and classifying text. In the past few years, many researches in Deep Neural Network such as Convolutional Neural Network (CNN) have demonstrated that it performs remarkably well in text categorization. However, most researches use CNN mainly for English text categorization [7, 8, 9] and a few studies explore multilingual text categorization with CNN mixing alphabetical and non-alphabetical languages, like Chinese and English [10], Japanese and English [11]. This research aims to explore this gap by using the special CNN model to categorize Twitter posts in English, Portuguese and Japanese about the same subject to identify useful information in public health emergency

* This research is partially supported under the grant 311441/2017-3, by CNPq/Brazil.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

occurred in 2014/2015 Ebola outbreak. As the main contribution of this research, our cross languages test suggests that English, Japanese and Portuguese text can use the same model with few hyperparameters changes.

## 2 Convolutional Neural Network and Related Works

### 2.1 Convolutional Neural Network

Convolutional Neural Network (CNN) consists of a sequence of one or multiple pairs of convolution and pooling layers. A convolution layer consists of several computational units, and each of which takes as input a region vector that represents a small region of the input image, and the small regions collectively cover the entire data [7]. A computational unit associates with the *l-th* region of input *x* calculates a feature $c_l$ (1), where $r_l(x)$ is the input region vector that represents the *l-th* region, *W* represents the weight matrix, *b* the bias, and $\sigma$ represents a nonlinear activation function such as Rectified Linear Units (ReLU).

$$c_l = \sigma ( W . r_l(x) + b ) \quad (1)$$

The matrix of weights *W* and the vector of biases *b* are learned through training, and they are shared by computational units in the same layer [7]. The output image of convolution layer is passed to a pooling layer, which shrinks each region of the image into one unit by computing the average or maximum value of each region [7]. The idea of pooling layer is to capture the most important feature of each region. These features from the last pooling layer are passed to a fully connected layer, which returns a prediction based on features learned internally by previous layers [8].

Using CNN for text analysis, each sentence of input data is transformed into a matrix of word embedding [12]. Word embedding is a distributed representation of words that reduce data sparsity problem [13] and can be trained as part of CNN training or adopt pre-trained corpus such as word2vec [14]. Each convolution layer has a variable number of computational units, each unit corresponding to a small region (one or more words) from the input text [8]. Similarly to CNN for image, CNN for text can be composed by one or multiple pairs of convolution and pooling layers followed by fully connected layer which returns the prediction result for the input.

### 2.2 Convolutional Neural Network for text categorization

Originally used for computer vision, CNN has shown to be effective for natural language processing achieving better results than traditional machine learning algorithms in text categorization. The shallow CNN model proposed by Kim et al. [9] composed by one convolution layer and one pooling layer built on top of pre-trained corpus word2vec performed better than SVM or even more sophisticated deep learning model with complex pooling schemes. The experiment of Johnson et al. [7] used parallel CNN to perform sentiment analysis and topic classification. The results suggested that with the parallel CNN model, several types of word embedding can be learned and combined for higher accuracy. Furthermore, Johnson et al. [7] argue that the strength of CNN is that n-grams (or region of n words) can contribute to accurate

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

prediction even if they did not appear in the training data, as long as their constituent words did.

Researches of CNN also explored valuable information posted on social media in disaster scenarios. Caragea et al. [8] used CNN to detect useful information posted on Twitter during some crisis events. Our CNN model is constructed as figure 1.

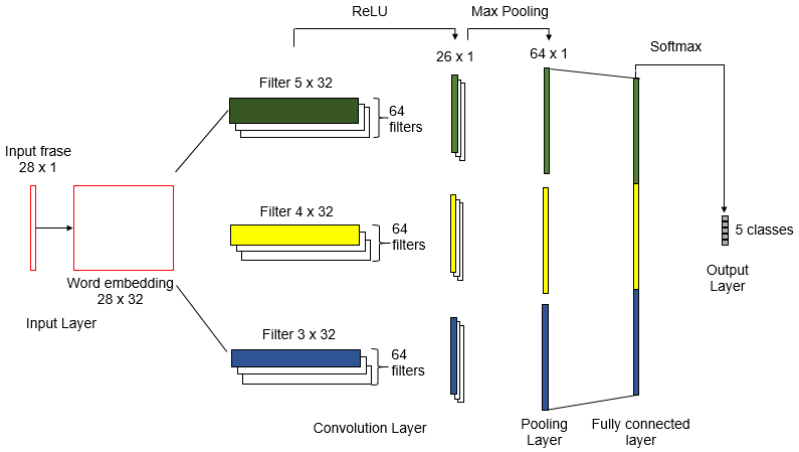


Fig. 1: CNN (1+1) model with one convolution layer and one pooling layer for text

## 3 Experiments

### 3.1 The Dataset

The datasets used in this research were collected from Twitter during the 2014/2015 Ebola outbreak using the keyword "ebola". From total of one million tweets downloaded during 6 months, 1162 tweets in English, 246 tweets in Portuguese and 157 tweets in Japanese where manually annotated into five categories following the rules described in Table 1.

| Category | Description |
|---|---|
| Outbreak situation report | Reports about newly confirmed Ebola cases and deaths, official announcement about Ebola free countries. |
| Informative posts | Hospitals prepared for Ebola patients, vaccine researches, virus prevention information, donation campaign. |
| Negative impact | Social and economic impact caused by Ebola. |
| Negative information | Criticism against the government, panic, racism. |
| Need for preparedness | Lack of hospitals, body bags, food, and safety funeral protocols. |

Table 1: Twitter posts classification criteria.

Table 2 details each dataset used in this experiment. The column #Tweets shows the number of tweets, the column Examples shows Twitter posts about the outbreak,

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

and the column Translation shows the English translation of Portuguese and Japanese tweets. Even representing a small dataset, the annotated Twitter posts used in this experiment is public available for further multilingual text classification researches[†].

| Dataset | #Tweets | Examples | Translation |
|---|---|---|---|
| English tweets | 1162 | Ebola death toll tops 10000. | - |
| Portuguese tweets | 246 | Ebola mata mais de 10 mil pessoas | Ebola kills more than 10000 people.. |
| Japanese tweets | 157 | 死者８０００人超に＝西アフリカの#エボラ熱ー | More than 8000 deaths in West Africa #Ebola |

Table 2: Twitter datasets.

## 3.2 **CNN Model and Hyperparameters**

The CNN model was used in the experiment, and denoted hereinafter as CNN (1+1) was composed by one convolution layer and one pooling layer based on Kim et al. [9] previous study, also see figure 1.

The CNN architecture is the same for English, Portuguese and Japanese datasets except for few details. The vocabulary was built based on the words contained in each dataset. For English dataset word level approach was used resulting in a vocabulary size of 2504 words and maximum tweet length of 28 words. For Portuguese dataset the same word level approach was used with vocabulary size of 1217 words and maximum tweet length of 31 words. Table 3 details the hyperparameters used in each dataset. For English and Portuguese datasets the dimensionality of input embedding was set to 32, filter windows of {3, 4, 5} with 64 filters each, so that filters slides over 3, 4, and 5 words with no padding and stride set to 1. Batch-size of 128 for English and 64 for Portuguese. For all datasets Rectified Linear Units was applied as activation function, Adam optimizer with learning rate 0.001 and #epoch 400 was used. For regularization purpose, dropout rate was set to 0.4 and L2 lambda to 0.4. Max pooling was used in the pooling layer. Then the pooling layer vectors were concatenated into a fully connected layer where softmax was applied to finally categorize each input into one of five classes described in Table 1.

| Dataset | Vocab. size | Max. length | Embedding size | Filter size | Batch size | Drop out | L2 lambda |
|---|---|---|---|---|---|---|---|
| English | 2504 | 28 | 32 | 3,4,5 | 128 | 0.4 | 0.4 |
| Portuguese | 1217 | 31 | 32 | 3,4,5 | 64 | 0.4 | 0.4 |
| Japanese | 769 | 140 | 128 | 3,4,5 | 64 | 0.4 | 0.2 |

Table 3: Dataset details and hyperparameters.

Japanese text usually does not have space separation between words as shown in Table 2 example making difficult to apply language processing methods that assume

[†] https://github.com/enmili/multilingualDataset

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

word as the basic construct. In order to produce better results, character level approach was used for Japanese dataset as suggested by Zhang et al. [15]. The dimensionality of input embedding was set to 128, filter windows of {3, 4, 5} with 64 filters each with no padding and stride set to 1, batch-size of 64. For regularization purpose, dropout was set to 0.4 and L2 lambda to 0.2.

## 4 Results

This section presents the results of CNN (1+1) model and compare the results with SVM baseline model for English tweets, Portuguese tweets and Japanese tweets. In addition, the English dataset used in Caragea et al. [8] research was processed into CNN (1+1) model for evaluation purpose.

Table 3 shows the experiment results where the rows contain the models and the columns contain the different datasets. The accuracy of SVM was 0.736 in English dataset, 0.625 in Portuguese dataset, and 0.600 in Japanese dataset. On the other hand, the accuracy of CNN (1+1) was 0.767 in English dataset, 0.791 in Portuguese dataset, and 0.8 in Japanese dataset. The experiment suggests that even with a small dataset and a simple CNN architecture with one convolution layer and one pooling layer, by adjusting the hyperparameters it is possible to get a superior result (4.21% higher in English dataset, 26.56% in Portuguese dataset and 33.33% in Japanese dataset) comparing with traditional SVM model.

For evaluation purpose, English Twitter dataset related to Philipine floods (2012), Colorado floods (2013), Queensland floods (2013) and Manila floods (2013) used in Caragea et al. [8] research and available at CrisisLex project where processed through the CNN (1+1) model. The flood dataset has 3671 tweets manually annotated into informative and noninformative posts. The accuracy of CNN (1+1) model was 0.837 which was slightly greater than 0.825 reported by Caragea et al. [8] which uses similar CNN architecture with one convolution layer and one pooling layer.

| Model | English tweets | Portuguese tweets | Japanese tweets | CrisisLex (English tweets) |
|---|---|---|---|---|
| CNN(1+1) | 0.767 | 0.791 | 0.800 | 0.837 |
| SVM | 0.736 | 0.625 | 0.600 | - |
| CNN[8] | - | - | - | 0.825 |

Table 4: Classification results.

## 5 Conclusion and Future Works

This article explored English, Japanese and Portuguese short text categorization about the same topic using a shallow CNN model.

The main contribution of this research is that with few hyperparameters changes the same CNN model outperforms SVM in multilingual text categorization even in small dataset. Crisis and emergency events that occurs in one country might cross borders and continents affecting people that communicate in different languages and

communication is one of the fundamental tools for emergency management. Exploring the same CNN model for multilingual short text categorization might help to reduce people anxiety and uncertainty during global emergency.

Currently there are three works in progress. The first work is to collect more data from online social media about a global topic which generate posts in different languages and create a more robust multilingual annotated dataset; the second work is to improve CNN (1+1) performance by model generalization and architecture change, and the last work is to use Recurrent Neural Network (RNN) to process multilingual short text. More contributions are expected as: generate a robust annotated dataset in English, Japanese and Portuguese to be used in other text categorization researches and improve our knowledge about CNN and RNN for multilingual text processing.

## References

[1] C. Chew and G. Eysenbach, Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak, *PloS one*, 5*(* 11), page e14118, November 2010.

[2] C. Corley, D. Cook, A. Mikler and K. Singh, Text and structural data mining of influenza mentions in web and social media, *International Journal of Environmental Research and Public Health*, 7:596-615, February 2010.

[3] K. Tefrie and K. Sohn, Automated disease outbreak detection and analysis, proceeding of the *Information Science and Applications (ICISA) 2016*, pages 985-993, Springer, Singapore, 2016.

[4] K. Hu, S. Bianco, S. Edlund and J. Kaufman, The impact of human behavioral changes in 2014 West Africa Ebola outbreak, proceeding of the *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, vol. 9021, pages 75-84, Springer, Cham, March 2015.

[5] T. Simon, A. Goldberg and B. Adini, Socializing in emergencies—A review of the use of social media in emergency situations, *International Journal of Information Management*, 35:609-619, October 2015.

[6] F. Sebastiani, Machine learning in automated text categorization, *ACM computing surveys (CSUR)*, 34:1-47, March 2002.

[7] R. Johnson and T. Zhang, Effective use of word order for text categorization with convolutional neural networks, *arXiv preprint arXiv:1412.1058*, December 2014.

[8] C. Caragea, A. Silvescu and A. Tapia, Identifying informative messages in disaster events using convolutional neural networks, proceeding of the *International Conference on Information Systems for Crisis Response and Management*, pages 137-147, May 2016.

[9] Y. Kim, Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882*, August 2014.

[10] X. Zhang, J. Zhao and Y. LeCun, Character-level convolutional networks for text classification, proceedings of the *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 649-657, 2015.

[11] M. Sato, R. Orihara, Y. Sei, Y. Tahara and A. Ohsuga Minato, Japanese text classification by character-level deep ConvNets and transfer learning, proceeding of the *International Conference on Agents and Artificial Intelligence*, pages 175-184, SCITEPRESS, 2017.

[12] J. Wang, Z. Wang, D. Zhang and J. Yan, Combining knowledge with deep convolutional neural networks for short text classification, proceedings of the *26th International Joint Conference on Artificial Intelligence*, pages 2915-2921. AAAI Press, August 2017.

[13] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, A neural probabilistic language model, *Journal of Machine Learning Research*, 3:1137-1155, February 2003.

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, proceedings of the *Advances in Neural Information Processing Systems 26 (NIPS 2013)* , pages 3111-3119, 2013.

[15] X. Zhang and Y. LeCun, Which encoding is the best for text classification in Chinese, English, Japanese and Korean?, *arXiv preprint arXiv:1708.02657*, August 2017