AMR-CNN: Abstract Meaning Representation with Convolution Neural Network for Toxic Content Detection

Ermal Elbasani[1] and Jeong-Dong Kim[1,2,*]

[1]*Department of Computer Science and Engineering, Sun Moon University, Asan, 31460 South Korea*
[2]*Genome-based BioIT Convergence Institute, Sun Moon University, Asan, 31460 South Korea*
E-mail: ermal.elbasani@gmail.com; kjdvhu@gmail.com
*Corresponding Author

Abstract

Recognizing the offensive, abusive, and profanity of multimedia content on the web has been a challenge to keep the web environment for user's freedom of speech. As profanity filtering function has been developed and applied in text, audio, and video context in platforms such as social media, entertainment, and education, the number of methods to trick the web-based application also has been increased and became a new issue to be solved. Compared to commonly developed toxic content detection systems that use lexicon and keyword-based detection, this work tries to embrace a different approach by the meaning of the sentence. Meaning representation is a way to grasp the meaning of linguistic input. This work proposed a data-driven approach utilizing Abstract meaning Representation to extract the meaning of the online text content into a convolutional neural network to detect level profanity. This work implements the proposed model in two kinds of datasets

from the Offensive Language Identification Dataset and other datasets from the Offensive Hate dataset merged with the Twitter Sentiment Analysis dataset. The results indicate that the proposed model performs effectively, and can achieve a satisfactory accuracy in recognizing the level of online text content toxicity.

1 Introduction

Since computer networks have grown rapidly and the many communication and business transactions have changed their system in an online way, the number of internet users worldwide also increases rapidly. According to the Internet World Stats (Internet World Stats, 2019), the number of internet users worldwide has increased up from 3.92 billion to 4.13 billion in 2019. Along with that, in modern culture, social networking or social media has become commonplace to connect with others, share information, and pursue common interests with other worldwide users. On the other side, this advancement also brought an issue with the safety of online social media users from cyber-bullying or abuse in online media. In order to tackle the issue and to build a healthy and safe online environment, social media companies like Facebook, Google or Yahoo are actively taking part in researching for automated toxic content detection. The term 'toxic content language' is one of the terms that determine other 'offensive language' terms, which are similar to other terms like a curse word, bad words, or swear words. Other past works use different labels to identify toxic content. Some past works include the Crowdflower1 dataset [1–3].

Natural Language Processing (NLP) is a subfield of artificial intelligence, computer science, and linguistics that has a purpose to make computers/machines understand human language using natural language. Under the term NLP, there are Natural Language Understanding (NLU) which more specifically about the meaning or semantic, and Natural Language Generation (NLG). The tasks in NLP include communicating with untrained individuals and understanding their intent, more specifically NLP aims to understand the meaning despite common human error like mispronunciations or transposed letters or words. NLP as a field that directly works with computationally analyzing language, is in a unique position to develop methods to detect and

filter toxic content. In the tasks on natural language understanding, the term 'meaning' is a multi-faceted concept with semantic, pragmatic, cognitive, and social aspects. The basic idea of 'Meaning representation' is to know what words mean, how words interconnect, and the context of the sentence. Meaning representation techniques aim to teach computers/machines to understand the meaning of sentences like how humans understand the meaning of the sentences. From the past research on automatic profanity content detection tasks, mostly using a different approach to extract and represent the features to enable the classification algorithms to perform the task. Since feature representation is the task to encode the human natural language into some form that computers and machines can understand, a variety of NLP methods have been employed to represent the feature of the input sentences. An effective model for text analysis in NLP is Abstract Meaning Representation(AMR), which designs an acrylic directed graph from a content of a sentence. The elements of the AMR graph are related to the structure of the sentence. AMR has been applied in several domains such as event extraction, [22] text summarization, and generation [23, 24]. The web multimedia has to deal with a large amount of data that need to be processed and recognize specific patterns in our case toxic meaning or content that is offensive to the majority and alongside the techniques such as AMR, machine learning methods have provided effective solutions for data-driven approaches. The state of art model is Convolution Neural Networks (CNN) a deep learning method that operates directly on graph structure [25]. Integration of AMR with CNN makes an efficient tool for representing and recognizing toxic meaning from text content, providing a solution for application to protect users by some irresponsible other causing a serious problem that could affect security and also could destroy the mental health of the unprotected. Since the number of social media or social networking users has been increasing rapidly, the power of human moderators to filter abusive or profane content in social media could not handle all the problems and automatic toxic content detection and filtering system became highly demanded.

The remainder of this paper is organized as follows. Section 2 presents the background and related studies. Section 3 describes the anomaly detection methods and the characteristics of the data used for the experiments. The experimental results are presented and discussed in Sections 4 and 5. Section 6 provides the conclusion and future work.

2 Related Work

To predict the toxic content of the sentence, after preparing the text to work with a machine, the next step using classifiers like the machine learning approach to work with the detection. In machine learning approaches, it could be categorized into unsupervised, semi-supervised, and supervised approaches. While supervised learning depends on manual labeling of a large volume dataset, on the other side unsupervised learning does not rely on human labor and it's a domain-independent approach that is capable of handling a diversity of the content. Supervised learning effectively to capture small scale event but in the contrast, it needs a manual label of the data set, while unsupervised learning has limited power to handle small scale event, that's why Hua et al. [4] proposed to use semi-supervised learning to using labeled data in conjunction with unlabeled data. In supervised classification, Burnap and Williams [5] worked on a variety of supervised classifiers and resulted in a similar performance of each classifier but the accuracy could be changed depending on the setting of the features. Moreover, in the semi-supervised learning domain, Xiang et al. [6] claimed that their proposed approach could be a good alternative to tackle the cost problem of supervised approaches to detect toxic content. One of the works in unsupervised learning by Giftari et al. [7] proposed to utilize a bootstrapping approach to build the toxic content lexicon and their best results were obtained with combined semantic hate and their based features.

Deep learning methods perform at state of art level in the text mining and text classification task. In toxic content detection challenge and similar task with other terms like toxic comment detection or hate speech detection. In the task of online hate detection, many models of deep learning have been applied in the research architecture, like recurrent neural network (RNN) [8–10], convolutional neural network (CNN) [11], or a combination of both RNN and CNN. For example, works from Badjatiya et al. [12] applied deep neural networks to classify the hatefulness of tweets and stated that performance from CNN showed a better result compared to the baseline methods (n-gram, TF-IDF, BOW). Other work from Park and Fung [13] applied CNN methods to detect racist and sexist language. This work classifies 20K tweets and puts them on 3 kinds of CNN models (CharCNN, WordCNN, and HybridCNN) achieving the best performance from HybridCNN. In the challenge for detecting types of toxicity in comments, Georgeakopoulos et al. [14] proposed a deep learning approach involving CNN similar to Chu et al. [15] but with this work applying character level embedding also for detecting types of toxicity.

Other works also involve Long-Short Term Memory Networks (LSTM) [16] and RNN-LSTM with custom embedding [17].

3 Proposed Method

One of the aims in meaning representation is to build a representation that able to give a similar format of representation even if the sentence is paraphrased. The meaning representations are often based on underlying formalism or grammar. There's a variety of meaning representation formalism including typed lambda-calculus expressions, semantic role labeling, and also AMR. In this work, we utilize AMR as a meaning representation in integration with CNN's.

3.1 Toxic Content to Graph Representation

AMR is a tool that represents the whole sentence as graph structure, where the nodes are the "concepts" and the edges are the "relations". AMRs were introduced by Banarescu et al. [18] and are acyclic, directed graphs that represent the meaning of a sentence. AMR of the sentence could be represented by its textual PENMAN notation format [19] as shown in Figure 1 or a graph format shown in Figure 2. One of the purposes of AMR is to normalize and remove the effect of grammatical skewing, meaning there are some grammatical twists or distortions between semantics and grammar.

   AMR parsing is the annotation of natural language with linguistic structure in the graph structure. AMR represents the input sentence as a graph $G = \langle N, E \rangle$ where nodes and edges are given labels from sets L_N and L_E. The graph is constructed with 2 stages, the first stage identifies the concepts (nodes) from the words in input sentence $W = \langle w_1, \ldots, w_n \rangle$, where w_i a

(v2\want-01

: ARG_0 (v1/girl)

: ARG_1 (v3/go-02)

: ARG_4 (v4/library)

: ARG_0 (v1))

Figure 1   PENMAN format for the sentence (ex. The girl wants to go to the library.).
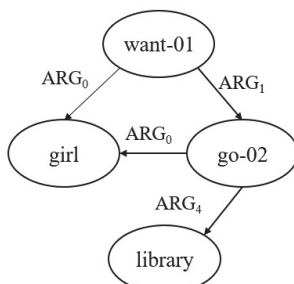
Figure 2   AMR for sentence (ex. The girl wants to go to the library.).

Table 1   AMR sentence meaning representation

Sentence	AMR Representation
The girl wanted to eat all candies.	(w / want-01
	: ARG_0 (g / girl)
The girl's wish was to eat all candies.	: ARG_1 (e /eat-01
	: ARG_0 g
	: ARG_1 (c / candy
Eating all candies was what children desired.	:mod (a / all))))

member of vocabulary *W*. Second stage is connecting the concepts by adding labeled edges capturing its relation with concepts, and select root in *G* corresponding to the focus of the sentence *w*. The principle of AMR is abstract away from the syntactic and morphological variation of the sentence. So, if we employed AMR to represent a sentence with different morphological, if their meaning is the same then the representation of their graph is going to be the same. Considering a simple example as given in Table 1, the paraphrase sentences, even they have a different form, but AMRs can abstract away from word to the concept or the meaning, the AMR representation of that three sentences will be the same.

3.2 AMR-CNN for Toxic Meaning Detection

This paper combines the AMR to capture the meaning representation from the input, which is the sentence that is annotated as a profane phrase or offensive phrase, and then uses CNN methods to do with the detection of whether the input is a profane phrase or not, demonstrated in Figure 3.

The first step is to create the AMRs graph for the Toxic content phrase data, we used the JAMR parser [20] to generate the Toxic meaning representation graphs. These parsers work in two parts algorithms, which first identifies

concepts using a Semi-Markov Model and then identifies the relation by searching the maximum spanning connected subgraph. After generating the AMR graph, in the second step, we feed the graph into a two-layer CNN and then fed it into a *softmax* classifier to classify the class. CNN is among the neural network methods that can calculate directly on a graph and induces embedding vectors of nodes based on the properties of their neighborhoods.

For a formal equation, graph $G = (P, E)$ where $(|P| =)$ represents the point (vertex) of the graph as are set of nodes and E are sets of edges. Usually, in Convolutional Network other works use the adjacency matrix A and its degree matrix D to represent graph G, where $D_{ij} = \sum_j A_i$. Where diagonal elements of A are set to 1 because of self-loops. And the formula for one convolutional layer is

$$H = \tilde{A}XW \tag{1}$$

Where $X \in R^{n \times m}$ is the number of nodes, m is the dimension of the feature, $\tilde{A} = D^{1\backslash 2}\overline{A}D^{-1/2}$ is the normalized symmetric adjacency matrix and $W \in R^{m \times h}$ is a weight of a matrix. The convolution layer function is

$$Convolution = \tilde{A}Relu(\tilde{A}XW_0)W_1 \tag{2}$$

Where W_0 and W_1 are weight matrices. For the second layer node (word) embedding pas into *softmax* classifier

$$Output = softmax(Convolution) \tag{3}$$

4 Experiments

AMR-CNN model performs for 2 kinds of datasets, the first dataset is Offensive Language Identification Dataset (OLID) [21] the second dataset is Twitter from Davidson [3] Twitter Sentiment Analysis Dataset. The OLID uses a hierarchical annotation schema to split the level of offensiveness into three levels. *Level_A* classifies the input data into two classes which are 'Offensive' labeled 'OFF' and 'Not Offensive' labeled by 'NOT'. *Level_B* categorized the type of offense, which is 'Targeted Insult' that contains insult/threat to an individual, a group, or other and also 'Untargeted' that sentence that contains insult but they are not targeted. The last level, *Level_C* more into categorizes the targets of insults/threats whether it's 'Individual', 'Group' or 'Other'. In the OLID, from the whole 13,240 tweets data, 4,400 data is labeled as a tweet that contains offensive words or offensive meaning, and 8,840 data labeled as neutral or tweets without offensive words or meaning (Figure 3).
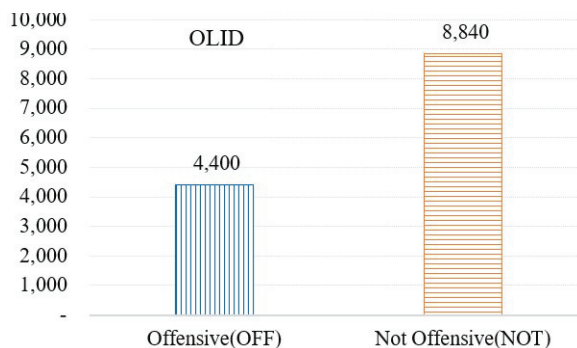
Figure 3   OLID distribution of toxic class (*Level_A*).

Besides using the OLID, this paper also tried the network with the dataset from Davidson et al. for hate and offensive tweets. This data came from the tweets that contained terms in the hate lexicon from Hatebase.org, and then manually annotated by CrowFlower(CF) workers. Since the start this work already defined a profane sentence as a sentence that includes offensive word or hate speech, that's why we change the label of the class into 3 classes. The 'Offensive' and 'Hate' data we consider as 'Toxic' and the tweets that neither contain hate or offensive words we put the label as 'Non-Toxic'. In the experiment, combine the data with Twitter Sentiment Analysis dataset. This kind of sentiment dataset is usually more used in the NLP domain to determine the emotional tone of the sentence or author's attitude toward something. This data is a twitter data with the label '0' for a positive and '1' for tweets that have a negative sentiment. As shown in the Figure 4, we choose to merge both datasets to create a balanced distribution of class. From the Figure 4(a) we could know that both datasets have an imbalance data distribution. The data from Davidson et al. have more Toxic, Non Toxic Davidson et al. have more Toxic tweet with 23,353 tweets data and only 1,430 data for Non-Toxic tweets. Meanwhile, the data from Twitter Sentiment Analysis have more data for the tweets that not have toxic meaning or profane word. With only 2,242 tweets that labelled as toxic and 29,720 data that considered as neutral or do not have toxic meaning. With merged both data, this experiment used both of dataset as our second dataset that also will be used to test the performance of the model if we used multiple kinds of dataset. The second datasets number is 56, 45 tweets dataset with 25,595 tweets are the tweet that have profanity meaning and 31,150 data are the tweet that neutral and not have a toxic words or profane meaning.
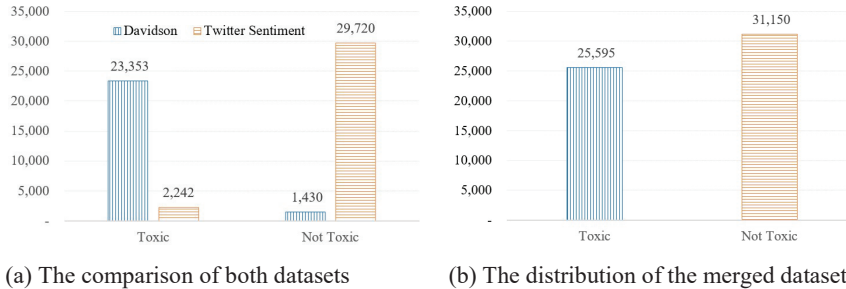
(a) The comparison of both datasets   (b) The distribution of the merged dataset

Figure 4   The distribution from Davidson-Twitter sentiment dataset.

Table 2   Hyperparameters used in AMR-CNN

Hyper-Paremeters	Values
Learning Rate	[0.5, 0.001, 0.002]
Dropout rate	[0.5, 0.2]

After we parse each input sentence into its graph representation, the model feed the graph meaning representation into the neural network. This work using CNN to learn the local features of the graph and find a toxic meaning of the sentence. Since AMR able to describe the correlation of each word and state what kind of correlation the word has, CNN able to detect the Toxic meaning from the text. And or the detailed information we trained this network using learning rate 0.02 and 200 epoch, with the number of hidden layers is 200. To optimizing the hyper-parameters of the model, we performed a fine-tuning of the hyper-parameters and 5-fold cross-validation of the model. Cross-validation is used to better understand the learning process and to estimate the learning variance since large variance tends to leads to biased model and overfitting problems. Using the sci-kit-learn library, we implement *Kfold*() class to measure the performance of the model for each fold. For fine-tuning the hyper-parameter, we also use a variant hyperparameter to improve the performance of the model. In this experiment, we performed tuning in learning rate and dropout rate as shown in Table 2.

5 Result and Discussion

In this experiment, we train 2 kinds of datasets to test the performance of the network in classification tasks. Based on our experiment before, we performed the model with the tuned hyper-parameters dropout rate 0,2 and also
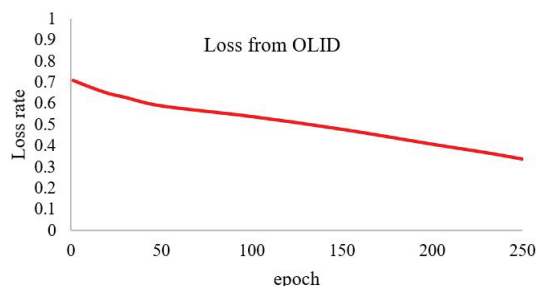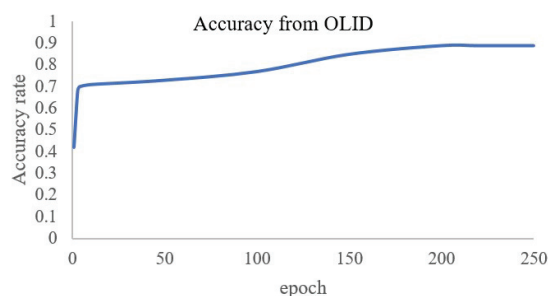
Figure 5 Loss rate for OLID.

Figure 6 Accuracy rate from OLID.

learning rate 0,001. The first dataset that we used is from OLID (Offensive Language Identification Dataset) and the second dataset is a merged dataset from Davidson et al. and Twitter Sentiment Analysis dataset. In the first experiment, we train the OLID dataset in our proposed network. We can see in Figures 5 and 6, the performance of our network in this dataset reaches the highest accuracy for 84% and the lowest loss score for 0.38 The second experiment with dataset 2, that is a merged dataset of Offensive tweet from Davidson et al. and data from Twitter Sentiment Analysis. The accuracy and loss result from the training for each epoch present in Figures 7 and 8. Based on the training result using dataset 2, our proposed method could achieve the highest performance in training with the highest accuracy 91%, and the lowest Loss score of 0.21.

Based on the experiment with both different datasets, our result showed that the performance of our network was slightly higher when trained on the merged dataset (Tables 3 and 4) but the score still showed that our approach is reliable to tackle the task on the toxic meaning detection. We can also say if our model shows promising result for both datasets that can be applied
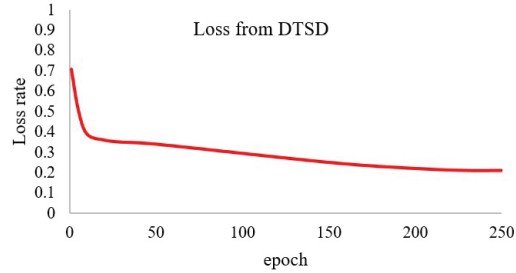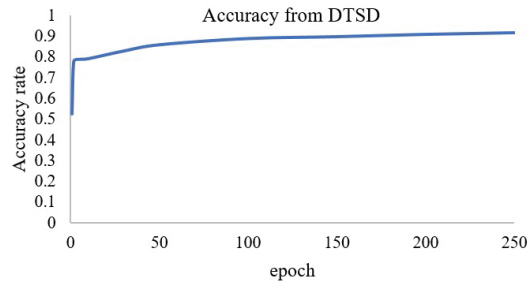
Figure 7 Loss rate for DTSD.

Figure 8 Accuracy rate from DTSD.

Table 3 Cross-Validation of model for OLID

Model	CV k = 1	CV k = 2	CV k = 3	CV k = 4	CV k = 5	Avg.
Loss	0.385	0.392	0.387	0.377	0.390	0.386
Accuracy	0.843	0.840	0.952	0.842	0.841	0.841

Table 4 Cross-Validation of model for DTSD

Model	CV k = 1	CV k = 2	CV k = 3	CV k = 4	CV k = 5	Avg.
Loss	0.213	0.211	0.215	0.210	0.215	0.212
Accuracy	0.912	0.911	0.918	0.913	0.911	0.913

to other datasets with a minor hyperparameter tuning or cross validation arrangement for estimate the learning variance of the data.

There are still some limitations in this research related to the data, comparison of the performance with other works and the development of the methods. First, we can say that our experiment still lacks data for training and testing the network. Since the domain of Toxic content detection that we choose is a broad domain and consider the data of social media content to grow rapidly, there's still many unexplored data that could help to increase

the performance of automated toxic content detection. Therefore, in the future work we will also focus on enriching the dataset for the classification and try to run the method on a few different datasets to test the network performance. Secondly, there's many other methods that work in this task. But our work still lacks in experiment to compare other robust methods with our proposed methods. Therefore, in the future research we will consider to also experiment the same dataset on a few other state-of-art methods to make a performance comparison with our works.

6 Conclusion

This paper proposed AMR-CNN as one method to represent the meaning of the text and build the graph of the input sentences based on its meaning.

Since the ability of AMR to handle semantic role, can identify entity types and also the coreference of one word with other words, thus make this method have a promising performance to tackle the task in online abuse/bad word or toxic meaning detection domain. The other advantages of AMR are its ability to handle modality, polarity and even connect every concept with the Knowledge of the world like wikification, that make AMR possible to extracted the meaning of concept based on its contents.

Our proposed new approach model showed a good performance on both datasets with the precision score for OLID dataset reach 85% and for Davidson-Tweeter dataset is 91%. This showed that our model can perform well with variety datasets and show similar results, even though we consider that we lack of more big and reliable datasets to test the performance, but in the future, we planning to train and test the model with bigger dataset. And also, we consider that using the Abstract Meaning Representation combined with CNN, the network could understand better the meaning of the sentence and this model able to detect the toxic meaning of the phrases and give a proper label 'Profane' and' 'Not Profane' for the input sentences.

References

[1] Gaydhani, A., Doma, V., Kendre, S. and Bhagwat, L. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. arXiv preprint arXiv:1809.08651, 2018.

[2] Watanabe, H., Bouazizi, M. and Ohtsuki, T. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, pp. 13825–13835, 2018.

[3] Davidson, T., Warmsley, D., Macy, M.W. and Weber, I., Automated hate speech detection and the problem of offensive language. CoRR, abs/1703.04009. URL: http://arxiv.org/abs/1703.04009, 2017.

[4] Hua, T., Chen, F., Zhao, L., Lu, C.T. and Ramakrishnan, N., STED: semi-supervised targeted-interest event detectionin in twitter. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1466–1469), August 2013.

[5] Burnap, P. and Williams, M.L., Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data science, 5(1), p. 11, 2016.

[6] Xiang, G., Fan, B., Wang, L., Hong, J. and Rose, C., October. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 1980–1984, 2012.

[7] Gitari, N.D., Zuping, Z., Damien, H. and Long, J., A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering, 10(4), pp. 215–230, 2015.

[8] Pavlopoulos, J., Malakasiotis, P. and Androutsopoulos, I., Deeper attention to abusive user content moderation. In Proceedings of the 2017 conference on empirical methods in natural language processing, pp. 1125–1135, September 2017.

[9] Pitsilis, G.K., Ramampiaro, H. and Langseth, H., Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433, 2018.

[10] Gao, L. and Huang, R. Detecting online hate speech using context aware models. arXiv preprint arXiv:1710.07395, 2017.

[11] Park, J.H. and Fung, P., One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206, 2017.

[12] Badjatiya, P., Gupta, S., Gupta, M. and Varma, V., April. Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759–760), 2017.

[13] Park, J.H. and Fung, P., One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206, 2017.

[14] Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G. and Plagianakos, V.P. Convolutional neural networks for toxic comment classification. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence pp. 1–6, July, 2018.

[15] Khieu, K. and Narwal, N., Detecting and classifying toxic comments. Web: https://web.stanford.edu/class/archive/cs/cs224n/cs224n,1184.

[16] Chu, T., Jue, K. and Wang, M., 2016. Comment abuse classification with deep learning. Von https://web.stanford.edu/class/cs224n/reports/2762092.pdf abgerufen.

[17] Kohli, M., Kuehler, E. and Palowitch, J., Paying attention to toxic comments online. Web: https://web.stanford.edu/class/archive/cs/cs224n/cs224n, 1184.

[18] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N., Abstract meaning representation for sembanking. In Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, pp. 178–186, 2013.

[19] Matthiessen, C.M.I.M. and BATEMAN, J., Systemic-Functional Linguistics in Language Generation: Penman, 1991.

[20] Flanigan, J., Thomson, S., Carbonell, J.G., Dyer, C. and Smith, N.A., June. A discriminative graph-based parser for the abstract meaning representation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1426–1436), 2014.

[21] Kipf, T.N. and Welling, M., Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[22] Rao, S., Marcu, D., Knight, K., Daumé III, H., Biomedical event extraction using abstract meaning representation. BioNLP 2017 pp. 126–135, 2017.

[23] Dohare, S., Karnick, H., Text summarization using abstract meaning representation. arXiv preprint arXiv:1706.01678, 2017.

[24] Song, L., Zhang, Y., Peng, X., Wang, Z., Gildea, D., Amr-to-text generation as a traveling salesman problem. In: EMNLP 2016.

[25] Tayal, Kshitij, Rao Nikhil, Saurabh Agarwal, and Karthik Subbian. "Short text classification using graph convolutional network." In NIPS workshop on Graph Representation Learning. 2019.

[26] Guo, Beibei, Yu Xiao, Chiping Zhang, and Yong Zhao. "Graph theory-based adaptive intermittent synchronization for stochastic delayed complex networks with semi-Markov jump." Applied Mathematics and Computation 366: 124739, 2020.

Biographies

Ermal Elbasani received his bachelor's and master's degree in Electronic Engineering in Electronic Engineering from Polytechnic University of Tirana, in Albania, 2011 and 2013 respectively. Currently attending the philosophy of doctorate in Computer and Electronics Engineering from Sunmoon University in 2021 South Korea. His research interest includes healthcare and wellness monitoring, graph and deep learning, and biological sequential data analysis.

Jeong-Dong Kim received the bachelor's degree in computer engineering from Sun Moon University in 2005. He received his M.S. and Ph.D. degrees in Computer Science from Korea University at Korea in 2008 and 2012, respectively. He is an associate professor in the department of computer science and engineering, Sun Moon University, Asan, Korea. His research interests include bigdata analysis based on deep learning, Healthcare, software & data engineering, and bioinformatics.