Computers, Materials & Continua
DOI:10.32604/cmc.2022.019521
Article

Tech Science Press

Integrating Deep Learning and Machine Translation for Understanding Unrefined Languages

HongGeun Ji[1,2], Soyoung Oh[1], Jina Kim[3], Seong Choi[1,2] and Eunil Park[1,2,*]

[1]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, 03063, Korea
[2]Raon Data, Seoul, 03073, Korea
[3]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, 55455, MN, USA
[*]Corresponding Author: Eunil Park. Email: eunilpark@skku.edu
Received: 16 April 2021; Accepted: 23 May 2021

Abstract: In the field of natural language processing (NLP), the advancement of neural machine translation has paved the way for cross-lingual research. Yet, most studies in NLP have evaluated the proposed language models on well-refined datasets. We investigate whether a machine translation approach is suitable for multilingual analysis of unrefined datasets, particularly, chat messages in *Twitch*. In order to address it, we collected the dataset, which included 7,066,854 and 3,365,569 chat messages from English and Korean streams, respectively. We employed several machine learning classifiers and neural networks with two different types of embedding: word-sequence embedding and the final layer of a pre-trained language model. The results of the employed models indicate that the accuracy difference between English, and English to Korean was relatively high, ranging from 3% to 12%. For Korean data (Korean, and Korean to English), it ranged from 0% to 2%. Therefore, the results imply that translation from a low-resource language (e.g., Korean) into a high-resource language (e.g., English) shows higher performance, in contrast to vice versa. Several implications and limitations of the presented results are also discussed. For instance, we suggest the feasibility of translation from resource-poor languages for using the tools of resource-rich languages in further analysis.

Keywords: Twitch; multilingual; machine translation; machine learning

1 Introduction

In linguistic and computer science research, one of the most challenging research topics is to develop systems for high-quality translation and multi-linguistic processing. Thus, many scholars have attempted to propose state-of-the-art translation services and systems to improve the results of translation.

In addition to some translation research, natural language processing (NLP) technologies have been rapidly improving. Because of international collaboration in research and development, the majority of NLP research aims to investigate resource-rich languages that are widely used in global society. Hence, NLP research is more focused on English rather than other languages [1].

Because of insufficient research and development in under-resourced languages, several scholars attempted to apply English NLP technologies to understand and investigate other languages [2–4]. For instance, Patel and colleagues used machine translation for sentiment analysis of movie reviews and then compared the results of the translation approach with native Hindustani NLP [3].

To employ NLP technologies for low-resource languages, a two-step approach can be used. First, well-constructed translation methodologies should be employed to translate the contents in low-resource language into high-resource language. Second, the translated content is represented as vectors by various word embedding algorithms. Therefore, improved translation methodologies can enhance the results of NLP technologies in other languages.

Within this trend, several studies have attempted to develop state-of-the-art translation techniques. One of the remarkable improvements is Google's neural machine translation system (GNMT) [5]. Compared with the phrase-based production system, GNMT reduced errors by 40% when using human evaluation [5]. Using rapidly improving machine translation techniques, Kocich et al. [6] successfully categorized the sentiments in an online social network dataset using an English sentiment library.

However, most recent studies have been conducted for well-refined content. With unrefined content, there can be some hindrances, for example, when chat messages are processed and explored. Communication in chat messages (known as "netspeak") has unique language characteristics in spelling and grammar, including the use of acronyms and abbreviations [7]. Moreover, because a lot of me-media channels, which are interactive media platforms for viewers and streamers, are globally introduced, a huge amount of chat messages and content in various languages is produced. Thus, we aim to investigate whether machine translation can be applicable for multilingual analysis of unrefined content. To address it, unrefined chat messages of both English and Korean streamers in *Twitch* [8], a widely used online streaming service, are collected for analysis.

2 Related Work

Machine learning and deep learning approaches have become mainstream in NLP research. Also, the cross-lingual approaches in NLP have also been extensively explored and achieved considerable results. Thanks to these approaches, diverse tasks can be performed for limited-resource languages (e.g., Spanish and Hindi) and not only for languages with rich resources (e.g., English) [2–4].

Among these tasks, a text categorization task using bilingual-corpus datasets was represented as the cost-effective methodology resulting in comparable accuracy [9].

Moreover, with the advancement of neural machine translation (NMT) beyond the conventional translation models, several cross-lingual approaches applied this technique [3,10,11]. Patel and colleagues showed comparable accuracy of sentiment classification by translating low-resource languages into English (as a high-resource language) [3]. Furthermore, performance of NMT models can be enhanced by focusing on topic-level attention during the process of translation [11].

Recent cross-lingual approaches have been improved by a pre-trained language model based on neural networks [12,13]. The pre-trained word-embedding techniques, such as *Skip-Gram* [14], and *GloVe* [15], capture different properties of the words. Moreover, in the case of learning the contextual meaning and structure of the syntax, several state-of-the-art pre-trained language models were introduced, including *CoVe* [16], *ELMo* [17], and *BERT* [18]. The transformer encoder enabled these models to handle the complex representation of contextual semantics. All the representative pre-trained language models were trained on refined large text corpora (such as *Wikipedia* in English, as a commonly used language).

By favor of these properties, several studies have applied pre-trained language models on large-scale data [19]. However, the majority of prior studies have been conducted using relatively well-refined datasets (*e.g., Wikipedia*, social networking sites, microblogs, or user reviews) [20]. As pre-trained language models were implemented to read the whole sequence of words and showed remarkable improvements in NLP tasks, we attempt to examine whether applying advanced pre-trained language models to the unrefined content to learn the entire context of words can be recommended in the field of machine translation.

Thus, we investigates whether machine translation approaches are applicable to the classification task of unrefined data compared with the evaluation of the original language.

3 Method

To validate our approach on unrefined data, we used chat messages in a representative live-streaming platform, *Twitch*. In *Twitch*, there are active interactions and communications between viewers and streamers [21]. We selected a straightforward binary classification task for chat messages: predicting whether a specific viewer in *Twitch* is a subscriber who pays for live game-streaming services.

3.1 Data Acquisition and Preprocessing

We collected the 50 most-followed English and Korean streamers from *TwitchMetrics* [22]. Specifically, we collected all chat messages from five recent streams of each streamer using an open-source crawler, *Twitch-Chat-Downloader* [23]. The dataset included 7,066,854 and 3,365,569 chat messages from English and Korean streams, respectively.

Fig. 1 shows the whole data preprocessing procedures. During the preprocessing, we first excluded the chat messages with URLs, user tags annotated with @, and emoticons. In addition, we eliminated the notifications which indicated who subscribed to the streamers. We did not apply stemming or lemmatization to prevent the information loss in short messages. In addition, we removed the chat messages less than five words which cannot convey the states of the viewers. Subsequently, we used *Google Translation API* to translate English chat messages to Korean and vice versa. The chat messages that were not translated properly were removed. Finally, we used 1,321,445 English (*EN*) and English-to-Korean (*EN2KO*) and 109,419 Korean (*KO*) and Korean-to-English (*KO2EN*) chat messages. Moreover, to classify whether a specific viewer is a subscriber, we identified the subscription badges of viewers, which were displayed in messages.

3.2 Embedding

We employed two techniques for embedding: word-sequence and sentence embedding.

3.2.1 Word-Sequence Embedding

We employed two tokenization techniques according to the target language. In the case of English (*EN* and *KO2EN*), we employed the *Tokenizer* of Python library *Keras* [24]. We tokenized the Korean chat messages (*KO* and *EN2KO*) using the *Open Korea Text* of Korean NLP library, *KoNLPy* [25]. After examining the tokenization techniques, we embedded the tokens in 256-dimensional vectors.

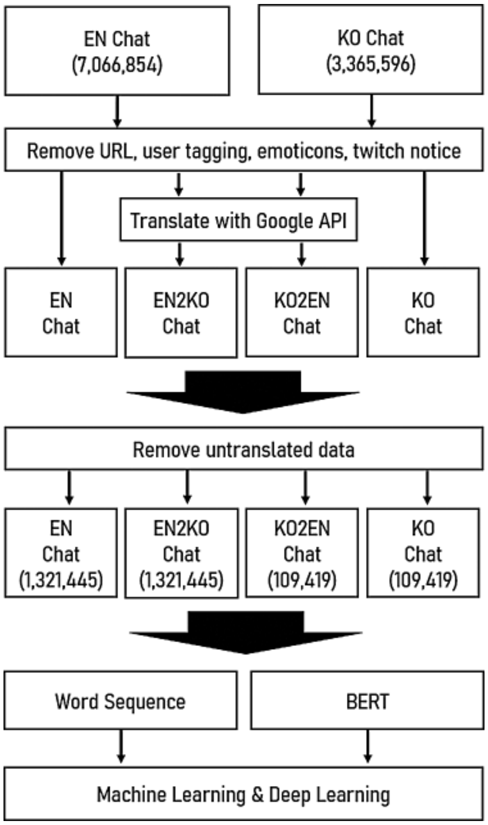

Figure 1: Workflow procedures

3.2.2 Sentence Embedding: BERT

We used the embedding vector extracted from the last layer of a widely-used pre-trained language model, *BERT*, which reflects the context of the sentences. Among the wide range of *BERT* model sizes, we chose *BERT-base-uncased* model to the English chat messages (*EN* and *KO2EN*) [26]. For Korean chat messages (*KO* and *EN2KO*), we applied *KoBERT* [27]. With the employment of *BERT* model, we used the hidden states of first token of input sequence (called [CLS] token) in the last layer of *BERT*, a 768-dimensional vector, as one of the embedding techniques.

3.2.3 Classification Models

We applied both machine learning classifiers and deep neural networks: *Logistic Regression*, *Naïve Bayes*, *Random Forest*, *XGBoost*, *Multilayer Perceptron* (MLP), *STACKED-LSTM*, and *CONV-LSTM*. The *STACKED-LSTM* model consists of two long short-term memory (LSTM) layers with 128 recurrent neurons and a fully connected layer. The *CONV-LSTM* has one-dimensional convolutional layer with 64 filters, max-pooling layer, *LSTM* layer with 128 recurrent neurons, and a fully connected layer. The output of the fully connected layer is passed through the softmax activation function.

We divided the collected chat messages into training (80%) and testing (20%) sets. Therefore, the training sets included 87,535 (KO) and 1,057,156 (EN) chat messages. The number of chat messages in the test dataset was 21,884 (KO) and 264,289 (EN). We applied the *synthetic minority over-sampling technique* (*SMOTE*) for the machine learning classifiers [28]; moreover, we adjusted class weights in the cross-entropy function of the deep neural networks to handle class imbalance (Fig. 2) [29,30].

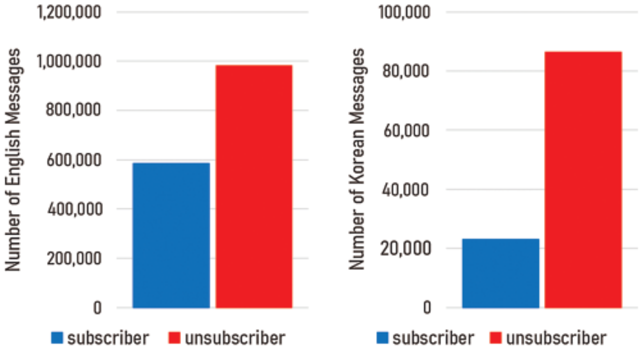

Figure 2: Class distribution for English and Korean datasets

4 Results

4.1 Classification Models with English Data

The accuracy of the classifiers using English data (*EN* and *EN2KO*) is summarized in Tab. 1. Among classifiers using untranslated English (*EN*), *Random Forest* with word-sequence embedding showed the highest performance, with the accuracy of 89.35%. The *STACKED-LSTM* model with word-sequence embedding showed the highest accuracy (82.03%) among the models with English-to-Korean input data (*EN2KO*).

The average accuracy of the models with word-sequence embedding was slightly higher with untranslated data (*EN*: 78.79%) compared with translated data (*EN2KO*: 73.30%). Similarly, in the case of *BERT* embedding, the models with untranslated data (*EN*: 80.17%) outperformed the models with translated data (*EN2KO*: 78.13%).

In the case of the *Naïve Bayes* classifier, performance was better with *BERT* embedding rather than word-sequence embedding, which was approximately 25% (*EN*) and 27% (*EN2KO*), respectively.

As shown on the left side of Fig. 3, the accuracy of classifiers with the word-sequence embedding of the untranslated data (*EN*) was higher than for *BERT* embedding (*Random Forest*, *XGBoost*, *CONV-LSTM*, *and STACKED-LSTM*).

4.2 Classification Models with Korean Data

Tab. 2 represents the accuracy of classifiers using Korean data as input (*KO* and *KO2EN*). *Random Forest* with *BERT* embedding showed the highest performance for both translated and untranslated data (*KO*: 86.92%, *KO2EN*: 86.70%). The average accuracy of classifiers was similar for untranslated and translated input data (*KO*: 73.74%, *KO2EN*: 72.11%). This aligns with the results of *BERT* embedding (*KO*: 80.30%, *KO2EN*: 79.33%).

Table 1: Classification metrics with English data

			English (EN)				English to Korean (EN2KO)			
Embedding	Model	Class	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Word Sequence	Logistic regression	Unsubscriber	76.12	93.16	83.79	76.89	64.92	94.77	77.05	64.66
		Subscriber	79.67	47.81	59.76		62.00	14.28	23.21	
	Random forest	Unsubscriber	86.75	98.42	92.22	89.35	81.15	92.80	86.58	81.57
		Subscriber	96.30	73.15	83.14		82.72	61.49	70.54	
	Naïve Bayes	Unsubscriber	82.30	6.76	12.49	39.29	82.08	10.50	18.63	41.15
		Subscriber	36.90	97.40	53.52		37.49	95.90	53.91	
	XGBoost	Unsubscriber	85.77	98.91	91.87	88.78	79.71	90.90	84.94	79.33
		Subscriber	97.32	70.69	81.90		78.31	58.68	67.09	
	MLP	Unsubscriber	83.01	87.55	85.22	80.53	77.87	94.03	85.19	79.53
		Subscriber	75.35	67.98	71.48		84.69	55.27	66.89	
	STACKED-LSTM	Unsubscriber	86.46	96.89	91.38	88.28	81.66	92.81	86.88	**82.03**
		Subscriber	92.91	72.91	81.70		83.02	62.77	71.49	
	CONV-LSTM	Unsubscriber	85.95	97.99	91.58	**88.44**	81.38	86.99	84.09	78.90
		Subscriber	95.22	71.38	81.59		73.50	64.45	68.68	
BERT	Logistic regression	Unsubscriber	82.10	95.43	88.27	83.73	79.84	91.01	85.06	79.51
		Subscriber	88.51	62.84	73.50		78.61	58.96	67.38	
	Random forest	Unsubscriber	81.50	97.90	88.95	84.41	80.55	94.50	86.97	81.85
		Subscriber	94.15	60.32	73.53		85.79	59.25	70.10	
	Naïve Bayes	Unsubscriber	80.92	57.96	67.55	64.29	79.25	69.88	74.27	68.96
		Subscriber	50.17	75.59	60.31		55.58	67.32	60.89	
	XGBoost	Unsubscriber	80.63	95.97	87.64	82.64	79.02	92.56	85.26	79.48
		Subscriber	89.11	58.83	70.88		80.85	56.12	66.26	
	MLP	Unsubscriber	82.24	99.36	89.99	85.83	79.19	95.22	86.47	80.89
		Subscriber	98.17	61.66	75.75		86.62	55.30	67.50	

In addition, the accuracy of *Naïve Bayes* was much higher with *BERT* embedding (*KO*: 76.42%, *KO2EN*: 79.95%) rather than word-sequence embedding (*KO*: 25.32%, *KO2EN*: 26.31%). The right side of Fig. 3 shows the accuracy of the classifiers trained on Korean data (*KO*, *KO2EN*). Overall, the classifiers with relatively high accuracy had different embedding methods.

5 Discussion

We aimed to validate whether machine-translated datasets are applicable in the NLP tasks. We conducted binary classification with unrefined data (chat messages in live-streaming platform, *Twitch*) by using several machine learning classifiers and neural networks. Moreover, we employed two different types of embedding: word-sequence embedding and the output layer of *BERT*. We chose both English (resource-rich) and Korean (resource-poor) languages for the validation and named the datasets as follows: *EN*, *KO*, *EN2KO*, and *KO2EN*.

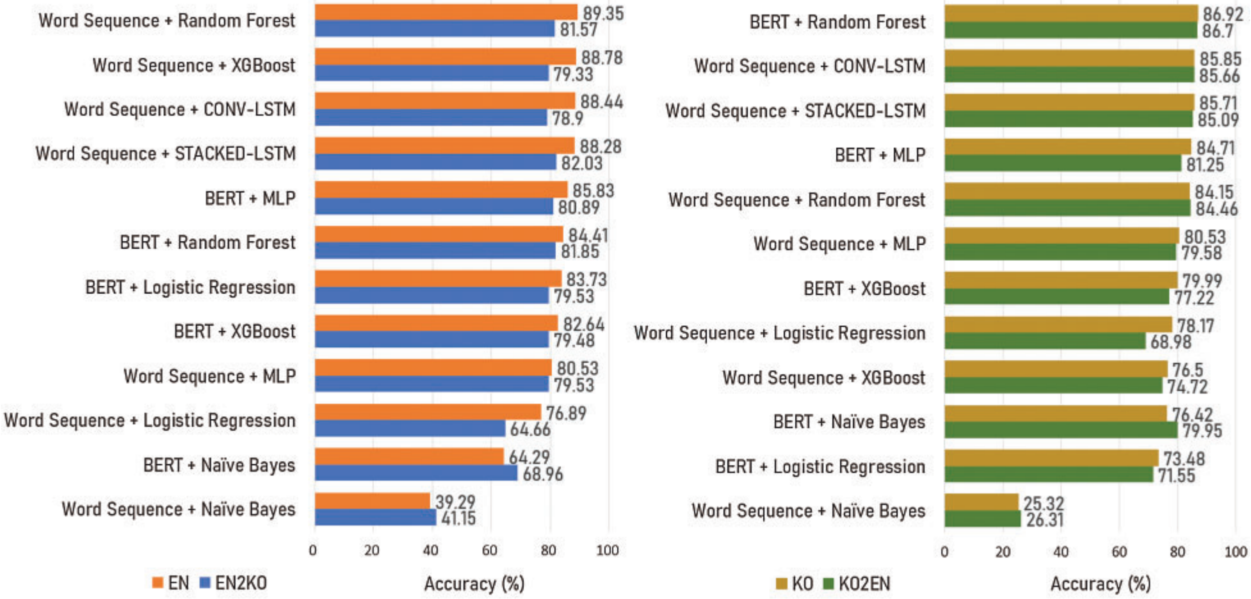

Figure 3: Classification accuracy for English data (EN, EN2KO) and Korean data (KO, KO2EN)

According to our results, the accuracy difference between *EN* and *EN2KO* was relatively high, ranging from 3% to 12%. For Korean data (*KO* and *KO2EN*), it ranged from 0% to 2%. Therefore, the results imply that translation from a low-resource language (e.g., Korean) into a high-resource language (e.g., English) shows higher performance, in contrast to vice versa.

Among the classifiers showing high accuracy for English (*EN* and *EN2KO*), the word-sequence embedding was highly employed. Meanwhile, in Korean (*KO* and *KO2EN*), there are no significant differences in dominance between word-sequence and *BERT* embedding. This shows that contextual approaches of *BERT* to unrefined data does not effectively impact the analysis.

In the case of classifiers resulting in low accuracy, *Naïve Bayes* in the current study, *BERT* embedding showed much higher accuracy compared to the word-sequence embedding in the multilingual analysis of unrefined content.

Overall, the evaluation of all classifiers implies that using machine translation from resource-poor (e.g., Korean) to resource-rich (e.g., English) language for input data (KO2EN) does not significantly affect the performance. This would suggest the feasibility of translation from resource-poor languages for using the tools of resource-rich languages in further analysis.

Although we investigated the efficacy of machine translation from a low-resource language to a high-resource language, several limitations must be considered. First, our evaluation of the task was limited to English and Korean. We may further investigate whether our approach produces comparable results in other languages. Also, using a highly improved classifier may be considered due to the rapid advancement in the field of machine learning. Therefore, these limitations can be addressed in future work.

Table 2: Classification metrics with Korean data

Embedding	Model	Class	Korean (KO)				Korean to English (KO2EN)			
			Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Word Sequence	Logistic regression	Unsubscriber	79.91	96.51	87.43	78.17	78.78	83.02	80.81	68.98
		Subscriber	44.90	10.48	16.99		21.51	17.16	19.09	
	Random forest	Unsubscriber	87.43	93.25	90.25	84.15	87.35	93.83	90.47	84.46
		Subscriber	67.01	50.55	57.63		68.68	49.87	57.78	
	Naïve Bayes	Unsubscriber	80.71	6.68	12.34	25.32	87.61	7.39	13.63	26.31
		Subscriber	21.46	94.10	34.95		21.95	96.14	35.74	
	XGBoost	Unsubscriber	86.42	83.20	84.78	76.50	86.64	80.24	83.32	74.72
		Subscriber	45.52	51.77	48.45		42.71	54.37	47.84	
	MLP	Unsubscriber	82.49	95.52	88.53	80.53	82.26	94.41	87.92	79.58
		Subscriber	60.42	25.22	35.59		54.67	24.86	34.18	
	STACKED-LSTM	Unsubscriber	85.50	98.54	91.56	85.71	85.20	98.08	91.19	85.09
		Subscriber	87.73	38.34	53.36		84.01	37.16	51.53	
	CONV-LSTM	Unsubscriber	85.41	98.90	91.66	85.85	84.69	99.82	91.63	85.66
		Subscriber	90.33	37.67	53.17		98.05	33.45	49.88	
BERT	Logistic regression	Unsubscriber	86.52	78.52	82.33	73.48	85.95	76.30	80.84	71.55
		Subscriber	40.91	54.86	46.87		38.18	54.00	44.73	
	Random forest	Unsubscriber	86.62	98.60	92.22	**86.92**	86.46	98.53	92.10	**86.70**
		Subscriber	89.49	43.82	58.84		88.82	43.07	58.01	
	Naïve Bayes	Unsubscriber	84.88	85.21	85.04	76.42	84.26	91.62	87.79	79.95
		Subscriber	44.64	43.99	44.31		54.41	36.88	43.96	
	XGBoost	Unsubscriber	85.43	89.90	87.61	79.99	85.15	86.06	85.60	77.22
		Subscriber	53.83	43.44	48.05		46.46	44.64	45.53	
	MLP	Unsubscriber	84.85	98.08	90.98	84.71	85.57	91.61	88.49	81.25
		Subscriber	83.34	35.38	49.67		58.14	42.99	49.43	

Acknowledgement: Prof. Eunil Park thanks to TwitchMetrics, and ICAN Program, IITP grant funded by the Korea government (MSIT) (No. IITP-2020-0-01816).

Funding Statement: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00358, AI·Big data based Cyber Security Orchestration and Automated Response Technology Development).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Al-Rfou, B. Perozzi and S. Skiena, "Polyglot: Distributed word representations for multilingual NLP," in *Proc. of the Seventeenth Conf. on Computational Natural Language Learning*, Sofia, Bulgaria, pp. 183–192, 2013.

[2] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis," in *Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing*, Honolulu, HI, USA, pp. 553–561, 2008.

[3] M. A. Hassonah, R. Al-Sayyed, A. Rodan, A. Z. Ala'M, I. Aljarah *et al.,* "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter," *Knowledge-Based Systems*, vol. 192, no. 1, pp. 105353, 2020.

[4] R. Xu, Y. Yang, H. Liu and A. His, "Cross-lingual text classification via model translation with limited dictionaries," in *Proc. of the 25th ACM Int. on Conf. on Information and Knowledge Management*, Indianapolis, IN, USA, pp. 95–104, 2016.

[5] Y. Wu, M. Schuster, Z. Chen, Q. V. Le and M. Norouzi, "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv, 2016. [Online]. Available: https://arxiv.org/abs/1609.08144.

[6] D. Kocich, "Multilingual sentiment mapping using Twitter, open source tools, and dictionary based machine translation approach," in *Proc. of GIS Ostrava*, Cham, Switzerland, pp. 223–238, 2017.

[7] M. Johnová, *The language of chat, Philologica.net*. Opava, Czech Republic: The Vilém Mathesius Society, 2004. [Online]. Available: http://philologica.net/studia/20040113000003.html.

[8] Twitch, [Online]. Available: https://www.twitch.tv/.

[9] N. Bel, C. H. Koster and M. Villegas, "Cross-lingual text categorization," in *Proc. of the Int. Conf. on Theory and Practice of Digital Libraries*, Berlin, Heidelberg, pp. 126–139, 2003.

[10] Z. Wu, H. Hou, Z. Guo, X. Wang and S. Sun, "Mongolian-Chinese unsupervised neural machine translation with lexical feature," in *Proc. of the China National Conf. on Chinese Computational Linguistics*, Cham, Switzerland, pp. 334–345, 2019.

[11] X. Wei, Y. Hu, L. Xing, Y. Wang and L. Gao, "Translating with bilingual topic knowledge for neural machine translation," in *Proc. of the AAAI Conf. on Artificial Intelligence*, Palo Alto, CA, USA, pp. 7257–7264, 2019.

[12] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness-of BERT," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*, Hong Kong, China, pp. 833–844, 2019.

[13] T. Pires, E. Schlinger and D. Garrette, "How multilingual is multilingual BERT?," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 4996–5001, 2019.

[14] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," arXiv, 2013. [Online]. Available: https://arxiv.org/abs/1301.3781.

[15] J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1532–1543, 2014.

[16] B. McCann, J. Bradbury, C. Xiong and R. Socher, "Learned in translation: Contextualized word vectors," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, New York, NY, USA, pp. 6297–6308, 2017.

[17] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark *et al.,* "Deep contextualized word representations," in *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, USA, pp. 2227–2237, 2018.

[18] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz *et al.,* "Fast and robust neural network joint models for statistical machine translation," in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, pp. 1370–1380, 2014.

[19] S. Ruder, M. E. Peters, S. Swayamdipta and T. Wolf, "Transfer learning in natural language processing," in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Tutorials*, Minneapolis, Minnesota, USA, pp. 15–18, 2019.

[20] M. H. Amirhosseini, H. B. Kazemian, K. Ouazzane and C. Chandler, "Natural language processing approach to NLP meta model automation," in *Proc. of 2018 Int. Joint Conf. on Neural Networks*, Rio de Janeiro, Brazil, pp. 1–8, 2018.

[21] Z. Hilvert-Bruce, J. T. Neill, M. Sjöblom and J. Hamari, "Social motivations of live-streaming viewer engagement on Twitch," *Computers in Human Behavior*, vol. 84, no. 3, pp. 58–67, 2018.

[22] TwitchMetrics, [Online]. Available: https://www.twitchmetrics.net/.

[23] Twitch-Chat-Downloader, [Online]. Available: https://github.com/PetterKraabol/Twitch-Chat-Downloader.

[24] Keras, [Online]. Available: https://keras.io/.

[25] KoNLPy, [Online]. Available: https://konlpy-ko.readthedocs.io/ko/v0.4.3/.

[26] BERT, [Online]. Available: https://github.com/google-research/bert.

[27] KoBERT-NER, [Online]. Available: https://github.com/monologg/KoBERT-NER.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[29] J. Kim, H. Ji, S. Oh, S. Hwang, E. Park *et al.,* "A deep hybrid learning model for customer repurchase behavior," *Journal of Retailing and Consumer Services*, vol. 59, pp. 102381, 2021.

[30] S. Hwang, J. Kim, E. Park and J. S., "Kwon Who will be your next customer: A machine learning approach to customer return visits in airline services," *Journal of Business Research*, vol. 121, pp. 121–126, 2020.