INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

IJIE

Vol. 16 No. 7 (2024) 12-25
https://publisher.uthm.edu.my/ojs/index.php/ijie

Demonstration of Efficacy of Exploiting ChatGPT Data to the Transformers-Based Models by Performing Bangla Intent Analysis

Al-Mahmud[1]*, Kazutaka Shimada[1]

[1] *Dept. of Artificial Intelligence,*
Kyushu Institute of Technology, 680-4 Kawazu Iizuka Fukuoka, 850-8502, JAPAN

*Corresponding Author: mahmud.al645@mail.kyutech.jp
DOI: https://doi.org/10.30880/ijie.2024.16.07.002

Article Info

Received: 24 July 2024
Accepted: 5 October 2024
Available online: 27 November 2024

Keywords

Intent analysis, conventional machine-learning models, transformers-based models, combined data technique, semi-supervised learning approach, Stepwise learning approach

Abstract

With the expanding mode of online opinion sharing, an automatic approach to intent analysis is necessary and useful in the practical scenario. Intent analysis inspects persons' and entities' viewpoints from online user-created texts. Conventional sentiment analysis deals with two classes: positive and negative. In this study, to extend the conventional sentiment analysis task, intent analysis deals with more important classes to obtain deeper insights. Accordingly, this study deals with five classes: pessimism, optimism, suggestion, sarcastic, and miscellaneous. Intent analysis with machine learning essentially needs a massive amount of data to generate a robust model. However, manually accumulating the training data is expensive, particularly in less dominant languages like Bangla. Hence, to obtain sufficient training data, this study generates, collects, and pre-processs Bangla restaurant data for the task by OpenAI ChatGPT API through prompt and data augmentation. These data are called "source data". As no user-generated Bangla data is available in the literature, this study prepares and validates a new Bangla intent analysis dataset by collecting user-generated real data. These data are referred to as "target data". Source data is utilized to assist the target task (i.e., main task) performed on the target data. By utilizing both source and target data, three approaches are proposed: combined data approach, semi-supervised learning, and stepwise learning. Experimental results demonstrated that the proposed semi-supervised learning with transformers-based models is effective in improving the performance of the target data by exploiting ChatGPT-generated source data. The best *F1* score of the proposed semi-supervised learning is 0.74, while that of the baseline is 0.72. Additionally, we proposed some feature concatenation methods. In this case, the highest *F1* score is 0.75.

1. Introduction

As the Internet's role in daily communication expands, so does the practice of sharing opinions on online platforms. Subsequently, the proliferation of online opinion content extends to remarkable magnitudes [1]. Sentiment analysis recognizes impressions that reflect the persons' viewpoints/sentiments on textual data, such as positive or negative toward a particular product, topic, or issue [2].

Consider the below review examples in the restaurant domain;

e1) Sushi is good. Here, the author manifested a positive sentiment towards sushi or food.
e2) Service is unsatisfactory. In this case, the author manifested a negative sentiment towards service or waiters.

Conventional sentiment analysis categorizes only positive and negative labels. However, deeper and more concrete information cannot be obtained from this simple binary classification. Moreover, a review can contain sarcastic information which may affect the classification of positive/negative class. It is possible to obtain deeper and more concrete information by increasing the number of important classes and defining them more specifically. This motivates the introduction of intent analysis. In this study, a dataset is regarded in the restaurant domain in the Bangla language, and there are five classes for the task. For non-native Bangla readers, examples are provided in English. Note that the examples of real Bangla sentences that are prepared are shown in Section 2.1.4. Below is the definition of the five classes in the intent analysis task:

1. Pessimism.
   The pessimistic attitude of a customer about food, service, price, ambiance, etc.
   e3) Food is terrible.
2. Optimism.
   The optimistic attitude of a customer about food, service, price, ambiance, etc.
   e4) Service is great.
3. Suggestion.
   The suggestion is given by the customer to the restaurant authority about food, service, price, ambiance, etc.
   e5) You should increase the number of items.
4. Sarcastic.
   The sarcastic reviews of a customer about food, service, price, ambiance, etc.
   e6) The soup was sweet!
5. Miscellaneous.
   If not fallen in none of the above classes.
   e7) I drink water.

Pessimism and optimism classes seem to be similar to the negative and positive classes, respectively, as the conventional sentiment analysis. However, two important classes are included such as “sarcastic” and “suggestion” in this study as compared with conventional sentiment analysis. The “sarcastic” class is very crucial for avoiding misclassification of pessimism/negative and optimism/positive and hence, providing a more accurate classification model. According to recent psychological studies, sarcasm has an influential relationship with sentiment features [3,4]. Pessimism/negative and optimism/positive contain affect/sentiment/emotion features (according to the definition in this study and conventional sentiment analysis). Therefore, when the inference or generalization occurs, the new/unseen text might be misclassified. The author’s implicit information/intention might be the inverse or quite different meaning of what he/she conveyed. Sarcasm detection is more challenging in the textual data due to no exposure to body gestures, tone of voice, and facial expressions [5]. Moreover, the “suggestion” class is included in this study and it is also crucial. The suggestion reviews might be closely related in either positive or negative classes in the conventional sentiment analysis. For example, the review “You should increase the items”, is probably implicitly related to the negative class, but it is difficult for a model to firmly say the example text has fallen into the negative class. However, it is concretely classified in this study. Therefore, specific suggestions can be obtained from the customers' reviews.

Intent analysis is important for retailers’ point of view in e-business. For instance, the business owners inspect the consumers’ comments to understand their fragility and robustness. Then it would aid business owners in making the policies better than their competitors.

This study performed the task in the Bangla language. Various methods have been proposed for the English language. However, these methods require massive data to produce a strong model. However, accumulating these data is time-consuming and laborious, especially in less dominant languages such as Bangla even though it has 265 million talkers across the earth [6]. The principal cause is the insufficiency of publicly available datasets [7]. The task complexity is high in intent classification due to handling more classes as compared to a simple binary (i.e., positive or negative) sentiment classification. To get satisfying performance in such a situation, we need a massive amount of proper data to train a model. However, resource-constrained languages like Bangla lag with massive amounts of proper training data to generate a robust model. To tackle this situation, we generate and collect Bangla data in the restaurant domain for the task by ChatGPT API[1] using prompt and data augmentation. These data are referred to as “source data”. We intend to utilize ChatGPT data as it is convenient to generate much data rather than relying on machine or manual translation from other high-resource user-generated data like English. Again, we annotate, prepare, and validate a new Bangla intent analysis dataset in the restaurant domain by collecting user-generated data from different publicly available sources[2,3,4]. These data are

[1] https://platform.openai.com/docs/api-reference/
[2] https://github.com/atik-05/Bangla_ABSA_Datasets
[3] https://github.com/eftekhar-hossain/Bengali-Restaurant-Reviews
[4] https://github.com/sazzadcsedu/BanglaRestaurant

called “target data”. The target data has randomness as they are user-generated whereas the source data follows a pattern or augmentation type as they are generated by ChatGPT. Therefore, there is a divergence gap in the nature between the source and target data. The existing studies [8] and [9] performed intention detection in discussion forums by utilizing transfer learning and domain adaptation techniques, respectively. They performed the task in English without considering the sarcastic reviews for their classification which can significantly affect the intent/implicit opinion mining task. However, sarcasm is handled in this study. In addition, they utilized user-generated data (source data and target data) for transfer learning and domain adaptation techniques. As the source data in this study is not user-generated, rigorous techniques are required to reduce the gap in nature between the source and target data. Therefore, to bridge the gap in the nature between source and target data, we propose a semi-supervised self-training with BERT to inject noise into the training data through pseudo-labeling to avoid bias/overfitting.

By utilizing both source and target data, three approaches are proposed: combined data, semi-supervised learning, and stepwise learning. In the combined data approach, both source and target data are combined. Then, perform training (and validation) and testing. In this case, source data is utilized only for the training (and validation) but not for the testing as they are generated by the ChatGPT model. In semi-supervised learning, a self-training approach is employed. In this approach, at first, some portion of the data from the training set of the combined source and target data is deliberately unlabeled. Then a machine learning (e.g., Logistic Regression) or transformers-based classifier (e.g., BanglaBERT) is trained with the rest of the labeled data. Then, the model predicts and labels the unlabeled data. These data are called pseudo-label data. After that, this pseudo-labeled data is combined with the already labeled training set. The training process, and generating and adding pseudo-label data to the training set are repeated until a certain number of iterations are reached. At last, the trained model is utilized for the inference on the test set of the target data. Stepwise learning comprises two stages: source/auxiliary and target tasks. For the source/auxiliary task, a BERT-based model is trained by utilizing source data. In this case, the model does not use the target data. Then, the model is re-used after the source/auxiliary task. Then, it is re-trained and evaluated with the target dataset. The study [10] proved that in stepwise learning, trained knowledge is conveyed from the source task to the target task to boost performance.

In this study, the main issue is the divergence nature between the source and target datasets. Based on the data-related issue, four research questions (RQ) and their hypotheses are formulated. The RQs and their hypotheses are given below:

RQ1: Does a naïve combination of the source and target data improve the model performance as compared to the baseline (i.e., without source data) because the combined data contains much data? Usually, machine learning models require massive data for training to generate robust models.

Hypothesis (H1):
NO: for conventional machine-learning-based models; because they have no pre-trained knowledge and there is a divergence between source and target dataset. There is a possibility that the model does not learn well from the source data as it is generated by ChatGPT which has pre-trained knowledge in Common Crawl, Wikipedia, books, online news, journals, etc. [11]. In other words, the trained model becomes more biased and overfitted towards target data. Note that the proportion of source and target data in the combined data is 4.88:1.
Hypothesis (H2):
YES: for transformers-based models (e.g., BanglaBERT); because they have similar kind of pre-trained knowledge of ChatGPT. In addition, there is a high proportion of the source data in the combined data as mentioned earlier.

RQ2: Does semi-supervised learning improve the model performance than the combined data approach?

Hypothesis (H3):
YES: because it predicts and labels data (i.e., pseudo-labeled) based on the trained knowledge from both source and target data. Then these pseudo-labeled data are later utilized for the training. Thus, the model can mitigate the divergence gap between the source and target data. This case is true for both conventional machine-learning- and transformers-based models. One study demonstrated [12] that the semi-supervised self-training approach works well even when labeled data is adequate although the study performed the task for image data. Similarly, in this study, the labeled data is much more than the unlabeled data when the semi-supervised self-training is performed.

RQ3: Does transformers-based semi-supervised learning perform better than conventional machine-learning-based semi-supervised learning?
Hypothesis (H4):

YES: because transformers-based models (e.g., BanglaBERT) have a similar type of pre-trained knowledge like ChatGPT. On the other hand, conventional machine-learning models have no pre-trained knowledge, and they are trained from randomly generated initial parameters. Pseudo-label is unable to achieve satisfying performance on models trained from scratch but assists in preventing overfitting in pre-trained models [13]. However, the study [13] performed the task for image data although this study performed the task on the textual data.

RQ4: Does stepwise learning [10] enhance the baseline performance utilizing source and target datasets?

Hypothesis (H5):
NO: in the study [10], stepwise learning improved the performance as both source and target data are user-generated data. Therefore, no divergence gap between the nature of the data. However, in this study, when the already trained model (in the second stage) is re-used for the target task, the model may not transfer better/useful knowledge to the target model due to the divergence in the nature between the two datasets.

The overview of the RQs and their corresponding hypothesis are illustrated in Figure 1.

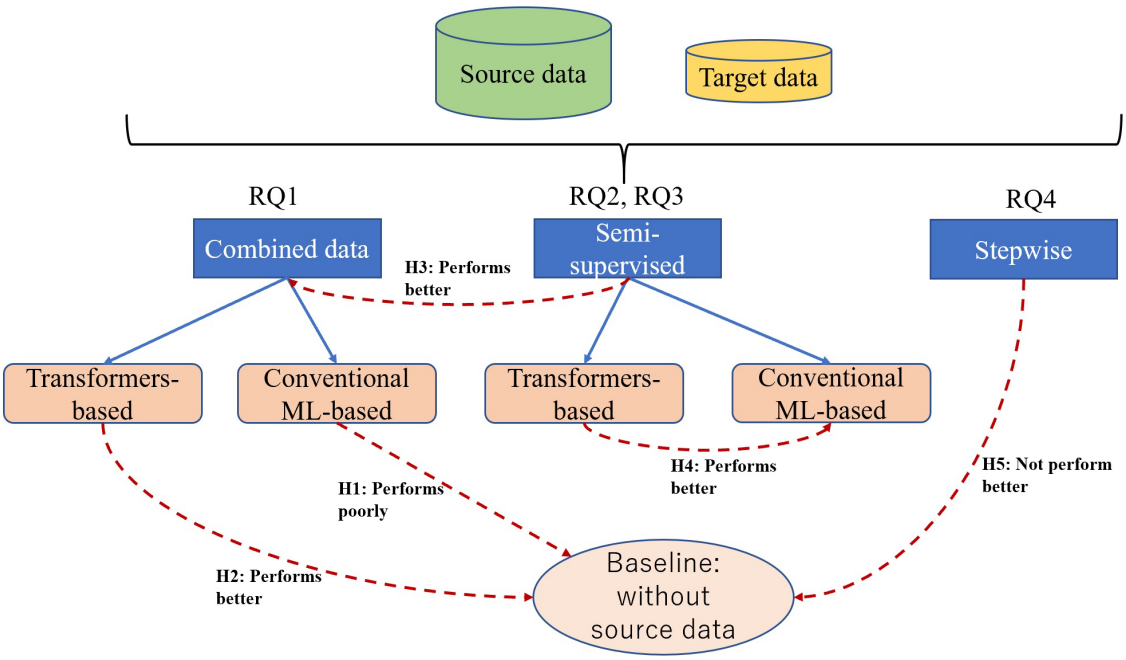

Fig. 1 *Flowchart of research questions and their corresponding hypothesis*

The main contributions to this study are summarized below:

1. We generated Bangla data in the restaurant domain by ChatGPT API, then collected and pre-processed the data to make a source dataset.
2. We collected real data and then prepared a new Bangla intent analysis dataset in the restaurant domain, and made it publicly available for further research.
3. Different baselines are set utilizing conventional machine-learning- and transformers-based methods.
4. The baseline results are compared with the proposed approaches and demonstrate the efficacy of these approaches.

2. Dataset

2.1 Target Data

2.1.1 Data Collection

1,250 sentences are collected which are publicly available for the restaurant domain in the Bangla language[2,3]. The datasets are elaborately discussed in the study [14] and [15], respectively. 75 sentences are also collected that are publicly available for the restaurant domain in the English language[4]. This dataset is discussed in detail

Int. Journal of Integrated Engineering Vol. 16 No. 7 (2024) p. 12-25

in the study [16]. As the target language in this study is Bangla, these sentences are translated into Bangla by machine translation, then verified the correctness of the translation and merged the sentences with 1,250 sentences. Therefore, the total number of instances is 1,325.

2.1.2 Data Annotation

In this study, intent analysis is considered a sentence-level classification task. The first author of this paper (a native Bangla speaker) annotated class tags in the data. Five tags are applied for the task: 0 (pessimism), 1 (optimism), 2 (suggestion), 3 (sarcastic), and 4 (miscellaneous).

2.1.3 Data Verification and Statistics

For verification of data, a native Bangla speaker is nominated. He knows the intent analysis task. Fleiss Kappa is calculated between two persons: this person and the corresponding author of this paper. The obtained Kappa score is 0.89 which indicates the strength of agreement is excellent. There is a total of 404 pessimism, 460 optimism, 146 suggestion, 75 sarcastic, and 240 miscellaneous tags in the prepared dataset. It is made openly accessible for future research[5].

2.1.4 Data Examples

Five examples are presented from the prepared dataset in Table 1. To make it comprehensible to common readers, its corresponding English-version sentences are supplied as well.

Table 1 *Examples from the target dataset*

No.	Bangla sentence	English-translated sentence	Tag (class)
1	টাকা অপচয় ...	Waste of money...	0 (pessimism)
2	শ্রেষ্ঠ খাবার!	Best food!	1 (optimism)
3	আইটেম বৃদ্ধি করা উচিত	The item should increase	2 (suggestion)
4	সর্বকালের সেরা মিষ্টি এবং টক স্যুপ!	The best sweet and sour soup ever!	3 (sarcastic)
5	আমাদের বন্ধু এখনো আসে নি।	Our friend has not come yet.	4 (miscellaneous)

2.2 Source Data

2.2.1 Direct Prompt for Data Generation and Pre-processing

We generated data for each class through prompts. For this purpose, OpenAI ChatGPT API[1] is utilized. The direct prompt is given in Appendix A. We set model = "gpt-3.5-turbo", temperature = 0.8, and the rest of the hyper-parameters default.

There are unnecessary tokens, English sentences, and duplicate Bangla sentences that appear in the generated data. Therefore, the unnecessary tokens, English sentences, and duplicate data are removed. After that, a total cleaned 5,000 Bangla sentences are retained where each class has an equal number of sentences (i.e., 1,000 sentences per class).

2.2.2 Prompt for Augmentation and Pre-processing

The distribution of the class of the prepared target dataset is uneven. To make class distribution fairly even, concerning the pessimism and optimism classes, data of suggestion, sarcastic, and miscellaneous classes are increased by augmentation. However, these augmented data are also regarded as source data. OpenAI ChatGPT API[1] is utilized for this data augmentation process. The prompt for augmentation is given in **Appendix B**. We set model = "gpt-3.5-turbo", temperature = 0.8, and the rest of the hyper-parameters default. Note that the test set (10%) of the target data is drawn away before the augmentation to avoid data leakage issues between the models' training and testing. 264 suggestion, 335 sarcastic, and 216 miscellaneous sentences are generated based on the training set of the target data. There are unnecessary tokens and English sentences that appear along with the generated data. They are removed. Therefore, the total cleaned Bangla sentences is 815. Now, these sentences are merged with 5,000 sentences (mentioned in Section 2.2.1) to prepare the source data. Hence, the final source dataset contains a total of 5,815 Bangla sentences.

[5] https://github.com/al-mahmud28/Bangla-intent-analysis

3. Related Work

A traditional classification model like a support vector machine (SVM) is proposed for commercial intent identification in the study [17]. For the classification task, WEKA is employed. In the study [18], a semi-supervised approach is utilized to categorize tweets into six classes. A neural network (NN)-based approach is proposed for medical text queries in the study [19]. A consumption intention mining model (CIMM) is proposed in the study [20] which focuses on mining implicit intent utilizing a convolutional neural network (CNN). A deep learning technique such as long short-term memory (LSTM) is presented in the study [21] for predicting customer purchase intention. An encoder-decoder with attention model is proposed by the study [22] for implicit intention identification. The study [23] proposed several feature extractors that deal with microblog classification. An ensemble method is utilized to merge the feature extractors. In the study [24], a weakly-supervised approach is presented for consumption intent detection. The task is accomplished as a binary classification problem. The study [25] proposed a graph-based ranking method that jointly models relevance and associativity. The study [26] presented a novel approach for suggestion intent understanding and classification. There are three major stages in this system: suggestion refining, suggestion zone recognition, and argument specification. In the study [27], several methods such as recurrent neural networks (RNN) or CNN are proposed to incorporate context information. In the study [28], a joint intent classification and slot-filling model is proposed based on a transformers-based model and demonstrated significant improvement on several public benchmark datasets. The study [29] proposed an attention-based multi-task model for intent analysis of Chinese online medical questions and established that the use of both attention and multi-task learning is effective. The study [30] proposed a novel multimodal-based approach to determine the marketing intent and demonstrated the merits of the method from multiple aspects. In the study [31], SVM, naïve Bayes (NB), and deep-learning-based models are employed for intention analysis. The authors combined consumption intention with traditional features used in the problem of box office prediction to achieve better performance. The study [32] proposed a framework for swift and precise intent classification for dialogue systems. The study [33] utilized a Joint BERT model which yielded high accuracy in intent analysis. In the study [34], different models, namely SVM, Stochastic Gradient Descent (SGD), and NB are proposed for intent classification. The authors of the study [35] introduced several heuristic and machine-learning approaches that have been considered for optimum results for intent classification. The authors of the study [36] proposed a multi-task learning (MTL) approach utilizing a transformer-based model with an attention mechanism. The study [37] proposed contrastive learning with k-nearest neighbor (KNN) algorithm for out-of-domain intent classification. The study [38] proposed several transformers-based models for fine-grained intent classification in the legal domain. The authors of the study [39] proposed a prompting-based approach to generate labeled training data for intent classification with off-the-shelf language models (LMs) such as GPT-3.

This study employed transformers-based models in addition to the conventional machine-learning-based models. To the best of our awareness, no existing studies worked on the Bangla intent analysis study. Hence, no Bangla dataset is available. Therefore, a new Bangla dataset is prepared in the restaurant domain by collecting user-generated data. Then the prepared dataset is made publicly available. Moreover, some crucial classes such as suggestion and sarcastic are considered in this study, which have various merits, as discussed in Section 1.

4. Method

4.1 Conventional Machine-learning-based Method

Five popular conventional machine-learning methods, namely support vector machine (SVM), random forest (RF), logistic regression (LR), multinomial naive Bayes (MNB), and multilayer perceptron classifier (MPC) are employed. For the extraction of feature, the bag of words (BOW), the term frequency–inverse document frequency (TF-IDF), BERT, BERT + BOW, BERT + TF-IDF, and BERT + BOW + TF-IDF are utilized. Here, the operator "+" denotes the concatenation of the features.

4.2 Transformers-based Method

4.2.1 mBERT

mBERT is a multilingual BERT-based model [40]. The pre-training procedure of mBERT is the same as BERT. It is fine-tuned by adding linear layers on top of it. Thus, linear layers are added to accomplish the task in this study.

4.2.2 BanglaBERT

BanglaBERT is a pre-trained BERT-based model in Bangla [41]. Like the BERT model, it is possible to fine-tune it by adding linear layers on top of it. Subsequently, linear layers are incorporated to accomplish the task.

4.2.3 BanglishBERT

BanglishBERT is a BERT-based bilingual model pre-trained in the Bangla and English languages so that it can transfer knowledge from English to the Bangla language [41]. Similar to mBERT and BanglaBERT, BanglishBERT is also applied in this study.

4.3 Combined Data Approach

Step I:

1. Split the source data into training (and validation) sets.
2. Merge with the corresponding training (and validation) set of the target data.
3. Train the model with the merged data.

Step II: Perform testing using the test set of the target data.

This technique is performed on both machine-learning-based and transformers-based methods.

4.4 Semi-supervised Approach

Here a self-training approach is utilized. The following steps are involved with this approach.
Step I: Merge the source and training sets of target data.
Step II: Unlabel some portion of data deliberately from the training set of the merged data. This is called an unlabeled dataset.
Step III: Train a machine learning (e.g., LR) or transformers-based classifier (e.g., BanglaBERT) with the labeled data of the merged training set. This trained model is called the base model.
Step IV: Utilize the base model to predict and label the unlabeled data based on some level of confidence (i.e., pseudo-labeling).
Step V: Add this pseudo-labeled data to the labeled training set.
Step VI: The textual data corresponding to the pseudo-labels are now discarded from the unlabeled dataset for generating the next pseudo-label in the next iteration.
Step VII: Now the currently trained model becomes the base model for the next iteration.
Step VIII: Repeat Steps (III-VII) with the current training data until a certain number of iterations is reached.
Step IX: Finally, the trained model is employed for the inference on the test set of the target data.

This technique is performed on both machine-learning-based and transformers-based methods.

4.5 Stepwise Learning

The stepwise learning approach is implemented utilizing the source and target data. It enhances performance for the main task by taking advantage of source/auxiliary tasks. It is useful when the target tasks' data are scarce. It has two phases: source and target tasks. In the first phase (i.e., source task), a transformers-based model is trained by the source data to update the pre-trained parameters. For the second phase (i.e., target task), the model is which is obtained from the first phase, and then re-trained and evaluated using target data. Note that stepwise learning utilizes the transformers-based models, not the conventional machine-learning-based models [10].

5. Experiment and Analysis

5.1 Baselines

A few well-known conventional machine-learning-based and pre-trained vanilla transformers-based models are implemented for the baselines. Without source data (i.e., only utilizing the target data) approach is considered as a baseline for both conventional machine-learning-based and transformers-based models.

5.2 Experimental and Hyper-parameters Settings

The experiments were executed on a Linux server (CPU: Xeon E5-2620@2.10GHz 32proc, Mem: 256GB, GPU: Quadro RTX8000 (48GB)) and accomplished in Python 3.6. The *F*1 (macro) score is used as an assessment metric.

The hyper-parameters settings are as follows:

1. Conventional machine-learning (both without source and combined data):
   Data splitting = 90:10 and default settings and default settings (for without source data case).
   10% of the target data for testing and the rest of the data for training, and default settings (for combined data case).
2. Transformers-based (without source data):
   Data splitting = 80:10:10, epoch = 5, batch size = 32, learning rate = 1e-3, and optimizer = AdamW.

3. Transformers-based approach (combined data):
   Source data splitting = 80:20 (training:validation), target data splitting = 80:10:10, epoch = 5, batch size = 32, learning rate = 1e-3, and optimizer = AdamW.
4. Semi-supervised learning:
   Conventional machine-learning approach: unlabeled training data = 8%, threshold = 0.95, and max_iter=3.
   Transformers-based approach: source data splitting = 80:20 (training:validation), target data splitting = 80:10:10, epoch = 3, batch size = 32, learning rate = 1e-3, and optimizer = AdamW, unlabeled training data = 8%, epoch for self-training = 3, and level of confidence for pseudo-labeling = 0.95.
5. Stepwise learning:
   In the auxiliary/source task, data splitting (training:validation) = 90:10, epoch = 1, batch size = 32, learning rate = 1e-3, and optimizer = AdamW.
   In the target/main tasks, data splitting = 80:10:10, epoch = 5, batch size = 32, learning rate = 1e-3, and optimizer = AdamW.

5.3 Experimental Result Analysis

Table 2 presents the experimental results of popular conventional machine-learning-based methods. Here three approaches are applied: without source data, with source data (i.e., combined data), and semi-supervised self-training. For this purpose, we considered five classifiers: support vector machine (SVM), random forest (RF), logistic regression (LR), multi-nominal naive Bayes (MNB), and multi-layer perceptron classifier (MPC). The "+" symbol indicates concatenation of features.

Table 2 *Experimental results using conventional machine-learning-based methods (boldface is the best model and score)*

Approach	Model	BOW	TF-IDF	BERT	BERT + BOW	BERT + TF-IDF	BERT + BOW + TF-IDF
					*F*1 score		
Without Source Data (Baseline)	SVM	0.56	0.56	0.48	0.58	0.51	0.59
	RF	0.63	0.63	0.47	0.50	0.47	0.56
	LR	0.65	0.57	0.63	**0.72**	0.70	**0.72**
	MNB	0.58	0.45	0.52	0.56	0.52	0.53
	MPC	0.65	0.64	0.63	0.69	0.71	0.69
Combined Data (Proposed)	SVM	0.59	0.62	0.52	0.56	0.53	0.61
	RF	0.61	0.59	0.49	0.47	0.50	0.45
	LR	0.60	0.59	0.52	0.64	0.56	0.66
	MNB	0.65	0.64	0.46	0.45	0.45	0.45
	MPC	0.66	0.62	0.55	0.63	0.55	0.59
Semi-Supervised Learning (Proposed)	SVM	0.60	0.65	0.50	0.57	0.52	0.61
	RF	0.61	0.59	0.53	0.52	0.51	0.49
	LR	0.61	0.60	0.55	0.65	0.55	0.68
	MNB	0.64	0.62	0.44	0.42	0.45	0.44
	MPC	0.62	0.64	0.52	0.71	0.63	0.66

For the without source data case, we can see that in most of the cases, the concatenation of the features outperformed the non-concatenated approaches. The concatenation of different feature embeddings can capture different characteristics and representations of the input sentences. For example, BERT captures local contextual information within a sentence, and BOW and TF-IDF contribute more global information about the overall dataset. Thus, concatenating them enables the model to be more adaptable to diverse linguistic patterns. Hence, it performed well, and the same thing occurred in this experiment. In that case, the BERT + BOW and BERT + BOW + TF-IDF jointly yielded the best result for LR, and the score was 0.72 (shown in bold). For the combined data approach case, we also see that in most of the cases, the concatenation outperformed the non-concatenated approaches. However, the results were not improved except for some cases as compared with without source data approach. It indicated that the first hypothesis (H1) was true for RQ1. In that case, the BOW and BERT + BOW + TF-IDF jointly yielded the best result for MPC and LR, respectively and the score was 0.66. For semi-supervised cases, the majority of the cases the concatenation performed better than the non-concatenated approaches. Moreover, in most cases, the semi-supervised approach performed better than the combined data approach. It proved that the hypothesis (H3) for RQ2 was true. However, the obtained results were not satisfying among all the tested models. In this case, the best score is 0.71 for BERT + BOW with MPC.

Int. Journal of Integrated Engineering Vol. 16 No. 7 (2024) p. 12-25

The graphical plots of Table 2 are illustrated in Figure 2. The best F1 score (i.e., 0.72) is obtained for BERT + BOW and BERT + BOW + TF-IDF jointly with LR in the case of without source data approach, i.e., the baseline. Here both the proposed approaches failed to provide improved results than the baseline.

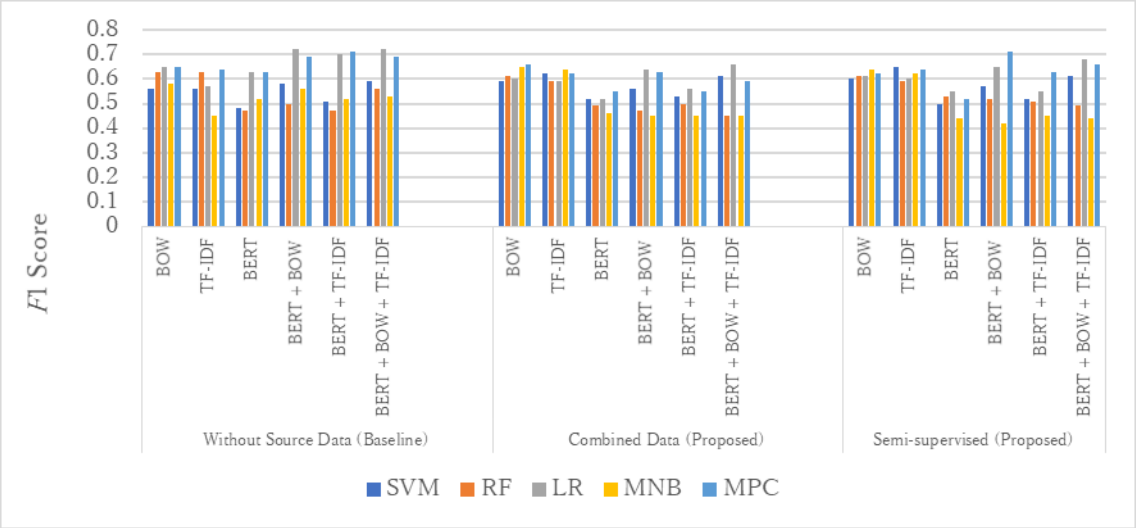

Fig. 2 *Bar diagram of the results using conventional machine-learning-based methods*

Table 3 displays the experimental results of the transformers-based method. For this, four approaches are applied: without source data, combined data, semi-supervised self-training, and stepwise learning. For each case, three models are utilized: mBERT, BanglaBERT, and BanglishBERT.

Here we can see the combined data approach improved the model performance as compared to the baseline. However, the situation was contrary to the conventional machine-learning-based models (shown in Table 2). It demonstrated that the second hypothesis (H2) for RQ1 was also true. The BERT-based models have pre-trained knowledge about Wikipedia and BooksCorpus [40]. It can understand the source data (i.e., ChatGPT data) better than conventional machine-learning-based methods such as SVM, and ChatGPT is also pre-trained with Common Crawl, Wikipedia, online news, journals, books, etc. [11]. Here the semi-supervised learning performed better than the combined data approach. It indicated the hypothesis (H3) for RQ2 was true. We also see that transformers-based semi-supervised learning performed better than conventional machine-learning-based semi-supervised learning (best score: 0.74 vs. 0.71 shown in Tables 3 and 2, respectively). It confirmed the hypothesis (H4) in RQ3. The semi-supervised self-training method performed best for BanglaBERT among all the transformers-based methods, and the score was 0.74 (shown in bold). The stepwise learning did not perform well in this study. Therefore, it provided the evidence for the hypothesis (H5) in RQ4.

Table 3 *Experimental results using transformers-based methods (boldface is the best model and score)*

Approach	Model	*F*1 score
Without Source Data (Baseline)	mBERT	0.58
	BanglaBERT	0.63
	BanglishBERT	0.67
Combined Data (Proposed)	mBERT	0.67
	BanglaBERT	0.73
	BanglishBERT	0.71
Semi-Supervised Learning (Proposed)	mBERT	0.70
	BanglaBERT	**0.74**
	BanglishBERT	0.71
Stepwise Learning (Proposed)	mBERT	0.67
	BanglaBERT	0.56
	BanglishBERT	0.51

BERT + BOW + TF-IDF with the LR model performed best for the conventional-machine-learning-based methods (shown in Table 2). In that case, the BERT feature was the pre-trained BERT feature. Therefore, we intended to further improve the performance as an additional experiment by using the fine-tuned BERT feature instead of the pre-trained BERT feature because the fine-tuned BERT can provide a better representation of the features. The results are reported in Table 4.

Table 4 *Experimental results using BERT + BOW + TF-IDF with the LR (boldface is the best model and score)*

Data Usage	Feature	Fine-tuned BERT	*F*1 score
Fine-tuned BERT with source data and passed the target data to the fine-tuned BERT for the feature extraction.	Fine-tuned BERT feature + BOW + TF-IDF (Proposed)	**mBERT**	**0.74**
		BanglaBERT	0.67
		BanglishBERT	0.65
Fine-tuned BERT with the combined source and target data, and extracted features for the combined data.	Fine-tuned BERT feature + BOW + TF-IDF (Proposed)	**mBERT**	**0.74**
		BanglaBERT	0.70
		BanglishBERT	0.70

Here the best result was obtained for fine-tuning the mBERT model. Note that in Table 2, mBERT was utilized as BERT. The reason is it provided better results than BanglaBERT and BanglishBERT when concatenated with BOW and TF-IDF. Between the two techniques of data usage in Table 4, the training time and memory allocation were less in the first technique as the fine-tuning was performed only with the source data.

The concatenation of the last four hidden states of pre-trained BERT features is effective which is illustrated in the study [40]. Therefore, for further improvement of the results, the concatenation of the last four hidden states of BERT features was obtained by passing the target data to the pre-trained BERT model. Then the extracted features were concatenated followed by the features mentioned in the first method in Table 4 and after that, the LR model was trained. The first method in Table 4 was utilized rather than the second method due to less training time and memory allocation as mentioned above. The results were significantly improved, even better than semi-supervised learning in Table 3. However, two types of BERT models were needed: (1) pre-trained and (2) fine-tuned. In addition, the dimension of the feature became high. The results are shown in Table 5.

Table 5 *Experimental results using BERT + BOW + TF-IDF with the LR (boldface is the best model and score)*

Data Usage	Feature	Pre-trained and fine-tuned BERT	*F*1 score
a. Passing target data to the pre-trained BERT model for the feature extraction. b. Fine-tuned BERT with source data and passed the target data to the fine-tuned BERT for the feature extraction.	Concatenation of the last four hidden states of pre-trained BERT features+ Fine-tuned BERT feature + BOW + TF-IDF (Proposed)	mBERT	0.73
		BanglaBERT	**0.75**
		BanglishBERT	0.74

In this experiment, the best score was 0.75 for BanglaBERT. This score outperformed all tested methods in this study.

6. Conclusion

To the best of our awareness, no previous study accomplished an intent analysis task for the Bangla language, and no dataset was accessible. To this end, a new dataset was prepared. Then this study employed different conventional machine-learning- and transformers-based methods to perform the task. For the conventional

machine learning-based method, three approaches were applied: without source data, combined data, and semi-supervised self-training. For the transformers-based method, four approaches were implemented: without source data, combined data, semi-supervised self-training, and stepwise learning. This study proved that semi-supervised learning performed best in the transformers-based methods and yielded the best performance among the above-mentioned approaches. The answers to each research question (RQ) are summarized in Table 6.

Table 6 *Answer to each research question*

RQ	Answer
RQ1	The combined data approach improved the performance of transformers-based models but the scenario was the opposite for the conventional machine-learning-based models.
RQ2	Semi-supervised learning improved the model performance than the combined data approach.
RQ3	Transformers-based semi-supervised learning performed better than conventional machine-learning-based semi-supervised learning.
RQ4	Stepwise learning did not improve the baseline performance due to the divergence nature between the source and target data.

This study concluded that when combining data from different sources to create much data, the nature of the data should be considered. Moreover, the compatible model should be selected so that it can learn better from the data. Divergence between the nature may degrade model performance instead of improving it. The same thing happened in the experiment for the conventional machine-learning-based methods. However, the scenario was different for the transformers-based models. We handled the data scarcity problem (i.e., limited user-generated target data) in the Bangla language by exploiting ChatGPT-model-generated data by employing a semi-supervised self-training approach with transformers-based models.

The restaurant domain data was regraded in this study. In upcoming research, datasets from various domains can be used to extend the applicability of this study. Additionally, Bangla fine-grained intent classification by incorporating more classes can be an attractive study.

Acknowledgement

This research was conducted in the Shimada laboratory, Dept. of Artificial Intelligence, Kyushu Institute of Technology, Japan. The funding for this work is aided by JST SPRING, Grant Number JPMJSP2154.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm the contribution of the paper as follows: ***Dataset collection and construction, implementation, and manuscript preparation:*** *Al-Mahmud;* ***study conception and idea generation:*** *Kazutaka Shimada. Both authors reviewed and approved the final version of the manuscript.*

Appendix A

"Generate 30 pessimistic sentences about the food, service, price, ambiance, etc. on the restaurant domain in the Bangla Language. The sentences must be different from each other. The sentences must not contain any English language."

"Generate 30 optimistic sentences about the food, service, price, ambiance, etc. on the restaurant domain in the Bangla Language. The sentences must be different from each other. The sentences must not contain any English language."

"Generate 30 suggestion sentences to the restaurant authority from the customer about food, service, price, ambiance, etc. on the restaurant domain in the Bangla Language. The sentences must be different from each other. The sentences must not contain any English language."

"Generate 30 sarcastic sentences about the food, service, price, ambiance, etc. on the restaurant domain in the Bangla Language. The sentences must be different from each other. The sentences must not contain any English language."

"Generate 30 sentences in the Bangla Language. The sentences must be different from each other. The sentences must not contain any English language."

For the token limitation of the engine "gpt-3.5-turbo" (maximum 4,096), we generated 30 sentences at a time using the prompt. Hence, we utilized a loop in the program to generate all the required sentences for each particular class.

Appendix B

f"Generate 2 augmented sentences from {row}."

Here, *"row"* is each row of the data file i.e., each sentence of the suggestion sentences from the target dataset.

f"Generate 5 augmented sentences from {row}."

Here, *"row"* is each row of the data file i.e., each sentence of the sarcastic sentences from the target dataset.

f"Generate 1 augmented sentence from {row}."

Here, *"row"* is each row of the data file i.e., each sentence of the miscellaneous sentences from the target dataset.

References

[1] Al-Mahmud, & Shimada, K. (2023) Dataset Construction and Opinion Holder Detection Using Pre-trained Models, *IJSKM, 7(2)*, 1-17, https://doi.org/10.52731/ijskm.v7.i2.779

[2] Li, J. & Hovy, E. (2017) Reflections on sentiment/opinion analysis, *In: A practical guide to sentiment analysis,* 41–59, DOI: 10.48550/ARXIV.1507.01636

[3] Huang, L., Gino, F., & Galinsky, A. D. (2015) The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients, *Organizational Behavior and Human Decision Processes, 131*, 162-177, https://doi.org/10.1016/j.obhdp.2015.07.001

[4] Pickering, B., Thompson, D., & Filik, R. (2018) Examining the emotional impact of sarcasm using a virtual environment, *Metaphor and Symbol, 33(3)*, 185-197, DOI: 10.1080/10926488.2018.1481261

[5] Babanejad, N., Davoudi, H., An, A., & Papagelis, M. (2020) Affective and Contextual Embedding for Sarcasm Detection, *In Proceedings of the 28th International Conference on Computational Linguistics,* 225-243, DOI: 10.18653/v1/2020.coling-main.20

[6] Sen, O., Fuad, M., Islam, M. N., Rabbi, J., Hasan, M. K., Fime, A. A., Fuad, M. T. H., Sikder, D., & Iftee, M. A. R. (2021) Bangla Natural Language Processing: A Comprehensive Review of Classical, Machine Learning, and Deep Learning Based Methods, *CoRR*, abs/2105.14875, https://arxiv.org/abs/2105.14875

[7] Karim, M. A., Kaykobad, M., & Murshed, M. (2013) Technical Challenges and Design Issues in Bangla Language Processing, *IGI Global*. ISBN: 9781466639706, DOI: 10.4018/978-1-4666-3970-6

[8] Chen, Z., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013) Identifying Intention Posts in Discussion Forums, In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1041-1050, https://aclanthology.org/N13-1124

[9] Bach, N. X., Linh, L. C., & Phuong, T. M. (2017) Cross-Domain Intention Detection in Discussion Forums, In *Proceedings of the 8th International Symposium on Information and Communication Technology*, 173–180, https://api.semanticscholar.org/CorpusID:30080459

[10] Al-Mahmud, & Shimada, K. (2023) Demonstration of Effectiveness of Nativeness in Stepwise Learning by Performing Sentiment Analysis, In *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 1-6, DOI: 10.1109/NCIM59001.2023.10212970

[11] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020) Language models are few-shot learners, *arXiv preprint arXiv:*2005.14165, https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[12] Xie, Q., Hovy, E. H., Luong, M.-T., & Le, Q. V. (2019) Self-training with Noisy Student improves ImageNet classification, *CoRR*, volume abs/1911.04252, http://arxiv.org/abs/1911.04252

[13] Zhou, H.-Y., Oliver, A., Wu, J., & Zheng, Y. (2018) When Semi-Supervised Learning Meets Transfer Learning: Training Strategies, Models, and Datasets, *CoRR*, abs/1812.05313, http://arxiv.org/abs/1812.05313

[14] Rahman, M. A., & Kumar Dey, E. (2018) Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation, *Data, 3(2), 15*. ISSN: 2306-5729, DOI: 10.3390/data3020015

[15] Sharif, O., Hoque, M. M., & Hossain, E. (2019) Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Naïve Bayes, In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 1-6, DOI: 10.1109/ICASERT.2019.8934655

[16] Sazzed, S. (2021) A Hybrid Approach of Opinion Mining and Comparative Linguistic Analysis of Restaurant Reviews, In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1281-1288, https://aclanthology.org/2021.ranlp-1.144

[17] Hollerit, B., Kröll, M., & Strohmaier, M. (2013) Towards linking buyers and sellers: detecting commercial Intent on Twitter, In *Proceedings of the 22nd International Conference on World Wide Web*, 629–632, https://api.semanticscholar.org/CorpusID:310513

[18] Wang, J., Cong, G., Zhao, X., & Li, X. (2015) Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets, In *Proceedings of the AAAI Conference on Artificial Intelligence*, *29(1)*, 9196, DOI: 10.1609/aaai.v29i1.9196

[19] Zhang, C., Fan, W., Du, N., & Yu, P. S. (2016) Mining User Intentions from Medical Queries: A Neural Network Based Heterogeneous Jointly Modeling Approach, In *Proceedings of the 25th International Conference on World Wide Web*, 1373–1384, https://api.semanticscholar.org/CorpusID:1697236

[20] Ding, X., Liu, T., Duan, J., & Nie, J.-Y. (2015) Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network, In *Proceedings of the AAAI Conference on Artificial Intelligence, 29(1), 9529,* DOI: 10.1609/aaai.v29i1.9529

[21] Korpusik, M., Sakaki, S., Chen, F., & Chen, Y.-Y. (2016) Recurrent Neural Networks for Customer Purchase Prediction on Twitter, In *CBRecSys@RecSys*, 47–50, https://api.semanticscholar.org/CorpusID:16783231

[22] Li, C., Du, Y., & Wang, S. (2017) Mining Implicit Intention Using Attention-Based RNN Encoder-Decoder Model, In *Intelligent Computing Methodologies*, 413-424, https://link.springer.com/chapter/10.1007/978-3-319-63315-2_36

[23] Banerjee, N., Chakraborty, D., Joshi, A., Mittal, S., Rai, A., & Ravindran, B. (2021) Towards Analyzing Micro-Blogs for Detection and Classification of Real-Time Intentions, In *Proceedings of the International AAAI Conference on Web and Social Media, 6(1),* 391-394, https://doi.org/10.1609/icwsm.v6i1.14312

[24] Fu, B., & Liu, T. (2013) Weakly-supervised Consumption Intent Detection in Microblogs, *Journal of Computational Information Systems*, 2423–2431, https://api.semanticscholar.org/CorpusID:113425372

[25] Liu, T., Fu, B., & Chen, Y. (2015) Detecting consumption intention based on graph ranking in social media, Scientia Sinica Information*, 45(12):1523*, https://api.semanticscholar.org/CorpusID:168568365

[26] Ngo, T.-L., Pham, K. L., Takeda, H., Pham, S. B., & Phan, X. H. (2017) On the Identification of Suggestion Intents from Vietnamese Conversational Texts, In *Proceedings of the 8th International Symposium on Information and Communication Technology*, 417–424, https://api.semanticscholar.org/CorpusID:3652021

[27] Liu, Y., Han, K., Tan, Z., & Lei, Y. (2017) Using Context Information for Dialog Act Classification in DNN Framework, In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2170-2178, DOI: 10.18653/v1/D17-1231

[28] Chen, Q., Zhuo, Z., & Wang, W. (2019) BERT for Joint Intent Classification and Slot Filling, *CoRR, abs/1902.10909,* http://arxiv.org/abs/1902.10909

[29] Wu, C., Luo, G., Guo, C., Ren, Y., Zheng, A., & Yang, C. (2020) An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions, *Journal of Biomedical Informatics*, *108, 103511*, https://doi.org/10.1016/j.jbi.2020.103511

[30] Zhang, L., Shen, J., Zhang, J., Xu, J., Li, Z., Yao, Y., & Yu, L. (2022) Multimodal Marketing Intent Analysis for Effective Targeted Advertising, *IEEE Transactions on Multimedia, 24,* 830-1843, https://doi.org/10.1109/TMM.2021.3073267

[31] Wu, S. (2020) Internet Public Information Text Data Mining and Intelligence Influence Analysis for User Intent Understanding, *Journal of Intelligent & Fuzzy Systems, 38(1),* 487-494, https://doi.org/10.1016/j.jbi.2020.103511

[32] Hengst, F. den, Wolter, R., Altmeyer, P., & Kaygan, A. (2024) Conformal intent classification and clarification for fast and accurate intent recogni*tion, arXiv:2403.18973*, https://doi.org/10.48550/arXiv.2403.18973

[33] Park, S., Menassa, C. C., & Kamat, V. R. (2023) Joint BERT Model for Intent Classification and Slot Filling Analysis of Natural Language Instructions in Co-Robotic Field Construction Work, *In Computing in Civil Engineering 2023,* 453-460, https://doi.org/10.1061/9780784485224.055

[34] Mustafa, S. N. B., & Zakaria, L. Q. B. (2024) Intent Classification for Malaysian Academic Writers' Proofreader Chatbot Using Machine Learning, *Journal of Theoretical and Applied Information Technology, 102(9),* https://www.jatit.org/volumes/Vol102No9/10Vol102No9.pdf

[35] Chandrakala, C. B., Bhardwaj, Rohit, & Pujari, Chetana. (2024) An intent recognition pipeline for conversational AI, *Internationa*l Journal of Information Technology, 16(2), 731-743, https://doi.org/10.1007/s41870-023-01642-8

[36] Myint, P. Y. W., Lo, S. L., & Zhang, Y. (2024) Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction, *Information Processing & Management, 61(4), 103695*, https://doi.org/10.1016/j.ipm.2024.103695

[37] Zhou, Y., Liu, P., & Qiu, X. (2022). KNN-Contrastive Learning for Out-of-Domain Intent Classification. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,* 5129–5141. https://doi.org/10.18653/v1/2022.acl-long.352

[38] Mullick, A., Nandy, A., Kapadnis, M. N., Patnaik, S., & Raghav, R. (2022). Fine-grained Intent Classification in the Legal Domain, *arXiv preprint arXiv:2205.03509*. https://arxiv.org/abs/2205.03509

[39] Sahu, G., Rodriguez, P., Laradji, I. H., Atighehchian, P., Vazquez, D., & Bahdanau, D. (2022). Data Augmentation for Intent Classification with Off-the-shelf Large Language Models, *arXiv preprint arXiv:2204.01959*. https://arxiv.org/abs/2204.01959

[40] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805,* https://doi.org/10.48550/arXiv.1810.04805

[41] Bhattacharjee, A., Hasan, T., Ahmad, W., Mubasshir, K. S., Islam, M. S., Iqbal, A., Rahman, M. S., & Shahriyar, R. (2022) BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla, *In Findings of the Association for Computational Linguistics: NAACL 2022,* 1318-1327, DOI: 10.18653/v1/2022.findings-naacl.98