IMNTPU Dialogue System Evaluation at the NTCIR-16 DialEval-2 Dialogue Quality and Nugget Detection

Ting-Yun Hsiao
Information Management,
National Taipei University
New Taipei City, Taiwan
s711036112@gm.ntpu.edu.tw

Yung-Wei Teng
Information Management,
National Taipei University
New Taipei City, Taiwan
s711036115@gm.ntpu.edu.tw

Pei-Tz Chiu
Information Management,
National Taipei University
New Taipei City, Taiwan
s711036103@gm.ntpu.edu.tw

Mike Tian-Jian Jiang
Zeal Co., Ltd
Tokyo, Japan
tmjiang@gmail.com

Min-Yuh Day*
Information Management,
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

ABSTRACT

A surge in interest in the evaluation of the quality of chatbot conversation has been observed in recent years. We performed Dialogue Quality (DQ) and Nugget Detection (ND) subtasks in Chinese and English. However, the majority of existing conventional approaches are based on the long short-term memory (LSTM) model. The paper suggests a method for assisting customers in resolving problems. This subtask aims to automatically determine the status of dialogue sentences in a dialogue system's logs. In conversation tasks, we developed fine-tuning methodologies for the transformer model. To evaluate and show the concept, we created a wide framework for testing and displaying the XLM-RoBERTa model's performance on conversational texts. Finally, the experimental findings of the two subtasks demonstrated the efficacy of our strategy. The experimental findings for the DialEval-2 task showed that the suggested method's performance is reasonably equal to that of the LSTM-based baseline model. The main contribution of our study is our suggestion of two crucial elements, namely, tokenization methods and fine-tuning procedures, to increase the conversation quality and nugget identification subtasks in dialogue assessment.

CCS Concepts

• Information systems ~ Information retrieval ~ Retrieval tasks and goals ~ Question answering

Keywords

Tokenization, Fine-tuning, Transformers, Dialogue evaluation, Dialogue quality

TEAM NAME

IMNTPU

SUBTASKS

Dialogue Quality (Chinese, English),

Nugget Detection (Chinese, English).

1. INTRODUCTION

Chatbot has been more popular in a variety of areas, including marketing, health care, and entertainment, in recent years. Such popularity is mostly due to advancements in the disciplines of natural language processing (NLP) and artificial intelligence (AI). A chatbot is an online messaging account that may deliver services to consumers by utilizing instant messaging frameworks with the goal of providing effective conversational services. However, evaluating the systems necessitates a labor-intensive and time-consuming annotation procedure, which is costly and inefficient.

As a result, the task organizers of NTCIR-16 DialEval-2 devised Dialogue Quality (DQ) and Nugget Detection (ND) subtasks [1, 2, 3] to investigate the automatic assessment systems for helpdesk discussions in Chinese or English to bridge the gap. This challenge includes two subtasks: (1) Dialogue Quality (DQ), which aims to quantify a whole dialogue by quality score testing of the accomplishment and effectiveness of the dialogue and the customer's satisfaction, and (2) Nugget Detection (ND), which aims to predict the situational state of the dialogue and determines whether a turn of dialogue is a nugget. Subjective measures are used in the DQ subtask to assess the overall quality of a conversation. The organizers establish three score kinds using a five-degree rank system that sorts from -2 to 2:

1. A-score: Accomplishment

— to what extent an inquiry has resolved a dialogue;

2. S-score: Satisfaction

— how assured a customer is with the conversation;

3. E-score: Effectiveness

— how helpful and economical a dialogue is.

The ND subtask defines a nugget as a dialogue turn, assesses whether it belongs on the Customer or Helpdesk side, and then categorizes it into seven different sorts of four groups.

1. CNaN / HNaN: Non-nuggets from the customer or helpdesk that are unrelated to the problem-solving situation;

2. CNUG / HNUG: Regular nuggets from the customer or helpdesk that are important to the problem-solving situation;

3. CNUG* / HNUG*: Goal nuggets from the customer or helpdesk that confirm or suggest solutions, respectively;

Table 1 Nugget types at the ND subtask

	Customer	Helpdesk
Not a nugget	CNaN	HNaN
Regular	CNUG	HNUG
Goal	CNUG*	HNUG*
Trigger	CNUG0	

4. CNUG0: Customers' trigger nuggets that start a conversation with specific problem descriptions. Table 1 presents Nugget types at the DialEval-2 Nugget Detection (ND) subtask.

The remainder of this paper is structured as follows. Section 2 describes related studies, whereas Section 3 elaborates our methodology. Section 4 displays the findings: the evaluation metrics are given in Section 4.1, the hyperparameter settings that we selected are explained in Section 4.2, and all the data are displayed in Section 4.3. Finally, we present the conclusion in Section 5.

2. Related Work

Machine translation reviews by humans are thorough but costly, taking months to complete and requiring human labor that cannot be repeated. Automatic assessment measures are utilized in the assessment of simulated dialog corpora by human judgements, according to previous research [10]. The accuracy of these automated metrics, on the other hand, has yet to be demonstrated. The challenge of automatically evaluating such systems remains a difficult one [11]. Traditional reference-based measures, such as BLEU, are unhelpful given that numerous appropriate replies that have no common terms with reference responses may be provided for a given context [12]. As a result, numerous automatic conversation quality evaluation models have been presented by researchers to tackle the difficulties more efficiently and generate more effective outcomes.

Some feasible approaches to dialogue evaluation are described in the literature, for example, [13] the Hidden Markov Model approach for modeling dialogue acts in conversational speech [14] and a model based on recurrent neural networks and convolutional neural networks that incorporates short texts and receives the results on different datasets for dialog act prediction. As previously noted, they present an assessment approach for conversation systems that do not require manual involvement. We suggested a method of objective automated evaluation to decrease annotation mistakes and personal prejudice. We evaluated dialogue quality by using the LSTM-based baseline model as the input of attention layers following dialog embedding and nugget identification.

2.1 Bi-LSTM Baseline

When categorizing texts, the LSTM model may be used to overcome the problem that typical machine learning methods struggle to extract high-level semantics from. This model takes a text sequence matrix made of pre-trained distributed word vectors as input and then uses its unique memory structure to extract feature expressions comprising context information. A forward LSTM layer and a backward LSTM layer are combined in BiLSTM. The BiLSTM model uses topic information to learn the sensitive representation of a polysemous word in a specific situation. Topic information is created via Latent Dirichlet Allocation (LDA) and topic modeling. The model can automatically capture the meaning of the polysemous word and lengthy sequence information owing to the topic information-based BiLSTM network [16].Using this structure, BiLSTM may overcome the problem of LSTM's incapability to encode information from the rear to the front. Figure 1 shows the construction of the BiLSTM model [18].

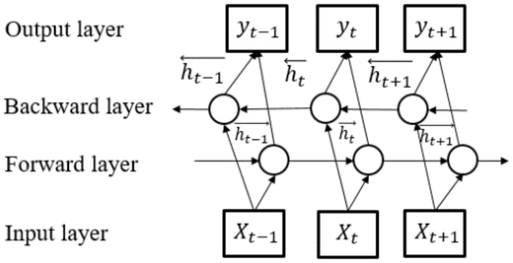

Figure 1 Architecture of the BiLSTM model

3. Proposed Approaches

Learning via trial and error is the most basic strategy for identifying current studies. This paper aim to describe and show the trial-and-error technique of learning and provide several of the findings produced. To manage the quality and speed of transfer learning, we leveraged pre-trained models from HuggingFace's Transformers and fastai. Extra specifics are available in the source papers for the models. We provided the suggested models and training methodologies for multilabel and multiclass classifications of the DQ and ND subtasks in this section.

3.1 Selected Models

The usage of BERT's pre-trained language models (PLMs) in NLP has had considerable success in a number of domains [1]. PLMs do not use any supervised data, but they aid in the creation of considerable performance increases for numerous NLP tasks, which has led to their current popularity [2].

BERT is a transformer-based language model meant to learn deep bidirectional representations by pre-training on a large unsupervised dataset. It may be fine-tuned to a variety of standards and produces cutting-edge outcomes. Mask Language Model (MLM) and Next Sentence Prediction (NSP) are the two subtasks that make up this task. They are utilized as pretraining schemes in the case of BERT. Pre-training is performed in the BERT by reducing the loss of the MLM and NSP tasks simultaneously [3]. MLM is a technique for masking certain words in an input sequence and then predicting the masked word based on context; NSP is a technique for predicting whether a sentence pair is continuous.

The proposed model predicts text sequences using the pre-trained XLM-RoBERTa model for text sequence categorization. All pre-trained models were the base versions to meet our research aim of conducting trials quickly. The interactions were divided into turns, with each turn containing one or more statements from either a client or a helpdesk representative. The XLM-RoBERTa model supports Chinese and English languages. XLM-RoBERTa retains subword token letter cases and integrates and improves approaches from cross-lingual language model pretraining schemes and a robustly optimized BERT pretraining methodology. RoBERTa changes the critical hyperparameters for optimization. RoBERTa uses a byte-level Byte Pair Encoding (BPE) tokenizer and dynamically adjusts the masking pattern used to the training data in terms of tokenization [4].

3.2 Tokenization Tricks

We not only used XLM-special RoBERTa's tokens for the beginning (<s>), conclusion (</s>), and separator (</s> </s>) of sentences but also tweaked a couple of tokens in the fastai convention of “xx” prefix which offers context to better depict the structure of the conversation. The special tokens are as follows:

xxlen: length of the dialogue in turns

xxtrn: position of each turn of the dialogue

xxsdr: differentiates whether the sender is Customer or Helpdesk

xxlen and xxtrn are special tokens that represent the duration of the discussion in turns and the position of each turn of the dialogue, respectively. The numbers adjacent to them indicate various characteristics of turns. xxsdr uses the same method to determine whether the sender is Customer or Helpdesk. The nugget type is usually always CNUG0 when the context of a turn contains “xxtrn _1 xxsdr _customer.”

The dialogue for the DQ subtask may be tokenized in a similar way, with xxlen being beneficial for specific quality scores, such as time/turns spent on resolving a problem:

xxlen 3 <s> xxtrn 1 xxsdr customer Since there is no lunar calendar in the phone's calendar, I installed a new calendar application, but the date displayed is different. @Smartisan Customer Service </s> </s> xxtrn 2 xxsdr helpdesk Hello, the problem of not displaying the lunar calendar in the view of the built-in calendar month will be updated in the later version. The external version of the calendar cannot display the dynamic icon at present. </s> </s> xxtrn 3 xxsdr customer I see. Thank you! </s>

Although we did not use the default tokenizer fastai, the “xx” prefix in the fastai convention identified special context tokens. Fastai tokenizes English texts using SpaCy by default, inserting special tokens before uncapitalized or previously repeated words/characters. If we use fastai’s default tokenization for both title case and word duplication, we obtain the following results. Given that pretrained transformer models are unaware of particular context tokens, we must consider whether they can still help in fine-tuning a certain job, regardless of how lossless the conversion is. Repetition and capitalization may be key indicators if the objective was to analyze the emotion of a sentence. However, how the repeating word or character may aid semantically or syntactically, especially given that XLM-RoBERTa already retains subword token letter cases, is difficult to conceive. We did not use them for the DialEval-2 task based on previous observations [4].

3.3 Fine-tuning Techniques

We also used different fine-tuning approaches for the learning rate and optimizer to efficiently assess multiple pre-trained models. We suggested progressive unfreezing to fine-tune the classifier in addition to discriminative fine-tuning and triangular learning rates. According to our preliminary testing, discriminative fine-tuning and fastai’s version of one-cycle policy function well. However, graduate unfreezing has a minimal effect, which is consistent with that of other research.

3.3.1 Discriminative Fine-tuning

Discriminative sentence modeling captures the meanings of sentences and categorizes them in accordance with particular criteria (e.g., sentiment) [5]. It has gained considerable interest in the NLP community because it is connected to numerous tasks of interest [6]. Varying layers should be fine-tuned to different extents because they capture different sorts of information [7].

As a result, we proposed discriminative fine-tuning, which allows us to adjust each layer with varied learning rates, effectively creating pre-trained models. We may adjust each layer with different learning rates using discriminative fine-tuning. We demonstrated that our technique enables the effective transfer learning and performs thorough analysis using state-of-the-art technology. We concentrated on fine-tuning to address the issue of manual labeling that they require.

3.3.2 One-cycle Policy

Different momentum has a negligible influence on the validation accuracy of fastai, according to one cycle policy [8]. The network’s performance will be improved by selecting an appropriate learning rate; a slight learning rate causes overfitting, whereas a high learning rate produces divergence. As a result, a “one-cycle” learning rate strategy may be adopted to address the problem.

In comparison with a constant learning rate, a one-cycle learning rate reduces the likelihood of overfitting and allows models to learn more rapidly and efficiently [9]. A cycle is a collection of epochs that have the same hyperparameter strategy, particularly for learning rates and momentums. A policy of cyclical learning rates, in which the step size is regularly increased and then decreased, may converge quicker and better when training a deep neural network with stochastic gradient descent or comparable methods [4].

3.4 Research framework

For DialEval-2’s DQ and ND subtasks, we presented the model definitions and training methodologies. Transfer learning and transformer-based language models play major roles in the present NLP research community, according to a four-phase study. For transfer learning, we first selected the transformer model.

The tokenized input sentence was then passed to the XLM-RoBERTa model in the second step. This method cannot only convert an unstructured text into a numerical data structure suited for machine learning but can also be used by a computer to initiate meaningful actions and replies. By evaluating the sequence of words, tokenization aids in deciphering the meaning of the text.

Finally, we provided fine-tuning approaches for the transformer model on conversation tasks. Figure 2 shows that we developed a broad framework to test and demonstrate the XLM-RoBERTa model’s performance on conversational texts to test and illustrate the concept.

4. RESULTS

Our team submitted for the Dialogue Quality (DQ) subtask and the Nugget Detection (ND) subtask. The Task Organizers for DialEval-2 provided three files: “Train set,” “Dev set,” and “Test set.” To evaluate the performance of models, we used the “Test set”, and provided the evaluation scores for our top models in it. “A-score,”

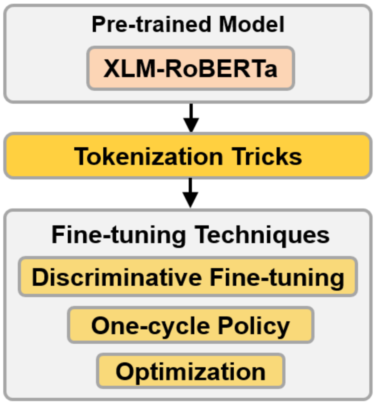

Figure 2 Proposed research architecture of IMNTPU at NTCIR-16 DialEval-2

Table 2 Hyperparameter Settings

Hyperparameter	Value
Learning rate	2e-6, 1e-6, 5e-7, 1e-7
Batch size	128
Optimizer	Adam

"S-score," and "E-score" refer to the average assessment scores for each conversation. Our runs were the ensemble models' prediction results from the test set. We trained a model that had a 5% reduction in cost.

4.1 Evaluation metrics

We applied the evaluation metrics from the DialEval-2 Task. This paper employed bin-by-bin and cross-bin measurements to evaluate the model's performance. Given the non-nominal nature of the DQ subtask classes. We used two cross-bin metrics: the Normalised Match Distance (NMD) and Root Symmetric Normalised Order-aware Divergence (RSNOD) [20]. The classes in the ND subtask are nominal. Therefore, bin-by-bin metrics are more appropriate than those in the DQ subtask. The Root Normalised Sum of Squares (RNSS) and Jensen Shannon Divergence (JSD) are the metrics employed in the ND subtask. For clarity, the distance scores (JSD, RNSS, RSNOD, and NMD) were transformed using -log(). As a result, the more converted scores there are, the more effective the model is.

4.2 Hyperparameter settings

The cycle schemes and their max_lr discriminative learning rates are important hyper parameters, but they all have the same reduction rate: the bottom bound is always max_lr/1000, and each cycle comprises one epoch. Table 2 lists the experiment's hyperparameters.

4.3 IMNTPU NTCIR-16 DialEval-2 Results

Table 3 presents the results of IMNTPU at NTCIR-16 DialEval-2 Chinese Dialogue Quality. In the Chinese DQ was in the S-score (customer satisfaction with the dialogue), and in NMD indication, IMNTPU-run0 test set was outperformer than Baseline-run0 test set and IMNTPU-run0 development set. In the A-score (task completed: Has the problem been solved?) and E-score (dialogue effectiveness: Does the speaker interact effectively to solve the problem effectively?), and in RSNOD and NMD indication, IMNTPU-run0 development set was outperformer than Baseline-run0 test set and IMNTPU-run0 test set.

Table 4 presents the results of IMNTPU at NTCIR-16 DialEval-2 English Dialogue Quality. In the English DQ, IMNTPU-run0 development set was outperformer than Baseline-run0 test set and IMNTPU-run0 test set.

Table 5 presents the results of IMNTPU at NTCIR-16 DialEval-2 Chinese Nugget Detection. In the JSD and RNSS indicators, Baseline-run0 test set was outperformer than IMNTPU-run0 development set.

Table 6 shows the results of MNTPU at NTCIR-16 DialEval-2 English Nugget Detection. In the JSD and RNSS indicators, IMNTPU-run0 test set outperformed the Baseline-run0 test set and IMNTPU-run0 development set.

5. Conclusion

In this paper, we provided XLM-RoBERTa, a fine-tuning approach for transformer models for conversational texts. We employed XLM-RoBERTa to extract texts between conversations to learn key information and to cope with text categorization challenges in the DQ and ND subtasks. We show that XLM-RoBERTa works well with Chinese and English data sets.

Table 3 Results of IMNTPU at NTCIR-16 DialEval-2 Chinese Dialogue Quality

NTCIR-16 DialEval-2 Chinese Dialogue Quality (DQ) Test set						
	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	0.2479	0.1618	0.2032	**0.1315**	0.1860	0.1427
Baseline-run0	0.2301	0.1772	**0.1998**	0.1523	0.1854	0.1579
NTCIR-16 DialEval-2 Chinese Dialogue Quality (DQ) Development set						
	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	**0.2262**	**0.1495**	0.2076	0.1344	**0.1694**	**0.1251**

Table 4 Results of IMNTPU at NTCIR-16 DialEval-2 English Dialogue Quality

NTCIR-16 DialEval-2 English Dialogue Quality (DQ) Test set						
	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	0.2535	0.1654	0.2020	0.1312	0.1826	0.1400
Baseline-run0	0.2321	0.1780	0.1986	0.1467	0.1745	0.1431
NTCIR-16 DialEval-2 English Dialogue Quality (DQ) Development set						
	A-score		S-score		E-score	
Model	RSNOD	NMD	RSNOD	NMD	RSNOD	NMD
IMNTPU-run0	**0.2102**	**0.1397**	**0.1879**	**0.1216**	**0.1617**	**0.1184**

The important contribution of this study is that we proposed two critical elements, namely, tokenization procedures and fine-tuning approaches, to improve the dialogue quality and nugget recognition subtasks in dialogue analysis.

6. ACKNOWLEDGMENT

This research was supported in part by the Ministry of Science and Technology (MOST), Taiwan under grant number 110-2410-H-305-013-MY2, and National Taipei University (NTPU) under grant number 110-NTPU-ORDA-F-001, 111-NTPU-ORDA-F-001, and 111-NTPU-ORDA-F-003.

7. REFERENCES

[1] Zeng, Z., Kato, S., Sakai, T., & Kang, I. (2020). Overview of the NTCIR-15 dialogue evaluation (DialEval-1) task. Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies.

[2] Zeng, Z., Kato, S., & Sakai, T. (2019). Overview of the NTCIR-14 short text conversation task: Dialogue quality and nugget detection subtasks. Proceedings of NTCIR-14, 289-315.

[3] Tao, S. & Sakai, T. (2022). Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task, Proceedings of NTCIR-16.

[4] 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.

[5] Lin, N., Fu, Y., Chen, C., Yang, Z., & Jiang, S. (2021). LaoPLM: Pre-trained Language Models for Lao. arXiv preprint arXiv:2110.05896.

Table 5 Results of IMNTPU at NTCIR-16 DialEval-2 Chinese Nugget Detection

NTCIR-16 DialEval-2 Chinese Nugget Detection (ND) Test set		
Model	JSD	RNSS
Baseline-run0	**0.0585**	**0.1651**
NTCIR-16 DialEval-2 Chinese Nugget Detection (ND) Development set		
Model	JSD	RNSS
IMNTPU-run0	2.0670	1.3969

Table 6 Results of MNTPU at NTCIR-16 DialEval-2 English Nugget Detection

NTCIR-16 DialEval-2 English Nugget Detection (ND) Test set		
Model	JSD	RNSS
IMNTPU-run0	**0.0601**	**0.1574**
Baseline-run0	0.0625	0.1722
NTCIR-16 DialEval-2 English Nugget Detection (ND) Development set		
Model	JSD	RNSS
IMNTPU-run0	0.0752	0.1727

[6] Zöllner, J., Sperfeld, K., Wick, C., & Labahn, R. (2021). Optimizing small BERTs trained for German NER. Information, 12(11), 443.

[7] Jiang, M. T. J., Wu, Y. C., Shaw, S. R., Gu, Z. X., Huang, Y. C., Day, M. Y., ... & Chiu, C. H. IMTKU Multi-Turn Dialogue System Evaluation at the NTCIR-15 DialEval-1 Dialogue Quality and Nugget Detection. Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-15), Tokyo Japan, December 8-11, 2020, pp. 68-74.

[8] Mou, L., Peng, H., Li, G., Xu, Y., Zhang, L., & Jin, Z. (2015). Discriminative neural sentence modeling by tree-based convolution. arXiv preprint arXiv:1504.01106.

[9] Wang, Y., Luo, J., Hao, S., Xu, H., Shin, A. Y., Jin, B., ... & Ling, X. B. (2015). NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. International journal of medical informatics, 84(12), 1039-1047.

[10] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. Advances in neural information processing systems, 27.

[11] Zhou, Q., Zhang, Y., Li, P., Liu, X., Yang, J., Wang, R., & Huang, R. (2020). DaSGD: Squeezing SGD Parallelization Performance in Distributed Training Using Delayed Averaging. arXiv preprint arXiv:2006.00441.

[12] Koay, H. V., Chuah, J. H., Chow, C. O., Chang, Y. L., & Rudrusamy, B. (2021). Optimally-weighted image-pose approach (OWIPA) for distracted driver detection and classification. Sensors, 21(14), 4837.

[13] Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

[14] Ai, H., & Litman, D. (2008). Assessing dialog system user simulation evaluation measures using human judges. In Proceedings of ACL-08: HLT (pp. 622-629).

[15] Ghazarian, S., Wei, J., Galstyan, A., & Peng, N. (2019). Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. arXiv preprint arXiv:1904.10635.

[16] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics, 26(3), 339-373.

[17] Lee, J. Y., & Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827.

[18] Zhang, J., Zi, L., Hou, Y., Deng, D., Jiang, W., & Wang, M. (2020). A C-BiLSTM approach to classify construction accident reports. Applied Sciences, 10(17), 5754.

[19] Huang, Y., Jiang, Y., Hasan, T., Jiang, Q., & Li, C. (2018). A topic BiLSTM model for sentiment classification. In proceedings of the 2nd international conference on innovation in artificial intelligence (pp. 143-147).

[20] Sakai, T. (2018). Comparing two binned probability distributions for information access evaluation. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 1073-1076).