Exploring RoBERTa model for cross-domain suggestion detection in online reviews

Anuradha Nandula[1,2], Panuganti Vijayapal Reddy[3]
[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India
[2]Faculty of Informatics, Osmania University, Hyderabad, India
[3]Department of Computer Science and Engineering, HoD, Matrusri Engineering College, Hyderabad, India

Article Info

Article history:

Received Feb 21, 2024
Revised Apr 15, 2024
Accepted May 6, 2024

Keywords:

Cross-domain generalization
Natural language processing
Suggestion mining
Transfer learning
Transformers

ABSTRACT

Detecting suggestions in online review requires contextual understanding of review text, which is an important real-world application of natural language processing. Given the disparate text domains found in product reviews, a common strategy involves fine-tuning bidirectional encoder representations from transformers (BERT) models using reviews from various domains. However, there hasn't been an empirical examination of how BERT models behave across different domains in tasks related to detecting suggestion sentences from online reviews. In this study, we explore BERT models for suggestion classification that have been fine-tuned using single-domain and cross-domain Amazon review datasets. Our results indicate that while single-domain models achieved slightly better performance within their respective domains compared to cross-domain models, the latter outperformed single-domain models when evaluated on cross-domain data. This was also observed for single-domain data not used for fine-tuning the single-domain model and on average across all tests. Although fine-tuning single-domain models can lead to minor accuracy improvements, employing multi-domain models that perform well across domains can help in cold start problems and reduce annotation costs.

Corresponding Author:

Anuradha Nandula
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Aziz Nagar, Hyderabad, 500075, India
Email: evanianuradha2002@gmail.com

1. INTRODUCTION

Online services such as Suggestion Forums and TripAdvisor have review systems where users can provide comments and suggestions for products, offering valuable information to both buyers and companies as shown in Figure 1. Analysing the large volume of review data manually is impractical, so companies use data science and machine learning to extract knowledge efficiently [1]. An important task in review understanding is suggestion sentences classification, akin to sentiment analysis [2], where the text of a review is used to determine the corresponding sentiment. The task of suggestion mining from online reviews and forums, termed as Task 9 [3] was introduced at the International Workshop on Semantic Evaluation 2019 Suggestion sentence detection aids in identifying the suggestion conveyed by the customer along with the sentiment expressed in the text, addressing potential improvements for a product or a service [4]-[6] have used Ensemble classifier (Logistic, GRU, FFA, CNN), with BERT and achieved an F1 score of 85%. Goldberg *et al.* [7] achieved a score of 81% with Ensemble classifier CNN and LSTM. BERT. Wicaksono and Myaeng [8] reported a score of 64.8% by implementing Ensemble classifier with Attention sentence encoder, BERT and CNN based word encoder technique. In today's digital age, a vast volume of textual content floods online platforms such as reviews and discussion forums. These platforms serve as invaluable resources for gauging public opinions on products or services [9]. The insights gleaned from such sources not only benefit readers but also provide crucial suggestions for stakeholders [10], [11]. These suggestions can be instrumental in enhancing product quality or offering pertinent recommendations [12]. However, sifting through numerous reviews or comments to pinpoint useful suggestions demands considerable time and effort. Additionally, the unstructured nature of online texts complicates the task further [11]. Consequently, the automated extraction of suggestions from such texts poses a significant and challenging endeavor [10]. Suggestion mining can adapt methods like CNN-based models [13] or RNN-based models [14] used for traditional sentiment classification systems.

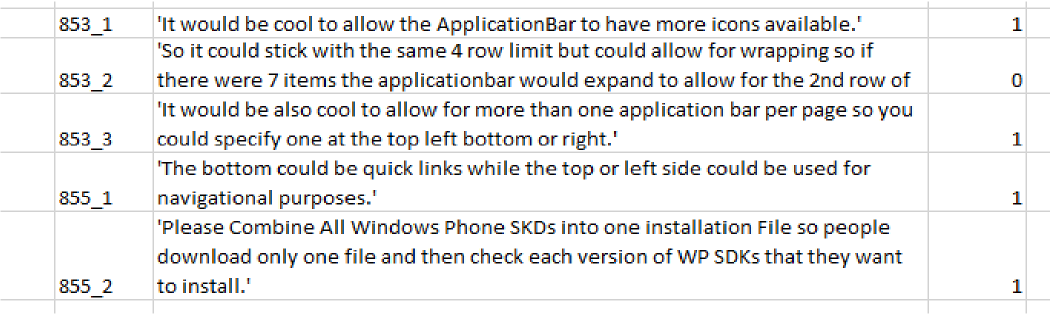
Markov and Clergerie [15] suggested that cross-domain sentiment analysis needs huge training data to fine tune BERT models. Work done showed fine-tuned BERT models are more efficient over state-of-the-art domain adaptation methods [16], [17] Performed a comprehensive empirical investigation into single- and multisource domain adaptation, both unsupervised and supervised, as well as the development of a sentiment-sensitive thesaurus that accurately captures words conveying similar sentiments.

Pecar *et al.* [18] on domain-specific knowledge distillation for conversational commerce is closely related to this work. Knowledge distillation addresses the issue of large and computationally expensive models being too costly to be deployed through a "teacher-student" training approach, in which a trained "teacher" model facilitates the training of a smaller "student" model [18], [19] show knowledge distillation can both reduce model size and adapt models to a specific domain. In this work, we show that BERT base models are capable of adapting to a domain through fine-tuning on domain-specific, instead of knowledge distillation. However, we also show that, compared to single-domain models, cross-domain models, though incurring a slight accuracy penalty, still retain high performance towards a broad range of data. We also propose multi-domain models as an alternative solution to cost of computation. Multidomain models and knowledge distillation could be used in tandem to further reduce computational resources, using knowledge distillation to train a "student" model with a broad spectrum of data domains, creating a model that is both small in size and applicable to multiple domains.

Bidirectional encoder representations from transformers (BERT) models [20], renowned for their prowess in natural language understanding, possess the capability to undergo fine-tuning to classify suggestion sentences in online reviews [6]. BERT leverages a transformer architecture, which enables the simultaneous processing of entire sequences of text, thereby capturing intricate contextual dependencies among words [21]. Comprising stacked transformer encoder layers equipped with self-attention mechanisms, BERT undergoes pre-training on two primary tasks: masked language modelling and next sentence prediction. Following a phase of general pre-training, BERT is fine-tuned for specific downstream tasks by incorporating task-specific layers and adjusting neural network parameters based on labeled data [22], with suggestion classification being an important downstream task. The presence of diverse domains poses a significant challenge for extracting suggestion reviews. Platforms such as Amazon encompass a wide array of product categories, necessitating models to analyse reviews specific to each domain separately. Despite endeavours to devise cross-domain models, it remains uncertain whether a single model can efficiently perform across all domains [23], [24]. In addressing this challenge, BERT can undergo fine-tuning using domain-specific training sets or a combination of training sets spanning various domains [25]. Fine-tuning with data tailored to specific domains enables models to concentrate on predicting relevant information encountered within their domain, thus avoiding the counterproductive task of predicting information irrelevant to the domain at hand. However, training distinct domain-specific models involves extra resource consumption compared to training a unified model for multiple domains, under the assumption that the multi-domain model exhibits satisfactory performance across diverse domains. To comprehend the trade-offs between a domain-specific versus a generalized approach for multiple domains, conducting an empirical investigation analysing RoBERTa’s cross-domain behaviours in the realm of review comprehension becomes crucial.

In practical scenarios, the domain-specific nature of suggestion mining frequently poses challenges when there isn't enough data available for a particular domain to adequately train a high-performing single-domain model. This issue is commonly addressed through domain adaptation, where a model is trained on data from a relevant secondary domain and then applied to the primary, data-deficient domain. Previous studies have demonstrated that fine-tuned BERT models outperform state-of-the-art domain adaptation methods [15]. Our research similarly highlights the ability of cross-domain models to adapt to different domains.

This research conducted a comprehensive empirical analysis of RoBERTa's cross-domain behaviors in Suggestion classification, contributing novel insights into model performance and domain adaptation strategies. The subsequent sections will delve into the proposed model, the methodology, experimental results, and implications of the study, demonstrating its relevance in advancing the field of suggestion mining and informing practical applications in consumer feedback analysis.



853_1	'It would be cool to allow the ApplicationBar to have more icons available.'	1
853_2	'So it could stick with the same 4 row limit but could allow for wrapping so if there were 7 items the applicationbar would expand to allow for the 2nd row of	0
853_3	'It would be also cool to allow for more than one application bar per page so you could specify one at the top left bottom or right.'	1
855_1	'The bottom could be quick links while the top or left side could be used for navigational purposes.'	1
855_2	'Please Combine All Windows Phone SKDs into one installation File so people download only one file and then check each version of WP SDKs that they want to install.'	1

Figure 1. Dataset example

2. THE PROPOSED RoBERTa MODEL ARCHITECTURE

2.1. RoBERTa model

Figure 2 depicts the RoBERTa model showing a sequential process, beginning with the tokenization of input sentences containing s words into a collection of n tokens using the BERT mechanism. Each token undergoes a transformation process, resulting in a 768-dimensional numeric vector, which serves as a contextualized representation for each character. The output from BERT is organized into a n×768 matrix. Next, a convolutional layer operates on this matrix, partitioning it and generating an output vector with dimensions n−k+1, where k represents the kernel size. Subsequently, a pooling layer employs max pooling to extract salient features from each output vector, producing a vector of size n−k+1p, where p denotes the pool size. These pooled vectors are then transformed into s distinct categorization categories, where s=2 in this study, encompassing suggestion and non-suggestion classes. Finally, the Linear activation function is applied to generate output values indicating the probability of an input sentence belonging to its respective suggestion class, as 0 to 1.

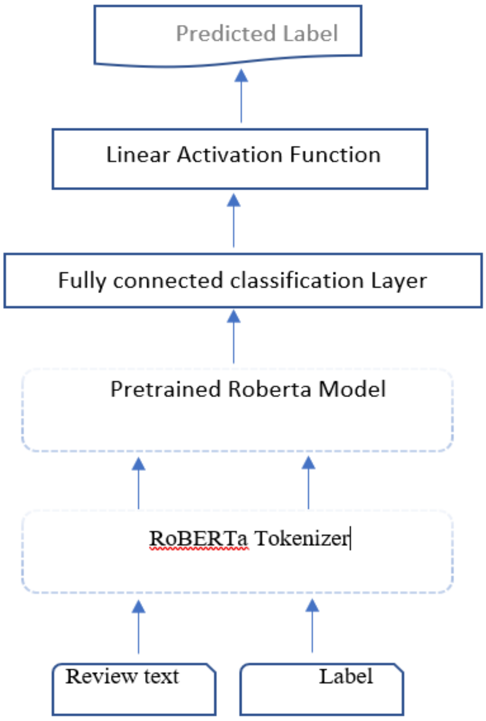


Figure 2. Fine tuning of the RoBERTa mode

3. METHOD

This section describes the text representation method used and its application in deep learning models for suggestion mining problems of online reviews in two different domains. This study evaluated the performance of the deep learning models with RoBERTa as a text representation method. Before doing

Suggestion classification using deep learning models, the textual data input is transformed into a numerical representation using RoBERTa tokenizer. An assessment is conducted after obtaining the textual representation and then fine tuning RoBERTa pretrained model for Indomain and Cross-Domain scenarios with electronics and hotel review datasets.

3.1. Robustly optimized BERT pretraining approach

Figure 3 shows RoBERTa by [26], is an enhancement of Google's BERT model from 2018. It modifies key hyperparameters, eliminating the next-sentence pretraining objective, and utilizes larger mini-batches and learning rates during training. RoBERTa shares the architecture of BERT but employs a byte-level byte-pair encoding (BPE) tokenizer, similar to GPT-2, and adopts a distinct pretraining approach. Unlike BERT, RoBERTa does not utilize token_type_ids, eliminating the need to specify segment tokens; segments are separated using the separation token tokenizer.sep_token (or </s>). It incorporates improved pretraining techniques similar to BERT, including dynamic masking, where tokens are masked differently in each epoch, contrary to BERT's fixed masking. Furthermore, RoBERTa combines sentences to reach 512 tokens, enabling sentences to span multiple documents, and trains with larger batches. Also, it utilizes BPE with bytes as subunits instead of characters due to the presence of Unicode characters.

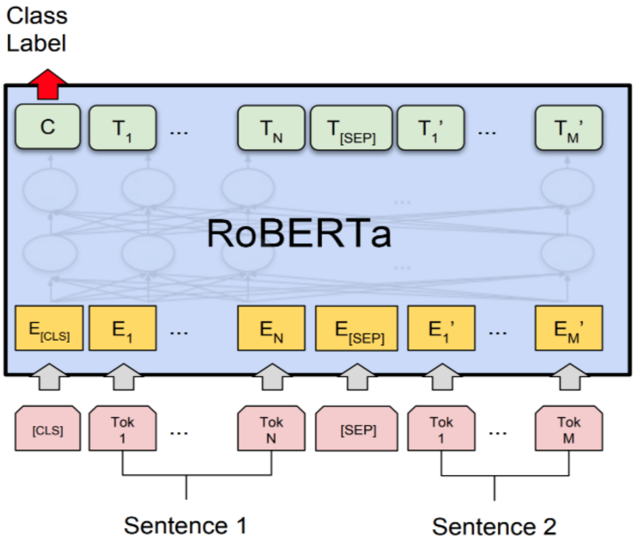


Figure 3. RoBERTa pretrained model

4. RESULTS AND DISCUSSION

4.1. Datasets

Two review datasets provided by [13], software forum and trip advisor with mutually dissimilar domains were chosen to fine-tune RoBERTa Model. Statistics of Electronics review data (Electronics), Hotel review data (Hotel) datasets are as shown in Table 1. Our pipeline process from data sampling to model evaluation is shown in Figure 4. Data for model training dataset Software reviews was random sampled to overcome class imbalance. Since the original Electronics domain data is significantly larger than the other datasets, 1,600 reviews were sampled from each domain for model fine-tuning, remedying domain imbalance. To reduce class imbalance, for Electronics domain, 1,600 training reviews consisting of 800 suggestion and 800 non suggestion reviews were randomly sampled. Validation set from hotel review is taken as training set for TripAdvisor domain. The test datasets provided for each domain are used for testing the model.

Table 1. Statistics of datasets

	Electronics			Hotel		
	Train	Eval	Test	Train	Eval	Test
Suggestions	2,085	296	87	-----	404	348
Non-Suggestions	6,415	296	746	-----	404	476
ALL	8,500	592	833	-----	808	824

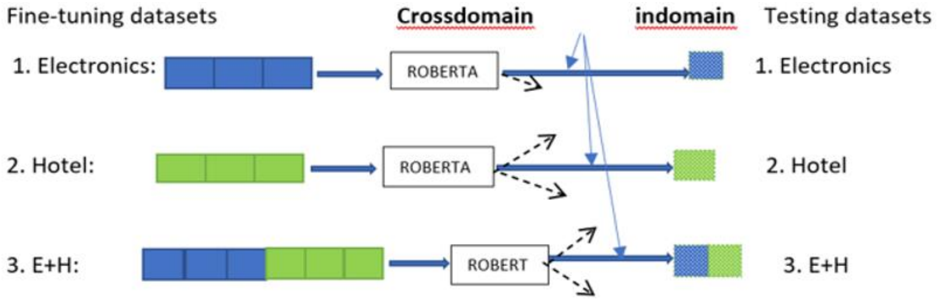


Figure 4. Two datasets with varying combinations of domains were made, all of the same size. RoBERTa models were trained upon all datasets and cross-domain tested

4.2. Preparing text data for RoBERTa

A RoBERTa tokenizer is used to ensure data is according to the format the model can understand. Firstly, the text is broken down into tokens using the WorldPiece model, with special tokens like [CLS] and [SEP] added for context. Then, to ensure uniformity, padding is applied by adding [PAD] tokens until each sentence reaches a set length of 128 tokens. Next, tokens are converted into unique integers using RoBERTa's vocabulary, which contains 30,522 token-integer pairs. This process, called numericization, assigns a token ID to each token, enabling RoBERTa to process them into numeric vectors for analysis.

4.3. Suggestion classification experiments

The data in this study is categorized into two distinct domains: electronics domain and Hotel domain. The balanced electronics training dataset and Hotel review training dataset are combined to form the cross-domain dataset. Processing resulted in a final training corpus of 3,200 reviews, divided evenly among the two domains. One two-domain datasets with 1,600 train examples each were formed through taking 800 examples from each combination of two single domain datasets. The series of Experiments conducted and the setup is shown in Figure 5.

The pre-trained BERT-base-cased model, RoBERTa, was utilized as the basis for fine-tuning. Multiple sets of fine-tuned models, each with different hyperparameters, were trained for individual datasets to identify the optimal model performance tailored to each dataset. Various combinations of two learning rates (1e-5, 2e-5) and two batch sizes (8, 16) were employed, alongside a training duration of 2 epochs. It was observed that higher learning rates and batch sizes led to enhanced model performance.

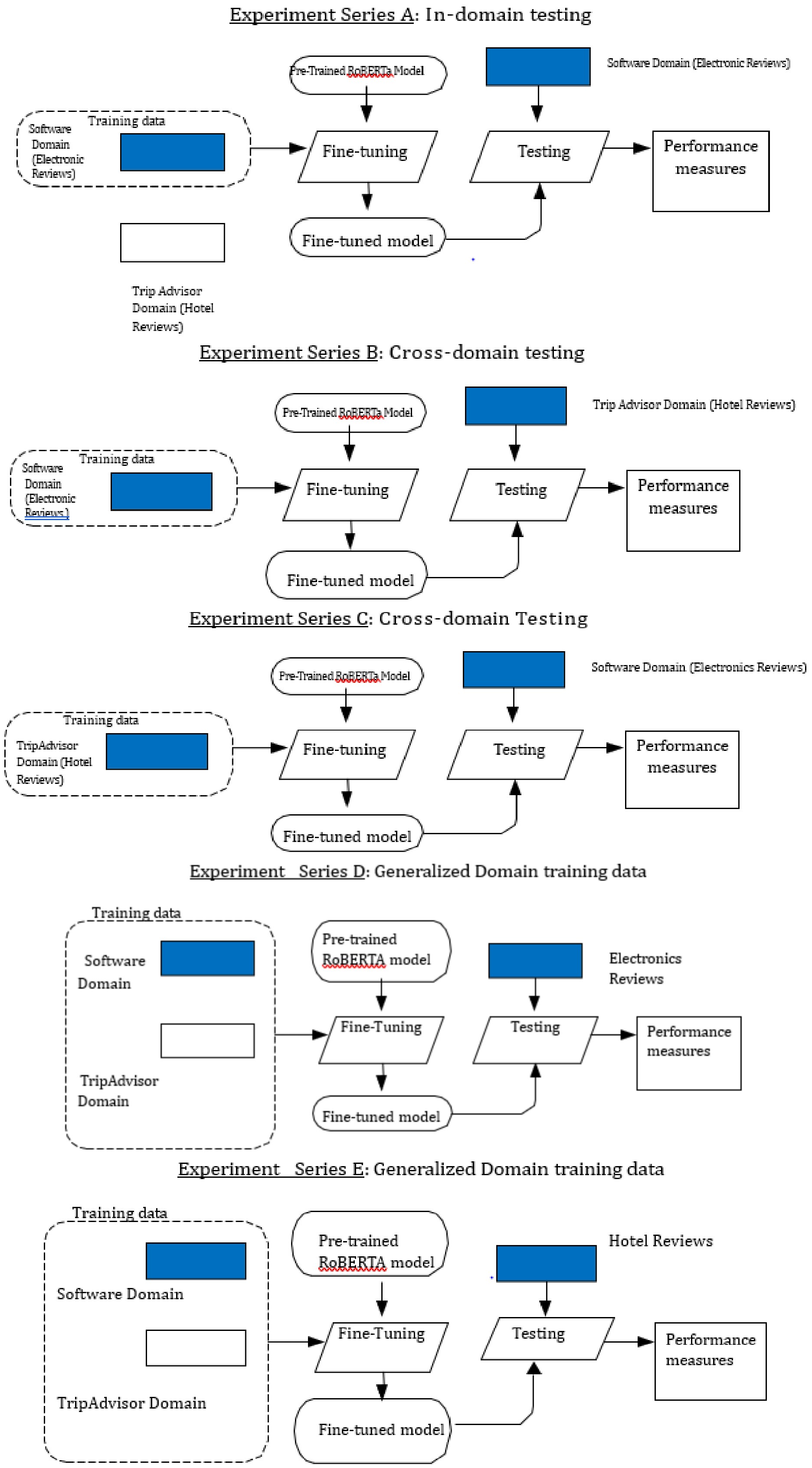
After selecting the best model from each batch, each of the three models was evaluated against the test split of every dataset. The F1 score metric was used to determine the optimal model in each dataset batch and evaluate model cross-domain testing. As F1 score is proportional to the difference between true and predicted values, a higher F1 score indicates higher performance. Models were trained and tested on a GPU server of provided by google Colaboratory.

4.4. Model performance evaluation

The results of testing each model against each dataset are shown in Table 2. Each row of the table represents one model being tested on all 3 datasets; each column represents one dataset having all 3 models evaluated upon it. Performance of the models was measured using F1 score. A higher F1 score indicates higher model accuracy. Bold table values represent the highest performance achieved in each column. Considering the horizontal rows, smaller F1 score indicates more difficult dataset. Considering the vertical column, higher F1 values mean a better performing model. Within-domain performance is indicated on the table through underlining. The bottom cells contain the F1 scores of cross-domain model tested on single-domain data, with right bottom cell contain the F1 scores of cross-domain data are shown.

Table 2. F1 scores of models trained on specific dataset training pools (rows) and predicting on testing datasets (columns)

	Test on datasets (F1 scores)		
Train on models	Electronics	Hotel	Electronics + Hotel(E+H)
Electronics	**86.5**	53.1	73
Hotel	54	**92**	74
Electronics + Hotel(E+H)	81	91	81.7



Experiment Series A: In-domain testing

Pre-Trained RoBERTa Model
Software Domain (Electronic Reviews)
Training data
Software Domain (Electronic Reviews)
Fine-tuning
Testing
Performance measures
Fine-tuned model
Trip Advisor Domain (Hotel Reviews)

Experiment Series B: Cross-domain testing

Pre-Trained RoBERTa Model
Trip Advisor Domain (Hotel Reviews)
Training data
Software Domain (Electronic Reviews)
Fine-tuning
Testing
Performance measures
Fine-tuned model

Experiment Series C: Cross-domain Testing

Pre-Trained RoBERTa Model
Software Domain (Electronics Reviews)
Training data
TripAdvisor Domain (Hotel Reviews)
Fine-tuning
Testing
Performance measures
Fine-tuned model

Experiment Series D: Generalized Domain training data

Experiment Series E: Generalized Domain training data

Figure 5. Experiment schemes. Series A–fine-tuning on in-domain corpus; series B–C– fine-tuning and cross domain testing; Series D–E fine-tuning on Generalized-domain corpora, and testing the obtained model using Electronics and Hotel reviews respectively

4.5. Discussion

The main objective of this study was to assess the advantages and disadvantages of employing specific versus general model domain scopes in RoBERTa based suggestion review classification. The disparity between the model's F1 score was relatively minor, fluctuating around 5-10%. The F1 scores of electronic model (In domain and cross domain) are 86.5% and 54% could be attributed to the nature of classifying electronics reviews, which may heavily rely on topic-specific vocabulary. While F1 score of Hotel model is 92% for in-domain and 74% for cross-domain model. This can be attributed to the fact that reviews taken from hotel reviews may encompass different verbs, the adverbs or modal verbs used to express suggestions are often similar. The BERT models likely identified the significance of suggestion-infused verbs such as "need" try "," which are applicable across various product categories. Consequently, the Pretrained models may have prioritized more general keywords over topic-specific terms and descriptions. Also, it is important to note that suggestion mining still necessitates a certain level of domain understanding, as evidenced by our findings.

5. CONCLUSION

In this study, we've presented an instance of how datasets formulated in varied contexts can be reconfigured and employed toward a shared aim. We illustrated a benefit in efficacy of pre-trained language models when utilized in domains dissimilar from those they were originally trained on. The pre-trained language models appear to possess the capability to grasp the fundamental concept behind a task to a greater degree. Moreover, training the models on cross-disciplinary data enhanced resilience and somewhat boosted performance. In summary, these findings indicate that pre-trained language models exhibit numerous desirable attributes when trained on diverse datasets, rendering them an optimal tool for effective utilization of existing online reviews datasets to bolster future exploration in innovative tasks.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," 2002, doi: 10.3115/1118693.1118704.

[2] B. Liu, "Sentiment analysis and opinion mining. morgan & claypool publishers 2012," *Google Scholar Google Scholar Digital Library Digital Library*, 2012.

[3] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, Mar. 2018, doi: 10.1002/widm.1253.

[4] A. Viswanathan, P. Venkatesh, B. Vasudevan, R. Balakrishnan, and L. Shastri, "Suggestion mining from customer reviews," *in AMCIS*, 2011.

[5] C. Brun and C. Hagège, "Suggestion mining: detecting suggestions for improvement in users' comments," *Research in Computing Science*, vol. 70, no. 1, pp. 199–209, Dec. 2013, doi: 10.13053/rcs-70-1-15.

[6] S. Negi and P. Buitelaar, "Towards the extraction of customer-to-customer suggestions from reviews," 2015, doi: 10.18653/v1/d15-1258.

[7] A. B. Goldberg, N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson, and X. Zhu, "May all your wishes come true: a study of wishes and how to recognize them," 2009, doi: 10.3115/1620754.1620793.

[8] A. F. Wicaksono and S.-H. Myaeng, "Mining advices from weblogs," Oct. 2012, doi: 10.1145/2396761.2398637.

[9] S. Moghaddam, "Beyond sentiment analysis: mining defects and improvements from customer feedback," in *Advances in Information Retrieval*, Springer International Publishing, 2015, pp. 400–410.

[10] S. Negi and P. Buitelaar, "Suggestion mining from opinionated text," in *Sentiment Analysis in Social Networks*, Elsevier, 2017, pp. 129–139.

[11] S. Negi and P. Buitelaar, "Inducing distant supervision in suggestion mining through part-of-speech embeddings," *arXiv preprint arXiv:1709.07403*, 2017.

[12] S. Negi, M. De Rijke, and P. Buitelaar, "Open domain suggestion mining: Problem definition and datasets," *arXiv preprint arXiv:1806.02179*, 2018.

[13] S. Negi, T. Daudert, and P. Buitelaar, "SemEval-2019 task 9: suggestion mining from online reviews and forums," 2019, doi: 10.18653/v1/s19-2151.

[14] T. Fatyanosa, A. H. A. M. Siagian, and M. Aritsugi, "DBMS-KU at SemEval-2019 Task 9: exploring machine learning approaches in classifying text as suggestion or non-suggestion," 2019, doi: 10.18653/v1/s19-2208.

[15] I. Markov and E. V. la Clergerie, "INRIA at SemEval-2019 Task 9: suggestion mining using SVM with handcrafted features," 2019, doi: 10.18653/v1/s19-2211.

[16] R. A. Potamias, A. Neofytou, and G. Siolas, "NTUA-ISLab at SemEval-2019 Task 9: mining suggestions in the wild," 2019, doi: 10.18653/v1/s19-2215.

[17] Y. Ding, X. Zhou, and X. Zhang, "YNU_DYX at SemEval-2019 Task 9: a stacked BiLSTM for suggestion mining classification," 2019, doi: 10.18653/v1/s19-2223.

[18] S. Pecar, M. Simko, and M. Bielikova, "NL-FIIT at SemEval-2019 Task 9: neural model ensemble for suggestion mining," 2019, doi: 10.18653/v1/s19-2214.

[19] R. S, A. Suseelan, S. M. Rajendram, and M. T T, "SSN-SPARKS at SemEval-2019 Task 9: mining suggestions from online reviews using deep learning techniques on augmented data," 2019, doi: 10.18653/v1/s19-2217.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[21] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Z. Gao, A. Feng, X. Song, and X. Wu, “Target-dependent sentiment classification with BERT,” *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/access.2019.2946594.

[23] C. C. Aggarwal, “Training deep neural networks,” in *Neural Networks and Deep Learning*, Springer International Publishing, 2018, pp. 105–167.

[24] “Understanding LSTM networks,” *Colah’s Blog*, 2015.

[25] P. Anki, A. Bustamam, H. S. Al-Ash, and D. Sarwinda, “High accuracy conversational AI chatbot using deep recurrent neural networks based on BiLSTM model,” Nov. 2020, doi: 10.1109/icoiact50329.2020.9332074.

[26] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv*, 2019, doi: 10.48550/arXiv.1907.11692.

BIOGRAPHIES OF AUTHORS

Ms. Anuradha Nandula working as an Assistant Professor in Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India and Research Scholar Osmania University, Hyderabad, India. Her research areas are natural language processing, text mining, text analytics, machine learning, and deep learning. She can be contacted at email: evanianuradha2002@gmail.com.

Dr. Panuganti Vijayapal Reddy working as a professor in the department of Computer Science and Engineering, Matrusri Engineering College, Hyderabad. He has 23 years of teaching experience at various administrative levels. He has published more than 50 research papers in reputed journals and conferences. He has authorised five text books and published 4 patents. He has guided 3 Ph.D. scholars and guiding 6 Ph.D. scholars. He has executed research funding projects worth of more than 15 lakhs. His research interests include natural language processing, text mining, data mining, and optical character recognition. He can be contacted at email: p.vijayapalreddy@matrusri.edu.in.