sustainability

MDPI

Article

Transformer Architecture-Based Transfer Learning for Politeness Prediction in Conversation

Shakir Khan [1,2,*], Mohd Fazil [3], Agbotiname Lucky Imoize [4], Bayan Ibrahimm Alabduallah [5,*], Bader M. Albahlal [1], Saad Abdullah Alajlan [1], Abrar Almjally [1] and Tamanna Siddiqui [6]

1 College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh 11432, Saudi Arabia
2 University Center for Research and Development, Department of Computer Science and Engineering, Chandigarh University, Mohali 140413, India
3 Center for Transformative Learning, University of Limerick, V94 T9PX Limerick, Ireland
4 Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Lagos, Akoka, Lagos 100213, Nigeria
5 Department of Information System, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11564, Saudi Arabia
6 Department of Computer Science, Aligarh Muslim University, Aligarh 202002, India
* Correspondence: sgkhan@imamu.edu.sa (S.K.); bialabdullah@pnu.edu.sa (B.I.A.)

Abstract: Politeness is an essential part of a conversation. Like verbal communication, politeness in textual conversation and social media posts is also stimulating. Therefore, the automatic detection of politeness is a significant and relevant problem. The existing literature generally employs classical machine learning-based models like naive Bayes and Support Vector-based trained models for politeness prediction. This paper exploits the state-of-the-art (SOTA) transformer architecture and transfer learning for respectability prediction. The proposed model employs the strengths of context-incorporating large language models, a feed-forward neural network, and an attention mechanism for representation learning of natural language requests. The trained representation is further classified using a softmax function into polite, impolite, and neutral classes. We evaluate the presented model employing two SOTA pre-trained large language models on two benchmark datasets. Our model outperformed the two SOTA and six baseline models, including two domain-specific transformer-based models using both the BERT and RoBERTa language models. The ablation investigation shows that the exclusion of the feed-forward layer displays the highest impact on the presented model. The analysis reveals the batch size and optimization algorithms as effective parameters affecting the model performance.

Keywords: politeness prediction; conversation AI; machine learning; transfer learning

Citation: Khan, S.; Fazil, M.; Imoize, A.L.; Alabduallah, B.I.; Albahlal, B.M.; Alajlan, S.A.; Almjally, A.; Siddiqui, T. Transformer Architecture-Based Transfer Learning for Politeness Prediction in Conversation. *Sustainability* **2023**, *15*, 10828. https://doi.org/10.3390/su151410828

Academic Editors: Mourade Azrour, Azidine Guezzaz, Imad Zeroual, Azeem Irshad, Jamal Mabrouki, Said Benkirane, Shehzad Ashraf Chaudhry and Rui Xiong

Received: 22 May 2023
Revised: 12 June 2023
Accepted: 14 June 2023
Published: 10 July 2023

1. Introduction

In literature, politeness refers to good manners or etiquette in human behavior. It is also reflected in online conversations while using social networks. Politeness is the fundamental etiquette of various communication methods, either verbal, textual, or something else [1]. In Online social media, politeness in conversation is more crucial due to anonymity, the abundance of fake profiles, and usability issues. Advancement in natural language understanding and machine learning-based large language models are employed in various domains [2–5]. These language models also provide the opportunity to encode and detect the presence of politeness in a conversation or social media post. Online messaging platforms are popular and used by millions of users. Large organizations and business houses also use intelligent chatbots to communicate and resolve the queries of their customers. Therefore, if a chatbot is not polite in its language, it may lead to customer dissatisfaction which will hamper the organization's business. Impolite language in a conversation may lead to conflict and hate speeches among the users. Online social media

platforms can avoid these conflicts through timely moderation of instigating language. Therefore, tracking, profiling & predicting politeness in a conversation is vital. This study presents a transformer architecture-based deep neural network model for classifying polite conversations from impolite ones.

Our Contributions

Recently, offensive speech problem has attracted researchers from across disciplines to analyze different aspects of hate speech problem. They also introduced various machine learning models to classify hate content into different categories [6,7]. Impolite language is a type of offensive content. In the existing literature, a closely related but contrasting research problem—politeness prediction, is understudied. Recently, this vital problem in textual content has received attention from researchers. They analyzed and addressed different aspects of an offensive content problem in conversational systems and chatbots [8–10]. However, existing approaches understudied the detection of politeness in social media posts. However, the politeness prediction methods don't employ the state-of-the-art transformer-based representation models like Bidirectional Encoder Representations from Transformers (BERT) [11], Robustly Optimized BERT (RoBERTa) [12]. To this end, this study presents a transformer employing deep model for politeness prediction in textual content. The preprocessed input text is passed through a BERT layer to encode the textual input into a numeric representation. It converts each word into a vector of the same length. Further, the model forwards the BERT representation to a feed-forward neural network (FFN) consisting of 2 dense layers to learn abstract feature extraction and representation. Finally, the encoded vector from FNN is passed to a sigmoid layer for classification to predict the final class of the text label.

In short, the proposed model can be summarized as employing the strength of transformer-based language architecture, feed-forward dense layer, and variable weight assignment to words based on their importance through an attention mechanism for efficient representation learning. Further, the trained representation is passed through a sigmoid layer having two neurons for politeness prediction in the content. We outline the contributions of the presented study as follows:

- Introduce a deep model for the understudied problem of politeness prediction by integrating the strength of transformer architecture-based large language models, feed-forward dense layer, and attention mechanism. Presented model learns an efficient text representation, passed to a sigmoid layer to classify into polite and impolite categories.
- Perform an in-depth investigation of the presented model by applying the initial weight assignment from transformer-based language models in politeness prediction over two benchmark datasets, including a blog dataset.
- Conduct an ablation study to investigate the impact of various neural network components like the attention mechanism towards the politeness prediction.
- Also, investigate the impact of different values of hyperparameters like batch size and optimization algorithm on the efficacy of the presented model to discover their optimal number.

The remaining manuscript is divided into the following sections. Section 2 explores the existing literature studying the different aspects of the politeness prediction. It also discusses the evolution of existing classification models presented over the year. Section 3 discusses the proposed model, including all its components. Further, Section 4 describes datasets, empirical evaluation, and analytical observations. Conclusion Section 6 presents the main findings of this study and a discussion of future research directions.

2. Literature Survey

The existing literature has analyzed different aspects of politeness in textual content generated from conversation systems and simple online posts. Thus, we group existing approaches into two categories: (i) politeness prediction methods developed for conversational systems and (ii) politeness prediction in textual content.

Researchers have presented politeness detection methods for conversational systems in the first group of approaches. These existing approaches generally employ different neural network components to propose architectures for politeness prediction [13–15]. In [14], the authors used emotions extracted from the multi-modal content to develop an end-to-end dialogue generation framework. The inclusion of sentiment made the dialogue system efficient and user-adaptive. In another interesting study, authors in [16] in-depth investigated the correlation between politeness and social interaction. In [17], the authors presented three weakly supervised machine learning models to generate diverse and polite-incorporating text. In [18], authors presented a dictionary matching and machine learning-based generative model. They used word- and sentence-level emotion captured from embeddings to generate a quality emotion dataset. Cristiana et al. [19] presented a sliding window-based strategy to compute the politeness score of each sentence in a conversation. It shows how politeness in a conversation evolves and tracks the chances of a conflict. Similarly, Zhang et al. [20] presented a politeness- and rhetoric-based strategy to early detection of derailing conversations. In another similar approach, Jonathan et al. [21] devised an unsupervised forecasting model to learn the representation of conversation dynamics and predict its probability of turning into a toxic conversation.

Authors in [8] studied the linguistic aspects of politeness. They devised lexical and syntactic features incorporating politeness-based theories like *difference* and *modality*. Authors found that polite Wikipedia authors have higher social power. In a vital work, Madaan et al. [9] presented a pipeline to convert impolite sentences into polite sentences. The proposed approach used the concept of *tag* and *generate*. In [22], the authors introduced a simple convolutional neural network-based model to predict politeness in natural language requests. In [23], researchers studied the linguistic features of politeness and presented a lexicon of politeness features called *PoliteLex*. They performed various empirical analyses to observe the politeness perception across cultures. Authors in [24] analyzed the impact of politeness in a vehicle speech interface on drivers' trustworthiness and found that polite speech interfaces have more trust among drivers. They also used activation clusters to identify the linguistic factors behind the polite text. In another method, authors in [25] presented a hierarchical transformer-based architecture to detect politeness in a goal-oriented dialog system. The authors also investigated the system efficacy against multiple baseline systems. As per our knowledge, no study in the existing literature employs the transformer-based context-incorporating language model for politeness prediction. It is the first study that integrates the strength of the context-aware language model, feed-forward neural network & attention mechanism for politeness prediction.

3. Models and Methods

The architecture of the presented model includes four layers: data-preprocessing, contextual embedding layer, feed-forward neural network layer, and output layer, as shown in Figure 1. We present a detailed introduction of these four layers in the subsequent subsections.

3.1. Data Pre-Processing

We investigate the efficacy of our model on two benchmark datasets from Wikipedia and Stack-Exchange question-answering forums provided by Mizil et al. [8]. We applied pre-processing steps to these datasets to filter noisy content and useless symbols. We performed the following pre-processing steps in sequence:

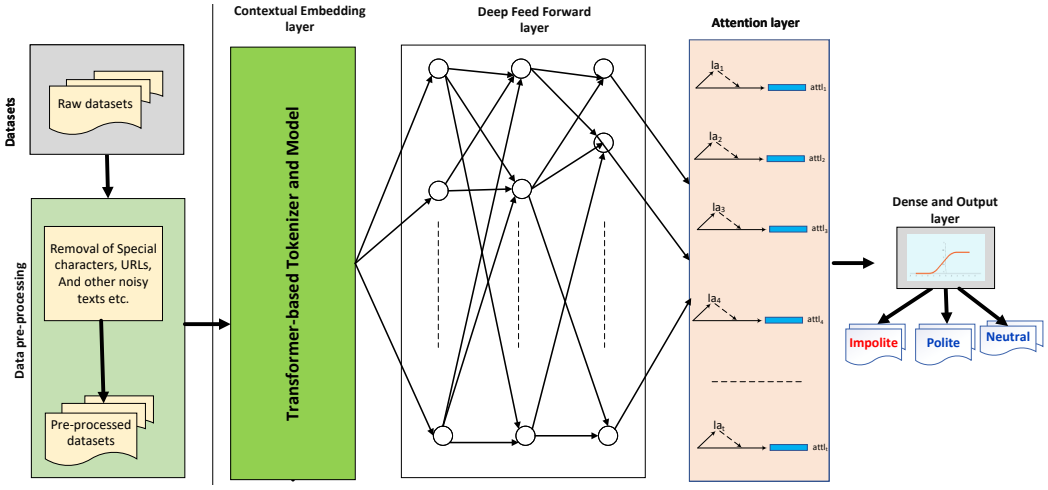

Figure 1. Workflow the model for politeness prediction.

3.1.1. Contraction Expansion

In English, many users use the contracted form of verbs, like I'm for I am, to shorten the sentence length. However, it is difficult for a language model to understand these words. Interestingly, negative contraction changes sentiment. Therefore expanding the contracted verb is essential. To this end, we identified 128 contracted verb forms like can't, could've with expansion to replace the contracted verb with their expanded form.

3.1.2. URL Filtration

In a text, URLs or hyperlinks don't contain much information. Therefore, we filtered all types of URLs.

3.1.3. Special Character Removal

The representation learning does not has any presentation for special characters. Therefore, we filter all the special characters and symbols like '#', '?' and ' ' to clean the text. We also remove the usernames (words starting with '@' symbol), replace two or more repetitions of the same character with a maximum of two, filter non-ASCII characters & numbers, and remove the extra white spaces. Finally, we convert the text into lower case to avoid the case-related issue.

3.2. Contextual Embedding Layer

Transformer-based large-language models are state-of-the-art models for text representation learning. These pre-trained models, trained over large corpora, are massive and demonstrate effective performance in different downstream tasks. For example, BERT (Bidirectional Encoder Representations from Transformers) [11,26] is trained on a large unlabelled corpus having the complete Wikipedia, approximately 2500 million words, and Book corpus, 800 million words. The two versions of the BERT: $BERT_{BASE}$ and $BERT_{LARGE}$, are used based on the complexity of the underlying problem. The first one, $BERT_{BASE}$, has 12 hidden layers along with 12 attention heads and 110 million parameters. On the other hand, the second one, $BERT_{LARGE}$, has 24 encode-decoder layers with 16 attention heads and 340 million parameters. Authors of the BERT model evaluated it on two tasks: (1) masked language modeling and (2) Next Sequence Prediction. Unlike traditional recurrent neural networks, which process information sequentially, BERT's success lies in training on a large corpus and parallel processing. Similarly, researchers have introduced many improved and domain-specific BERT, such as RoBERTa [12] and MentalBERT [27]. Training a BERT model from scratch requires massive computing resources and a large dataset. We generally fine-tune a model to avoid the requirement of computing resources and incorporate the domain-specific content in representation learning. In this study also, we fine-tune two transformer-based large language models—BERT and RoBERTa for effi-

cient representation learning to predict the politeness label of a text. Fine-tuning refers to using the weights of an already trained network as the starting values for training a new model, like weights from BERT and RoBERTa models as the starting weight in the presented model. Further, this weight is updated for politeness prediction during the training process. This whole process of using already trained weight on one problem and updating it for another problem is called fine-tuning.

3.3. Attention-Aware Deep Feed Forward Network Layer

The encoded output from the transformer-based language model passes through a feed-forward network consisting of two dense layers. The presented model includes two layers for efficient and effective representation learning. The encoded output passes through an attention layer, which assigns weights to encoded features based on their segregating power in classifying three target classes. If the input encoded representation of a feature f is f_n, then the attention layer learns f'_n representation using Equation (1). Further, it computes the similarity using dot product between f'_n and a high-level context tensor v_h. Finally, based on the computed similarity, Equation (2) calculates the attention score α_f of each feature f. The context tensor v_h is randomly initialized and updated during the training process [28]. The attention score is multiplied with each feature to assign the relative weight. Finally, the resultant representation vector, F_n, is the weighted sum of the hidden features, as shown in Equation (3).

$$f'_n = \tanh(wf_n + b) \tag{1}$$

$$\alpha_f = \frac{exp(f'_n v_h)}{\sum_f exp(f'_n v_h)} \tag{2}$$

$$F_n = \sum_f (\alpha_f f_n) \tag{3}$$

3.4. Output Layer

Finally, the learned feature vector from the attention-aware deep FFN layer is passed to a softmax layer to classify each text into one of three categories: *polite*, *impolite*, and *neutral* for final labeling.

4. Experiment

We investigate the efficacy of the presented model over two standard datasets related to Wikipedia and Stack-exchange. This section describes evaluation phases like dataset, hyper-parameter setting, and evaluation metrics. Finally, this section ends with a discussion of experimental results and performs the comparison with the SOTA and baseline methods.

4.1. Evaluation Datasets

We use the two publicly released datasets by Mizil et al. [8] to evaluate the presented model. The first one, D_w, is a Wikipedia request dataset. It initially has 35,991 text annotated by 219 annotators from Amazon Mechanical Turk (AMT), a crowdsourcing marketplace. Finally, the constructed dataset contains 4353 requests, each having exactly two-sentence, where the second sentence is the request. The annotated dataset contains both politeness score and class. Every Wikipedia request is given one of the three labels: *polite*, *impolite*, and *neutral* depending upon the politeness score. The first row of Table 1 gives brief statistics of D_w. The authors also constructed a dataset D_s from the Stack-Exchange forum, a question-answer community. These two data source platforms are standard for conversational datasets having user-to-user request information. Table 1 presents a brief description of datasets. We can see from the table that both datasets are almost balanced.

Table 1. A brief dataset statistics.

Dataset	Dataset Size	#Polite	#Impolite	Neutral
D_w	4353	1089	1089	2175
D_s	8254	3302	1651	3301

4.2. Experimental Setting

The experiments conducted in this study are coded in Python version 3.7.12 using Keras 2.7.2 framework over the freely available Google colab notebook. All the training is performed using a five-fold cross-validation strategy, which splits the evaluation dataset into five equal parts. Under this strategy, four parts train the underlying model, whereas the left part evaluates the trained model. This procedure is conducted five times to ensure the usability of each sample in training and testing. We train the model in the batches of 16 instances for 20 epochs. *Adam* and categorical cross-entropy are used as the optimization method and loss function, respectively. The learning rate is 0.001 during the training process. Table 2 provides various hyper-parameters used in the models and underlying adjusted values.

Table 2. Parameters in the proposed model.

Hyperparameter	Value
Model learning rate	0.001
Batch size	16
Loss method	Categorical Cross-Entropy
Optimization algorithm	Adam
Dropout	0.5
Epoch	20

4.3. Experimental Results

This section presents the experimental results over the two benchmark datasets. We compare the model to SOTA and six baselines considering accuracy to establish its efficacy. We use the recent transformer-based pre-trained language models—BERT and RoBERTa for encoding the textual content to classify the text into polite, impolite, and neutral categories. The first two rows of Table 3 present the results using the BERT and RoBERTa language models. We can see from the table that our model reports an accuracy of approximately 90%.

4.3.1. Comparative Evaluation

To establish the efficacy of the presented approach, we evaluate its comparison against two SOTA and six baseline methods for politeness prediction. We constructed six baselines: two domain-specific large language models: fBERT and HateBERT and four using neural network components like LSTM and BiLSTM to analyze their impact on politeness prediction. We use GloVe embedding of 200-d as input to the embedding layer in the last four baselines. The following paragraphs discuss the SOTA and baseline models briefly.

- Aubakirova and Bansal [22]: In this paper, the authors introduced a simple neural network employing the convolutional neural network to predict the politeness in requesting sentences. Further, the authors performed network visualization using activation clusters, first derivative saliency, and embedding space transformation to analyze the linguistic signals of politeness.
- Mizil et al. [8]: In this early study, the authors presented a computational framework employing the domain-independent lexicon and syntactic features to analyze the linguistic aspect of politeness. They further trained an SVM classifier to predict politeness. They also investigated the relationship between politeness and social power.

- BiLSTM: The first baseline, BiLSTM, is a simple RNN network. It incorporates both left-to-right and right-to-left contexts during representation learning. This model has an input layer to receive the embedding-based text representation followed by a BiLSTM layer having 128 neurons. Finally, a final softmax layer classifies the text into polite and non-polite categories.
- LSTM: The second baseline is an LSTM network to compare its performance against the presented model. The baseline network has an LSTM layer with 128 neurons. It also has an input layer and a softmax layer for classification. It also uses GloVe embedding of 200d as input to the model.
- BiGRU: It is the third baseline of this paper. It uses 128 neurons in the BiGRU layer for tweet representation learning. It also has an Embedding layer using 200d Glove embedding and a softmax layer to perform final classification.
- ANN: The model performance is also compared with a simple artificial neural network. This ANN model has 3 hidden layers having 128, 64, and 32 neurons. It also has an embedding layer and a final softmax layer for classification. This baseline model takes 200d GloVe embedding as input.
- fBERT [29]: It is a BERT model, pre-trained on a large English offensive language corpus (SOLID), containing more than 1.4 million offensive instances.
- HateBERT [30]: It is another BERT model, pretrained for abusive language detection. It is trained on a Reddit dataset of communities banned for being offensive, abusive, or hateful.

Table 3. Evaluation results considering accuracy over D_w and D_s.

Datasets →	D_w	D_s
Methods ↓	**Accuracy**	**Accuracy**
Proposed Model [BERT]	0.91	**0.87**
Proposed Model [RoBERTa]	**0.92**	0.84
Aubakirova and Bansal [22]	0.85	0.66
Mizil et al. [8]	0.83	0.78
fBERT [29]	0.90	0.87
HateBERT [30]	0.89	0.83
ANN	0.87	0.82
BiLSTM	0.88	0.76
LSTM	0.87	0.76
BiGRU	0.88	0.78

Table 3 presents the comparative evaluation results considering accuracy. We do not present the training accuracy because it is not viable. We can see that over both datasets, our model, employing the BERT and RoBERTa language models, shows the best result. The table demonstrates that over D_w, the proposed model with BERT performs best with an accuracy of 0.91. On the other hand, over D_s, the model performs best employing the RoBERTa model with an accuracy of 0.87. We can also see from the table that the model also significantly outperforms all the baseline models. The fBERT achieves an accuracy of 0.90 and 0.87 over D_w and D_s, respectively, and performs best among baselines. On contrary, the LSTM baseline shows the worst performance. We can observe from the table that the domain-specific transformer-based large language models show comparative performance. Overall, comparative models show relatively better performance over D_w.

4.3.2. Ablation Analysis

The presented model has two main neural components—(i) feed-forward neural layer (FNN) and (ii) attention mechanism. The impact of each neural network component is analyzed by conducting ablation analysis. We performed the ablation analysis with both BERT and RoBERTa. In the first ablation analysis, we exclude the feed-forward neural layer from the proposed model to construct a model having an input layer (embedding layer), an attention layer, and a final softmax layer. The underlying results for constructed models are given in the 5th and 6th rows of Table 4. In further ablation analysis, the attention mechanism is excluded from the proposed model to construct a model with a contextual-embedding layer, FNN layer, and a final softmax layer for prediction. The results for this ablation analysis are presented in the 7th and 8th row of Table 4. The ablation study establishes that excluding the attention mechanism has least impact on the proposed model with both BERT and RoBERTa language models. Further, exclusion of feed-forward layer has insignificant impact on the performance and reduces the RoBERTa-based accuracy by 3% over D_w. Interestingly, the model performance with RoBERTa increases on removal of FNN over D_s.

Table 4. Ablation evaluation results considering accuracy on two datasets.

Datasets →	D_w	D_s
Methods ↓	**Accuracy**	**Accuracy**
Proposed Model [BERT]	0.91	0.87
Proposed Model [RoBERTa]	0.92	0.84
Proposed model [BERT] (without FNN)	0.89	0.86
Proposed Model [RoBERTa] (without FNN)	0.89	0.87
Proposed model [BERT] (without Attention)	0.91	0.86
Proposed Model [RoBERTa] (without Attention)	0.91	0.83

5. Discussion: Evaluation of Hyperparameters Impact

In a neural network model, many hyper-parameters affect performance. To this end, we investigate to observe the impact of *batch size* and *optimization algorithms* on the model performance over D_w and D_s. We evaluate the model performance considering accuracy.

5.1. Batch Size

In a deep model, the number of training instances processed through it in one go is called *batch size*. It is a hyperparameter because a user can fine-tune it to optimize the model performance. Suppose a dataset contains 500 samples, and if the batch size is 50, then the model is trained using the first 50 samples, again trained using the next 50, and this process continues until the dataset exhaust. We investigate the model efficacy for both BERT and RoBERTa on 4 different batch sizes—16, 32, 64, and 128 over D_w and D_s and Figure 2 depicts the underlying results. It reveals as the batch size increases, the BERT-based model accuracy degrades. Our model best performs with 16 and 32 batch sizes on both datasets. The model performance with RoBERTa-based contextual embedding is shown in Figure 2b. The model over the D_w dataset shows the best performance with 16 batch size, whereas it shows the best result on the D_s using 32 batch size. We can conclude that the model demonstrates the best evaluation results when processed in batches of 16 instances. Therefore, evaluation results justify the adjustment of batch size to 16 in this paper.

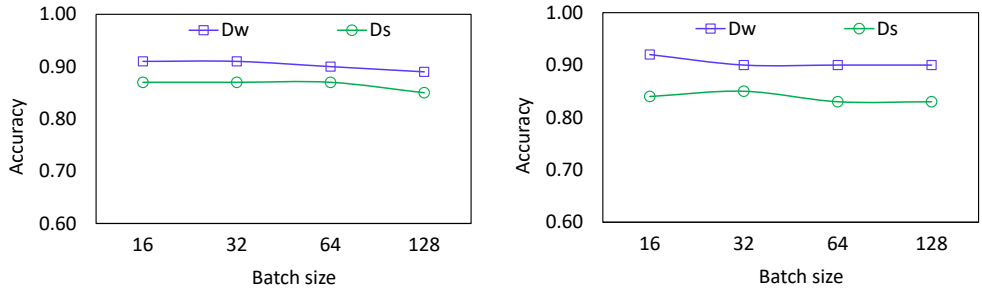

Figure 2. Evaluation results of proposed model on different batch sizes over D_w and D_s considering accuracy for (**a**) BERT (**b**) RoBERTa.

5.2. Optimization Algorithms

The optimization method used in a deep model is another parameter that affects the model performance. We investigate the impact of Adam, Adagrad, and Adadelta on the results of our model. Figure 3 displays the underlying evaluation results for both BERT and RoBERTa models the over the D_w and D_s datasets. The Figure 3a reveals that our model using BERT shows the most promising result with Adam and the worst result with Adadelta over both datasets. Figure 3b exhibits similar performance for the proposed model employing RoBERTa. The figure shows that the impact of the optimization algorithm is more significant over the D_w dataset. Overall, we can infer that the proposed model with Adam performs best and shows the worst performance with Adadelta. On the other hand, it shows a comparative performance with Adagrad. Finally, this evaluation result justifies the selection of Adam as an optimization algorithm.

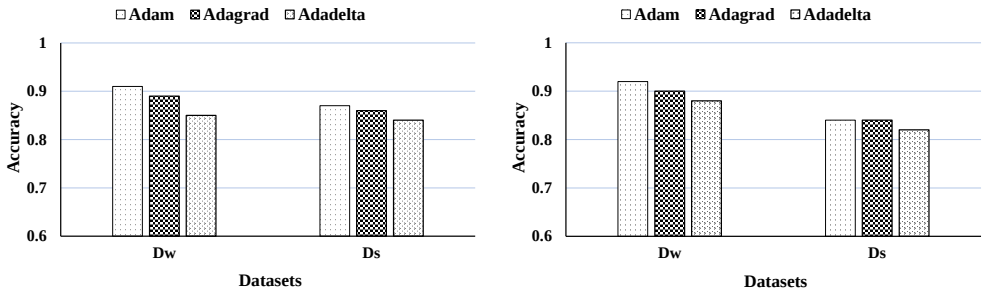

Figure 3. Empirical results of the model using the three optimization algorithms over D_w and D_s considering accuracy for (**a**) BERT (**b**) RoBERTa.

6. Conclusions and Future Directions of Work

In this research, we developed an advanced deep learning model that outperforms the existing methods. We introduced transformer-based architecture for representation learning toward politeness prediction in a text. The presented model integrated the strengths of context-aware language models, a feed-forward neural network, and an attention mechanism for the representation learning of natural language requests. The trained representation is further passed through a softmax layer and classified into polite, impolite, and neutral classes. We evaluated the model over the two benchmark datasets considering accuracy. In the comparative analysis, the proposed model outperformed the two SOTA and six baseline models, including two offensive content specific large language models. We also examined the hyperparameter effect on the model to ascertain the use of their optimal value in this study.

The proposed model lacks content, network, and profile-related features, which can be vital. In further research, we will incorporate these feature categories. Though the proposed model is highly effective for politeness prediction in English texts, it has limitations. Like, it has been evaluated only on English datasets. Its adaptation over multi-lingual or code-

mixed data is a promising future research. Second, it has not been evaluated for hate and offensive content detection.

Author Contributions: Conceptualization, M.F. and S.K.; methodology, M.F.; software, S.K.; validation, A.A. and S.A.A.; resources, B.I.A.; data curation, M.F.; writing—original draft preparation, S.K. and T.S.; writing—review and editing, M.F.; visualization, B.M.A.; supervision, A.L.I.; project administration, B.I.A.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Deanship of Scientific Research at Mohammad Ibn Saud Islamic University (IMSIU) through grant number RP-21-07-06 and Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R440).

Acknowledgments: The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding and supporting this work through Research Partnership Program no. RP-21-07-06 and Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R440), Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. The authors would also like to thank Aleem Ali from Department of Computer Science and Engineering, Chandigarh University for serving as a consultant to critically review the study proposal and participating in technical editing of the manuscript

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Khan, S. Business Intelligence Aspect for Emotions and Sentiments Analysis. In Proceedings of the First International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT, Trichy, India, 16–18 February 2022; pp. 1–5.
2. Haq, A.U.; Li, J.P.; Ahmad, S.; Khan, S.; Alshara, M.A.; Alotaibi, R.M. Diagnostic Approach for Accurate Diagnosis of COVID-19 Employing Deep Learning and Transfer Learning Techniques through Chest X-ray Images Clinical Data in E-Healthcare. *Sensors* **2021**, *21*, 8219. [CrossRef] [PubMed]
3. Qaisar, A.; Ibrahim, M.E.; Khan, S.; Baig, A.R. Hypo-Driver: A Multiview Driver Fatigue and Distraction Level Detection System. *Cmc-Comput. Mater. Contin.* **2021**, *71*, 1999–2017.
4. Abulaish, M.; Kumari, N.; Fazil, M.; Singh, B. A Graph-Theoretic Embedding-Based Approach for Rumor Detection in Twitter. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Thessaloniki, Greece, 14–17 October 2019; pp. 466–470.
5. Mahajan, S.; Pandit, A.K. Hybrid method to supervise feature selection using signal processing and complex algebra techniques. *Multimed. Tools Appl.* **2023**, *82*, 8213–8234. [CrossRef]
6. Khan, S.; Fazil, M.; Sejwal, V.K.; Alshara, M.A.; Alotaibi, R.M.; Kamal, A.; Baig, A. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 4335–4344. [CrossRef]
7. Khan, S.; Kamal, A.; Fazil, M.; Alshara, M.A.; Sejwal, V.K.; Alotaibi, R.M.; Baig, A.; Alqahtani, S. HCovBi-Caps: Hate Speech Detection using Convolutional and Bi-Directional Gated Recurrent Unit with Capsule Network. *IEEE Access* **2022**, *10*, 7881–7894. [CrossRef]
8. Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; Potts, C. A computational approach to politeness with application to social factors. In Proceedings of the International Conference of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; pp. 250–259.
9. Madaan, A.; Setlur, A.; Parekh, T.; Poczos, B.; Neubig, G.; Yang, Y.; Salakhutdinov, R.; Black, A.W.; Prabhumoye, S. Politeness Transfer: A Tag and Generate Approach. In Proceedings of the International Conference of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 1869–1881.
10. Niu, T.; Bansal, M. Polite Dialogue Generation Without Parallel Data. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 373–389. [CrossRef]
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
12. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In Proceedings of the ICLR, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 1–13.
13. Wen, T.H.; Vandyke, D.; Mrksic, N.; Gasic, M.; Rojas-Barahona, L.M.; Ultes, P.H.S.S.; Young, S. A Network-based End-to-End Trainable Task-oriented Dialogue System. In Proceedings of the International Conference of European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; pp. 438–449.
14. Shi, W.; Yu, Z. Sentiment Adaptive End-to-End Dialog Systems. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1509–1519.
15. Mishra, K.; Firdaus, M.; Ekbal, A. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing* **2022**, *494*, 242–254. [CrossRef]
16. Brown, P.; Levinson, S.C.; Levinson, S.C. *Politeness: Some Universals in Language Usage*; Cambridge University Press: Cambridge, UK, 1987; Volume 4.

17. Niu, T.; Bansal, M. Polite Dialogue Generation Without Parallel Data. In Proceedings of the the European Conference on Information Retrieval, Padua, Italy, 20–23 March 2018; Springer: Cham, Switzerland, 2018; pp. 810–817.
18. Peng, D.; Zhou, M.; Liu, C.; Ai, J. Human–machine dialogue modelling with the fusion of word- and sentence-level emotion. *Knowl.-Based Syst.* **2019**, *192*, 105319. [CrossRef]
19. Iordache, C.P.; Trausan-Matu, S. Analysis and prediction of politeness in conversations. In Proceedings of the International Conference on Human Computer Interaction, Bucharest, Romania, 16–17 September 2021; pp. 15–20.
20. Zhang, J.; Chang, J.P.; Danescu-Niculescu-Mizil, C. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1350–1361.
21. Chang, J.P.; Danescu-Niculescu-Mizil, C. Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, Hongkong, China, 3–7 November 2019; pp. 1–12.
22. Aubakirova, M.; Bansal, M. Interpreting Neural Networks to Improve Politeness Comprehension. In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2035–2041.
23. Li, M.; Hickman, L.; Tay, L.; Ungar, L.; Guntuku, S.C. Studying Politeness across Cultures using English Twitter and Mandarin Weibo. In Proceedings of the CSCW, Virtual, 17–21 October 2020; pp. 1–15.
24. Lee, J.G.; Lee, K.M. Polite speech strategies and their impact on drivers' trust in autonomous vehicles. *Comput. Hum. Behav.* **2022**, *127*, 107015. [CrossRef]
25. Mishra, K.; Firdaus, M.; Ekbal, A. Predicting Politeness Variations in Goal-Oriented Conversations. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 1–10. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 1–11.
27. Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; Cambria, E. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In Proceedings of the the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; European Language Resources Association: Paris, France, 2022; pp. 7184–7190.
28. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
29. Sarkar, D.; Zampieri, M.; Ranasinghe, T.; Ororbia, A. fBERT: A Neural Transformer for Identifying Offensive Content. In Proceedings of the Proc. of the EMNLP, Punta Cana, Dominican Republic, 1–6 August 2021; pp. 1792–1798.
30. Caselli, T.; Basile, V.; Mitrović, J.; Granitzer, M. HateBERT: Retraining BERT for Abusive Language Detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms, Online, 7–13 November 2021; pp. 17–25.