Deep Learning to Detect Computer-Generated Reviews in E-Commerce

1st Han Chau
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, NJ, U.S.
hchau@stevens.edu

***Abstract*—The rapid growth of online shopping platforms has led to an increase in the volume of user-generated content and online reviews, which have a significant influence on consumers' pre-purchase selections. In the e-commerce space, reviews and comments have a significant impact on sellers' and products' reputations. However, the prevalence of fraudulent reviews calls into question the legitimacy of online platforms, creating significant obstacles for users to discern between genuine and useful content. This paper addresses the crucial problem of distinguishing between computer-generated and human-authored evaluations and offers comprehensive solutions. By utilizing deep learning methods, the study seeks to enhance the reliability and usefulness of online product reviews by filtering reviews.**

For further code in this project see GitHub link or the next url: https://github.com/hanchau94/Deep-Learning-Project.git

I. Introduction

In 2021, the global landscape of retail e-commerce witnessed a staggering total of approximately 5.2 trillion U.S. dollars in sales.[1] A comprehensive survey, conducted in the second quarter of 2023, encompassed responses from 410 retail and 415 consumer packaged goods (CPG) companies worldwide, boasting revenues ranging from $50 million to over $10 billion.[2] Updated forecasts signal an anticipated 8.9% growth in worldwide e-commerce sales for 2023, building on the momentum of the most recently tracked period.[3] By 2024, projections indicate that 21.2% of total retail sales will transpire in the online domain.[4]

In the dynamic landscape of e-commerce, user reviews play a pivotal role, exerting considerable influence on the decision-making process of potential buyers. A recent study revealed that 87% of consumers attribute greater influence to real-life customer reviews and ratings compared to influencer or celebrity endorsements (50%).[5] However, the prevalence of fraudulent reviews, comprising approximately 50% of five-star ratings with a staggering count of 2.7 million detected in 2021, poses a significant challenge to the authenticity of online platforms.[6]

Recognizing the impact of these challenges, this project takes on the formidable task of semantic and pattern analysis, employing advanced deep learning techniques for discerning between reviews crafted by machines and those genuinely composed by humans. Using Hugging Face Transformers library, I am able to access the pre-trained BERT model that will serve as the foundation for the Fake Review Detection Model. BERT can understand words incredibly well. Consequently, it will help my model comprehend the review context more fully and enable it to forecast whether a review will be conducted by a human or a computer. PyTorch will be used to define, train, adjust, and assess the models.

Through a meticulous examination of the strengths and limitations inherent in both computational and human-generated sentiments, the aspiration is to contribute to the development of more resilient and trustworthy online platforms. By fostering consumer confidence and enabling informed decision-making in the digital marketplace, this project aims to address the pressing challenges posed by the proliferation of fake reviews and bolster the integrity of the e-commerce landscape.

II. Related Work

User-generated content, especially online reviews, has significantly influenced consumer behavior within the rapidly expanding realm of e-commerce. The impact of these reviews on purchasing decisions has made them a cornerstone of online marketplaces. However, the prevalence of counterfeit reviews has muddled the reliability of these platforms, making it challenging for users to distinguish between authentic and misleading content. Recognizing this issue, recent studies have delved into detecting fake reviews.

In 2022, Sherry He et al. examined the structural importance of products in online networks, employing measures like degree, eigenvector centrality, and PageRank. This research focused on understanding product connectivity and their relationship within the network, using metrics like clustering coefficients to assess reviewers' commonalities among product neighbors.[7]

Early in 2023, Choi et al. contributed by developing a model that evaluated review authenticity and usefulness using various machine learning techniques. Their work involved employing multiple supervised learning models, including SVC, LGBM Classifier, RandomForest Classifier, among others. Their findings showed promising accuracy ranging from 81% to 85% using these models to judge review usefulness.[8]

In May 2023, Lu et al. proposed a novel approach called BSTC for detecting fake reviews. This model amalgamated BERT, SKEP, and TextCNN, leveraging a pre-trained language model focused on sentiment knowledge enhancement. Their findings showcased BSTC outperforming existing methods,

achieving high accuracies across different datasets—Hotel, Restaurant, and Doctor—with rates of 93.44%, 91.25%, and 92.86%, respectively.[9]

In the context of identifying computer-generated reviews in e-commerce, the emphasis lies on utilizing a pre-trained BERT model sourced from Hugging Face Transformer library and fine-tuning a pretrained model in native PyTorch. This study aims to discern between computer-generated and human-generated reviews within a dataset generated by GPT-2, as detailed by Salminen et al. in 2022.[10] Additionally, the process involves training another model using a new dataset comprising 10,000 reviews sourced from the Amazon review dataset, serving as a foundation for training the original dataset.

III. Method

To distinguish between authentic and computer-generated reviews, the approach integrates cutting-edge techniques in sentiment analysis, utilizing a robust methodology rooted in deep learning. By harnessing the power of the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model from Hugging Face Transformers library, it establishes a strong foundation for the Fake Review Detection Model. BERT's remarkable language comprehension capabilities enable a nuanced grasp of context, empowering the model to make intelligent predictions regarding the authenticity of online reviews. The implementation of PyTorch serves as a versatile tool for defining, training, fine-tuning, and evaluating the models.

The methodology is structured into two primary phases: training and testing. In the training phase, the model undergoes exposure to a diverse labeled dataset encompassing both genuine and computer-generated reviews. Through supervised learning, the model hones its understanding of semantic analytics and patterns inherent to authentic human sentiments. The fine-tuning process involves iterative parameter adjustments, further enhancing the model's discernment capabilities.

Additionally, we will create a new dataset by applying cosine similarity to select 10,000 reviews from the Amazon review dataset. We will then predict labels using the model on the smaller subset of the original dataset. This new dataset is chosen based on the cosine similarity to the validation set within embedding a 768-dimensional dense vector space from the Sentence-Transformers library, providing a basis for training the original dataset. By utilizing the pre-trained model, known Model 2, as a starting point for Model 3, I leverage the knowledge and features learned during the previous training, potentially enhancing the model's ability to generalize patterns in the original data. The pre-trained model (Model 2) can act as a feature extractor, capturing relevant information from the Amazon dataset. This extracted knowledge can then be fine-tuned on the original dataset, potentially improving the model's performance on the specific task of distinguishing between human and computer-generated reviews. Generate predictions for each instance in the test set using all of the trained models by assigning the label with the majority of votes as the final prediction for Model 4.

For testing, the model is evaluated on a separate set of reviews, including a significant proportion of computer-generated content. Rigorous performance metrics are employed, such as precision, recall, and F1 score, to assess the model's effectiveness in accurately classifying reviews.

IV. Data

A. Description of Dataset

The dataset consists of 40,412 reviews, evenly divided between approximately 20,000 genuine product reviews marked as OR (presumed to be human-created and authentic) and 20,000 fake reviews labeled as CG (Computer-generated).[11] It encompasses four columns: Category, Rating, Label, and Text_. For this study, the primary focus revolves around the Label and Text_ columns.

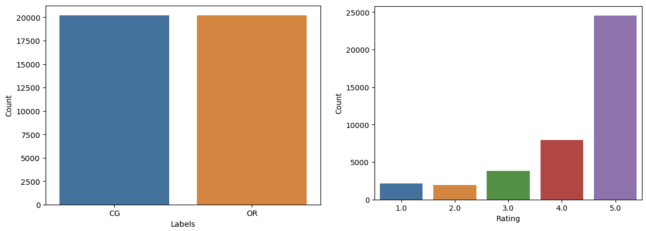

Figure: Plot the histogram of Labels and Ratings

Fig. 1. Plot the histogram of Labels and Ratings

Additionally, a vast dataset in fastText format, sourced from Kaggle, containing several million Amazon reviews, will be harnessed to create a new unlabeled dataset intended for use in constructing the second model.[12]

B. Data preprocessing

The original dataset requires encoding the labels ('computer' as 0 and 'human' as 1) for model training. The dataset will be segmented into three parts: the training set comprising roughly 32,000 reviews, a validation set with about 4,000 reviews, and a test set also containing approximately 4,000 reviews. These segments represent 80%, 10%, and 10% of the total dataset, respectively.

Utilizing the AutoTokenizer from the Hugging Face's Transformers library, the chosen model, 'bert-base-uncased' is preferred to limit the vocabulary size. Configuring the model involves setting tokens such as [CLS] and [SEP] while handling Padding and Truncation to ensure consistent input sizes. The focus remains on keeping text in lower case. The determination of the max_length is facilitated by counting the tokens using CountVectorizer for each review.

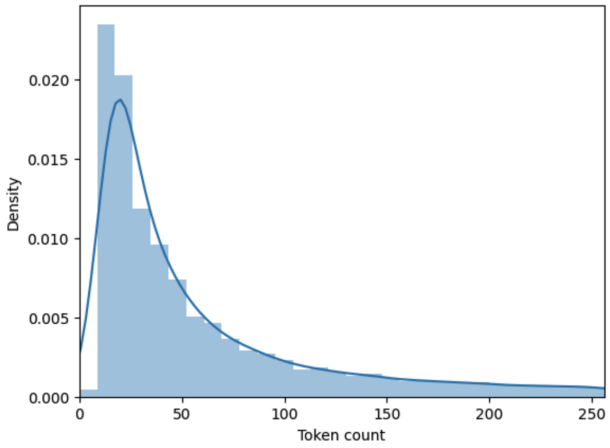

Fig. 2. Plot the distribution of review lengths

The distribution of review lengths indicates a noticeable peak around 30 tokens, with some reviews extending beyond 250 tokens. To strike a balance between computational efficiency and information retention, I've opted for a maximum length of 150 tokens per review. This decision aims to prevent excessive resource consumption while still capturing meaningful patterns for the model without compromising on memory or computational efficiency.

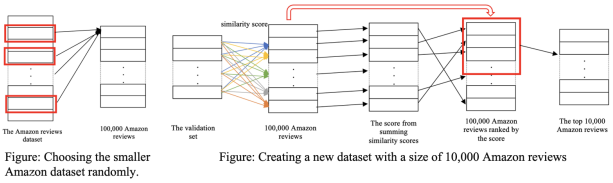

Figure: Choosing the smaller Amazon dataset randomly.

Figure: Creating a new dataset with a size of 10,000 Amazon reviews

Fig. 3. Creating a new dataset from Amazon reviews

For the additional Amazon dataset in Model 2, a subset of 100,000 reviews is randomly selected after omitting extraneous information. The SentenceTransformer is employed to map sentences into a 768-dimensional dense vector space. This facilitates applications such as clustering or determining sentence similarity using cosine similarity.[13] To create a new dataset, a similarity score is computed between the embeddings of the validation set and the 100,000 Amazon reviews. Reviews are ranked based on similarity scores, and the top 10,000 reviews are selected. This facilitates applications such as clustering or determining sentence similarity using cosine similarity. Employing the model trained on the smaller subset of the original dataset, these selected reviews are subsequently labeled based on predictions. However, it's worth noting that only approximately 500 reviews are predicted with label 0 out of the initial 10,000 reviews. To address this imbalance, I extended the dataset by including more reviews with label 0 from the next top 20,000 reviews, aiming to achieve a more balanced distribution, as illustrated in the figure below. Then, this dataset will be divided into two parts: 80% for the training dataset and 20% for the validation set.

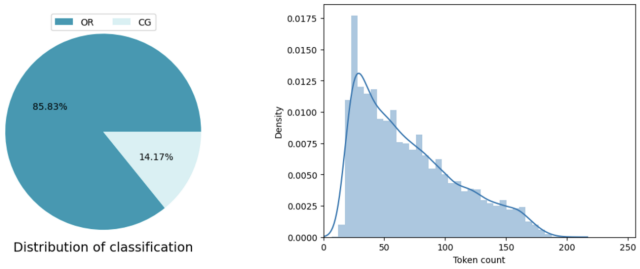

Fig. 4. The distribution of labels and review lengths labelled Amazon review dataset

In the labeled Amazon review dataset, the distribution of review lengths shows a distinct peak around 30 tokens, with no reviews extending beyond 200 tokens. However, considering the p-value of the distribution, it is determined that a maximum length of 160 tokens per review is deemed suitable for the model.

Datasets and DataLoaders play pivotal roles in facilitating efficient data loading and iteration for machine learning tasks, specifically deep learning. As system limitations are a consideration, batch loading is optimized to reduce computational time and memory usage. Additionally, the incorporation of Hugging Face's AutoTokenizer further enhances the efficiency of the text tokenization process. This tool dynamically selects the most appropriate tokenizer for the specified pre-trained model, eliminating the need for manual intervention and ensuring seamless compatibility with various architectures.

V. Tool and Technology

A. Software/Libraries

- Python: Primary programming language. - PyTorch: A deep learning framework facilitating the implementation of neural networks. - HuggingFace Transformers: Enables the utilization of pre-trained models, such as BERT. - SentenceTransformers: A library providing sentence embeddings, beneficial for tasks involving text similarity and representation. - Torchinfo: Offers detailed neural network model summaries, aiding in model architecture understanding and debugging. - Scikit-learn: A versatile machine learning library offering robust tools for data preprocessing, model building, and evaluation. - Pandas: Facilitates efficient data manipulation and analysis through its DataFrame structure. - NumPy: Fundamental for numerical operations and arrays. - Seaborn and Matplotlib: Data visualization libraries utilized to create insightful graphs and plots. - Google Colab: An interactive environment for coding and analysis.

B. Hardware

Cloud-based services: Utilizing services such as Google Colab for access to high-performance GPUs in the cloud.

VI. Experiments

A. Training details

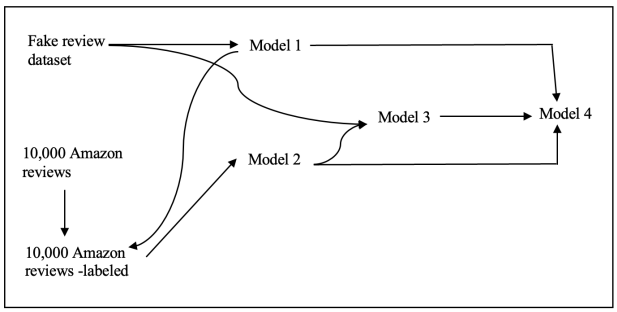

Fig. 5. A process map of experiments[14]

1) Model Architecture: Utilized a modified version of the BERT architecture from the Hugging Face Transformers library, specifically the AutoModelForSequenceClassification on 'bert-base-uncased'. The model comprises around 110 million parameters and is a bidirectional transformer pre-trained using masked language modeling and next sentence prediction on a large corpus involving the Toronto Book Corpus and Wikipedia. The primary focus of this project revolves around the masked language modeling objective.

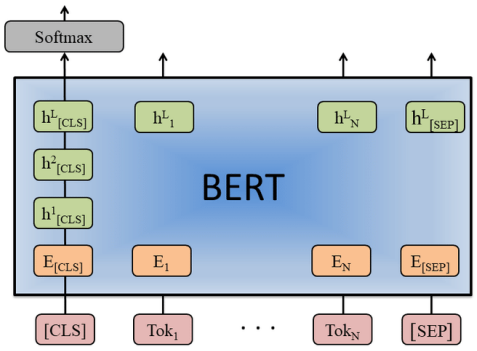

Fig. 6. Bert Model Architecture

2) Model 1: For Model 1, a training dataset of 40,000 reviews was employed, with 20,000 labeled as genuine product reviews ('OR') and 20,000 as fake reviews ('CG'). The preprocessing involved tokenization, padding, and encoding using Dataset and DataLoader with a batch size of 64. The training process incorporated an AdamW optimizer with an initial learning rate of 2e-5, coupled with a linear schedule that dynamically decreased the learning rate to 0 after a warm-up period.[15] The model underwent training for 5 epochs, consuming over 1 hour on a T4 GPU in Google Colab, utilizing a single GPU with approximately 9.5 GB memory. Cross-entropy loss was employed during training to optimize the model's performance

3) Model 2: Model 2 adopted a similar approach to Model 1, employing an optimizer and a learning rate scheduler for fine-tuning. However, it differed in utilizing the Amazon dataset as the training dataset, with a maximum review length set at 160 and a batch size of 64. After 5 epochs, the model required around 25 minutes and consumed approximately 10 GB of GPU memory.

4) Model 3: For Model 3, utilizing a fine-tuned BERT model (specifically Model 2) served as the pre-trained model on the original dataset as the training dataset. The fine-tuning process, with a max review length set to 160, a batch size of 32, an optimizer, and a learning rate scheduler, took around 1 hour and utilized approximately 6.7 GB of GPU memory over 5 epochs.

5) Model 4: Model 4 was designed with the objective of improving overall prediction accuracy by harnessing the collective decision-making capabilities of individual models. The ultimate prediction was established through a majority vote among the three models. However, due to the suboptimal performance of Model 2, which could potentially impact Model 4, an additional iteration of Model 4 was introduced. This revised version solely incorporates insights from Models 1 and 3, aiming to address the limitations associated with Model 2 and further refine the ensemble model.

B. Evaluation Method

Evaluation metrics included accuracy, precision, recall, and F1-score. The validation process involved an 80% training, 10% validation, and 10% test split for the original dataset, while the Amazon dataset adhered to an 80% training and 20% validation split. All models underwent testing on a separate test set, evaluating their performance based on the test set of the original dataset.

VII. Result

A. Model 1

The training process was conducted over five epochs, progressively refining the model's performance. Initially, remarkable accuracy and low loss were achieved, displaying a robust learning trend. However, in later epochs, although the accuracy continued to increase, the validation loss showed some fluctuations, indicating potential overfitting.

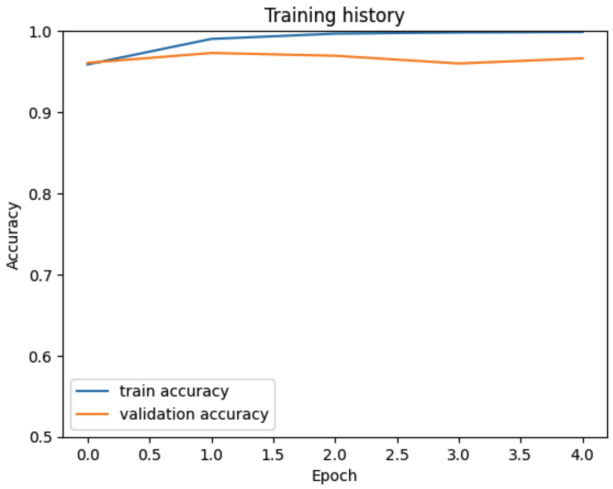

Fig. 7. Plot of the accuracy of Model 1 on training and validation set through 5 epochs

The validation set accuracies demonstrated consistency in the model's ability to generalize well to unseen data. However, epochs 4 and 5 showcased a rise in validation loss, potentially indicating a deviation from generalization due to overfitting. Test Set Evaluation: The test set evaluation confirmed the model's robustness, attaining an overall accuracy of 98%. The precision, recall, and F1-score metrics all indicate strong performance for both classes, 'CG' and 'OR,' demonstrating the model's ability to effectively distinguish between fake and genuine reviews.

	precision	recall	f1-score	support
CG	0.96	0.99	0.98	2017
OR	0.99	0.96	0.97	2027
accuracy			0.98	4044
macro avg	0.98	0.98	0.98	4044
weighted avg	0.98	0.98	0.98	4044

Fig. 8. The precision, recall, and F1-score metrics

On analyzing the data, the model demonstrated a high degree of accuracy and performed admirably in correctly identifying both groups. It exhibited a strong ability to correctly predict instances of class 'CG', with 1997 true positives and a minor misclassification of 20 instances as class 'OR'. Similarly, for class 'OR', the model's predictive capacity was notably accurate, correctly identifying 1947 instances, with a slight misclassification of 80 instances as class 'CG'. These results emphasize the model's proficiency in distinguishing between the two classes, reflecting strong precision and recall scores. The high accuracy, coupled with the precision and recall metrics, underlines the model's reliability in effectively identifying both class 'CG' and class 'OR' instances. This performance suggests a well-balanced model that can discern between the two classes with high precision, showcasing its efficacy in classification tasks.

B. Model 2

The training progression over the five epochs demonstrates a commendable improvement in the model's performance. Beginning with a solid 90.07% accuracy in the first epoch, the model consistently enhances its accuracy, reaching an impressive 99.70% by the final epoch. This positive trend in accuracy is reflected in the validation set, initiating at 88.89% in the first epoch and culminating at 92.27% by the end of training. The peak validation accuracy occurs at epoch 3, registering at 92.44%.

In terms of identifying computer-generated reviews ('CG'), the model achieves a precision of 0.77, denoting a 77% correctness when predicting a review as computer-generated. Notably, the recall for 'CG' is exceptionally high at 0.98, indicating the model effectively captures the majority of actual computer-generated reviews. The corresponding F1-score, at 0.86, offers a balanced assessment of the model's proficiency in correctly identifying computer-generated reviews.

The overall accuracy on the test set is 84%, representing the proportion of correctly classified reviews across both classes. While the model excels in identifying computer-generated reviews ('CG') with high precision and recall, there is room for improvement in the identification of human-generated reviews ('OR'), especially in increasing recall.

C. Model 3

The training of Model 3 unfolds over five epochs, revealing consistent improvements in performance across both the training and validation sets. In the initial epoch, the model achieves an accuracy of 96.78%, marking a robust starting point. This accuracy steadily increases in subsequent epochs, reaching an impressive 99.94% by the final epoch. Similarly, the validation accuracy, starting at a high 97.71%, rises to 97.84% by the end of training. However, the model's peak performance is observed at epoch 4, where it attains 98.10% accuracy.

The precision, recall, and F1-score metrics further underscore the model's exceptional performance. For both classes, 'CG' and 'OR,' the precision and recall scores are consistently high, resulting in F1-scores of 0.98. This balanced and high level of accuracy demonstrates the model's effectiveness in correctly classifying both computer-generated ('CG') and human-generated ('OR') reviews.

The confusion matrix provides additional insight into the model's predictive capabilities. With only a small number of misclassifications (16 'CG' reviews predicted as 'OR' and 82 'OR' reviews predicted as 'CG'), the model demonstrates robustness and reliability in distinguishing between the two classes. Consequently, this model outperforms the two previous models in detecting reviews generated by computers.

D. Model 4

The performance results of Model 4 - 1 showcase its remarkable accuracy, achieving an overall accuracy rate of 97%. The precision, recall, and F1-score metrics underscore the model's competence in effectively discerning between computer-generated (CG) and human-generated (OR) reviews.

Particularly noteworthy is the high recall of 0.99 for 'CG,' indicating the model's proficiency in capturing the majority of true computer-generated reviews. However, it's crucial to note that the precision scores, specifically 0.95 for 'CG,' highlight the model's accuracy when predicting this class. While Model 4-1 excels in detecting computer-generated reviews, there is a trade-off observed in terms of precision. The model tends to overpredict the 'CG' class, leading to a lower precision compared to Model 1 and Model 3. In summary, while Model 4-1 emerges as the top performer in identifying computer-generated reviews, its trade-off with precision suggests a careful consideration of the specific use case and priorities in model deployment.

Regarding Model 4-2, the results for the model reveal exceptional performance across various metrics. The precision scores for both classes, 'CG' and 'OR,' are impressively high, indicating that when the model predicts a review as computer-generated ('CG') or human-generated ('OR'), it is correct 98% and 99% of the time, respectively. The recall scores are also noteworthy, with 99% for 'CG' and 97% for 'OR,' indicating the model's ability to effectively capture the majority of actual instances of each class. The F1-scores for both classes are equally impressive at 0.98, reflecting a harmonious balance between precision and recall.

E. Comparison

	Accuracy	Precision	Recall	F1 score
Model 1	97.53%	96.15%	99.01%	97.56%
Model 2	84.25%	77.00%	97.77%	86.15%
Model 3	97.58%	96.06%	99.21%	97.60%
Model 4 - 1	97.20%	95.29%	99.30%	97.25%
Model 4 - 2	98.17%	97.50%	98.86%	98.17%

Fig. 9. The reports for various models in the "CG" label

In evaluating the performance of the models across multiple metrics, Model 4-2 stands out as the top performer, achieving the highest accuracy at 98.17%, precision at 97.50%, recall at 98.86%, and F1 score at 98.17%. This model demonstrates a balanced and robust performance, excelling in both precision and recall. Leveraging fine-tuning from Model 2, Model 3 emerges as a strong contender, showcasing remarkable accuracy, precision, and recall. Model 3 closely follows, showcasing remarkable accuracy (97.58%), precision (96.06%), recall (99.21%), and F1 score (97.60%). Models 1 and 4-1 exhibit consistent and strong results, surpassing 97% accuracy, precision, recall, and F1 score. The collective strength of individual models is harnessed in Model 4-1, resulting in an overall high performance, particularly in recall and F1 score. However, a discernible trade-off is observed in precision, indicating a tendency to overpredict positive instances. On the other hand, Model 2 lags behind with the lowest accuracy (84.25%), precision (77.00%), and recall (97.77%), resulting in a lower F1 score (86.15%). In conclusion, the model 4-2 achieves a high level of accuracy and effectiveness in distinguishing between computer-generated and human-generated reviews, making it a strong candidate for practical applications in fake review detection.

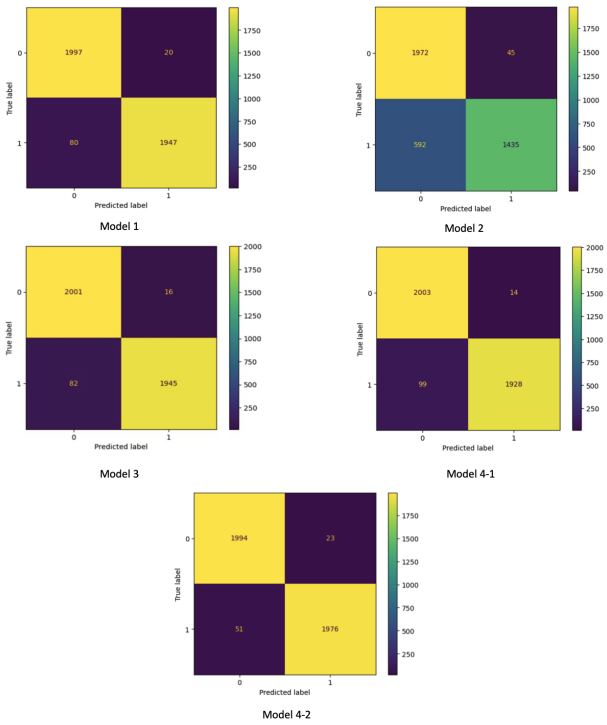

Fig. 10. The confusion matrix of 4 models on the test set

The collaborative power of individual models is effectively utilized in Model 4-1, contributing to an overall strong performance, especially in recall and F1 score. However, there is a noticeable trade-off in precision, suggesting a propensity to overpredict positive instances. Conversely, while Model 4-2 may not exhibit as high accuracy in detecting computer-generated reviews compared to Model 1, Model 3, and Model 4-1, it demonstrates noteworthy overall performance

VIII. Problem/Issues

During the preparatory phase for the experiments, significant hurdles were encountered while setting up the requisite libraries and packages. The foremost challenge revolved around seamless integration and compatibility issues among diverse libraries essential for the experiments. Notably, the project encountered a substantial setback primarily due to limited GPU memory constraints, particularly when utilizing the free Google Colab GPU. This memory constraint severely impeded the continuity of the project as it frequently led to unexpected disconnections and limitations in implementation time. The dataset, comprising over 40,000 reviews, proved significantly burdensome for the server, hindering smooth training due to its substantial size. Moreover, the complexity of the model itself contributed to memory wastage and accentuated the risk of overfitting, further exacerbating the memory limitations. The intricate nature of the model accentuated the challenge of managing memory constraints, posing a significant obstacle throughout the project's implementation.

Furthermore, the intent to implement Model 2 in the another labeled dataset to extract more information for developing Model 3 through transfer learning faced limitations in finding a suitable dataset. Consequently, the unlabeled Amazon dataset was utilized, and labels were predicted for this dataset. However, due to the inherent biases and limitations in the type of dataset used, particularly only relying on cosine similarity scores for evaluating reviews and identifying top reviews similar to the validation set in the Amazon dataset, Model 2's accuracy did not reach desired levels during training.

IX. Conclusion

The extensive Amazon reviews dataset provides not only scalability but also a realistic representation of diverse user opinions, contributing to the robustness of the model. Exposure to different datasets enables the extraction of features and patterns universally applicable across various scenarios. Training Model 2 on the Amazon reviews dataset aims to create a versatile model adept at handling the complexities of detecting computer-generated reviews in real-world situations. It's important to note that Model 2 exhibits the lowest accuracy among the four models due to its unique training approach. The incorporation of data based on cosine similarity to the validation set and predictions from Model 1 introduces a higher susceptibility to errors. This highlights the need to carefully consider the training approach's impact on model performance, indicating potential areas for refinement in future iterations.

Choosing the best model depends on specific task requirements. Model 1 and Model 3 excel across multiple metrics, while Model 2, despite its lower accuracy, performs well in recall. Model 4 offers a balanced performance, leveraging the strengths of individual models. The selection should align with the use case and priorities when deploying these models in practical applications.

In conclusion, this project rigorously evaluated the performance of various models in detecting fake reviews. Model 4-2 emerged as the top performer, exhibiting the highest accuracy, precision, and F1 score among the models considered. Its remarkable all-around performance, closely followed by Model 3, underscores the effectiveness of leveraging transfer learning and fine-tuning approaches. While Models 1 and 4-1 demonstrated consistently high scores, Model 2 lagged in accuracy and precision. The choice of the optimal model depends on the specific requirements of the task, considering factors such as overall accuracy, precision, recall, and F1 score. Despite challenges in memory constraints and dataset limitations, this project provides valuable insights and sets the stage for future enhancements. Refining model architectures, exploring diverse datasets, and fine-tuning hyperparameters could further optimize performance. Deploying these models in real-world scenarios and refining them based on evolving data and user feedback will be crucial for practical applicability. Overall, this project contributes significant findings to the evolving landscape of fake review detection.

References

[1] S. Chevalier, "Global retail e-commerce sales 2014-2026," https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/, 2022.

[2] E. Gregoire, "E-commerce poised to capture 41

[3] M. Keenan, "Global ecommerce statistics: Trends to guide your store in 2024," 2023, https://d1wqtxts1xzle7.cloudfront.net/84056081/A1301109119-libre.pdf?1649844210=&response-content-disposition=inline.

[4] a CommerceHub company, "Global retail ecommerce forecast," 2023, https://www.channeladvisor.com/resources/library-webinars/emarketer-global-retail-e-commerce-forecast/.

[5] Emplifi, "Emplifi reveals nearly 90

[6] S. J. Dixon, "Share of global fake online reviews removed 2021, by star rating," 2022, https://www.statista.com/statistics/1310797/global-fake-reviews-removed-by-star-rating/.

[7] S. He, B. Hollenbeck, G. Overgoor, D. Proserpio, and A. Tosyali, "Detecting fake-review buyers using network structure: Direct evidence from amazon," *Proceedings of the National Academy of Sciences*, vol. 119, no. 47, p. e2211932119, 2022. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2211932119

[8] W. Choi, K. Nam, M. Park, S. Yang, S. Hwang, and H. Oh, "Fake review identification and utility evaluation model using machine learning," *Frontiers in Artificial Intelligence*, vol. 5, 2023. [Online]. Available: https://doi.org/10.3389/frai.2022.1064371

[9] J. Lu, X. Zhan, G. Liu, X. Zhan, and X. Deng, "Bstc: A fake review detection model based on a pre-trained language model and convolutional neural network," *Electronics*, vol. 12, no. 10, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/10/2165

[10] J. Salminen, C. Kandpal, A. M. Kamel, S. gyo Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, 2022. [Online]. Available: https://doi.org/10.3389/frai.2022.1064371

[11] J. Salminen, "Fake reviews dataset," 2021, https://osf.io/tyue9/.

[12] A. Bittlingmayer, "Amazon reviews for sentiment analysis," https://www.kaggle.com/datasets/bittlingmayer/amazonreviews/data.

[13] Hugging-Face, "sentence-transformers/all-mpnet-base-v2," https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[14] P. Su and K. Vijay-Shanker, "Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism," https://www.researchgate.net/publication/345215818_Investigation_of_BERT_Model_on_Biomedical_Relation_Extraction_Based_on_Revised_Fine-tuning_Mechanism.

[15] Hugging-Face, "Fine-tune a pretrained model," https://huggingface.co/docs/transformers/training.