Exploring Transformer Models and Domain Adaptation for Detecting Opinion Spam in Reviews

Christopher G. Harris
University of Northern Colorado
Greeley, CO 80639 USA
christopher.harris@unco.edu

Abstract—**As online reviews play a crucial role in purchasing decisions, businesses are increasingly incentivized to generate positive reviews, sometimes resorting to fake reviews or opinion spam. Detecting opinion spam requires well-trained models, but obtaining annotated training data in the same domain (e.g., hotels) can be challenging. Transfer learning addresses this by leveraging training data from a similar domain (e.g., restaurants). This paper examines three popular transformer models—BERT, RoBERTa, and DistilBERT—to evaluate how training data from different domains, including imbalanced datasets, impacts Transformer model performance. Notably, our evaluation of hotel opinion spam detection achieved an AUC of 0.927 using RoBERTa trained on YelpChi restaurant data.**

I. INTRODUCTION

Businesses recognize that consumer-written reviews on websites like Amazon, TripAdvisor, and Yelp significantly influence purchase decisions. Positive reviews can increase hotel revenue, as shown by a Cornell University study, which found that a one-point increase in a hotel's online rating can boost booking probability by 11% [1]. This incentive has led some businesses to manipulate the system by posting fake reviews to either promote their products or discredit competitors.

This practice, known as *opinion spam*, has undermined the credibility of online review platforms. Surveys indicate that 82% of users encountered fake reviews for local businesses within the past year, and 74% of consumers say they can't always distinguish between genuine and fake reviews [2]. As a result, 75% of consumers have become distrustful of online reviews due to such spam [3]. This distrust can lead to adverse selection, where consumers struggle to differentiate between genuine and fake reviews.

Opinion spam is typically generated in two ways. The first involves hiring individuals to write fake reviews, often for products or services they have not experienced. The second method uses natural language processing (NLP) and deep learning (DL) techniques to generate fake reviews automatically. These automated systems can produce opinion spam cost-effectively and at scale and can be adjusted to evade advanced detection methods [4].

Detecting opinion spam is crucial as it ensures consumers receive accurate and honest information about products and services. Consumers rely heavily on reviews for purchasing decisions – nearly 95% of customers read online reviews before buying a product [5]; fake reviews can mislead them into buying subpar or misrepresented items. By identifying and removing opinion spam, platforms maintain the integrity of their information, helping consumers make informed choices and enhancing their overall shopping experience.

Additionally, detecting opinion spam helps businesses maintain their reputation and build customer trust. Honest businesses can suffer from negative opinion spam, which can unfairly damage their reputation and lead to a loss of customers. Conversely, competitors using fake positive reviews to inflate their ratings can gain an unfair advantage. By combating opinion spam, businesses ensure their reputation is based on genuine feedback, fostering a fair market environment. Moreover, robust opinion spam detection mechanisms uphold the credibility of review platforms, ensuring users continue to find value in the reviews and promoting long-term platform sustainability and growth.

The structure of this paper is as follows: In the next section, we introduce transfer learning and Transformer models and explain their effectiveness in addressing NLP challenges, such as detecting opinion spam. Section III reviews related work. Section IV details our experiments with different Transformer models and datasets, followed by a discussion of our findings and their implications in Section V. Finally, we conclude the paper in Section VI and outline potential future work in this area.

II. TRANSFER LEARNING AND TRANSFORMERS

Two recent developments have transformed the landscape of opinion spam detection – transfer learning and transformer models. We discuss them in turn below.

A. Transfer Learning

Often, we may have enough reviews in one domain (e.g., restaurants) but need to evaluate reviews in another domain (e.g., hotels) where resources are far more limited. In such cases, transfer learning is commonly used because it has been shown to be effective, efficient, and allows leverage of pre-existing knowledge.

Transfer learning is effective for detecting fake reviews because it can leverage pre-existing knowledge from large datasets [6-8]. Many DL models have already learned to understand complex language structures, context, syntax, and semantics. This provides a solid foundation for further fine-tuning on specific tasks such as fake review detection.

Another benefit of transfer learning is data efficiency. Collecting large, labeled datasets of fake reviews can be challenging, but transfer learning allows us to fine-tune a pre-trained model on a smaller dataset of labeled reviews. This reduces the amount of data needed and is more resource-efficient than training a model from scratch. The process saves computational costs and time.

Transfer learning models can be fine-tuned to adapt to the specific language and patterns found in reviews, enhancing their ability to distinguish between genuine and fake content. This adaptability allows them to generalize better across different types of reviews and writing styles. Additionally, transfer learning helps reduce overfitting, as the models start with a strong understanding of general language patterns, leading to improved accuracy when fine-tuned on specific datasets. This approach ensures robust and reliable detection of fake reviews in real-world applications.

How important is the domain used as a source during transfer learning? Recent work indicates that using a similar domain in a transfer learning task is important. Pan and Yang provide a comprehensive survey of transfer learning, including domain adaptation in [9]. This article discusses how transfer learning is more effective when the source and target domains are closely related. If the domains are dissimilar, negative transfer can occur, where the performance on the target task may degrade because the source domain knowledge is irrelevant. In [10], Blitzer et al. focused on domain adaptation in sentiment classification. They showed that models trained on one domain (e.g., movie reviews) perform better on a similar domain (e.g., product reviews) than on a dissimilar one. They found that the closer the source and target domains in terms of vocabulary and sentiment expressions, the better the transfer performance.

Ben-David et al. provide a theoretical analysis of domain adaptation in [11] and show that the effectiveness of transfer learning is highly dependent on the similarity between the source and target domains. They found that less divergence typically leads to better transfer. In this paper, we look at different domains and how they can assist with detecting opinion spam.

B. *Transformer Models*

Transformer models have revolutionized how machines understand and generate human language. Based on the Transformer architecture introduced in [12], these models have become the state-of-the-art method of addressing challenges in NLP, such as sentiment analysis and opinion spam detection, text classification, named entity recognition, and question-answering systems. Fig 1 shows the architecture of the Transformer model. We evaluate the performance of three transformer models on transfer learning tasks using the encoder portion of the model (i.e., the left side of Fig 1).

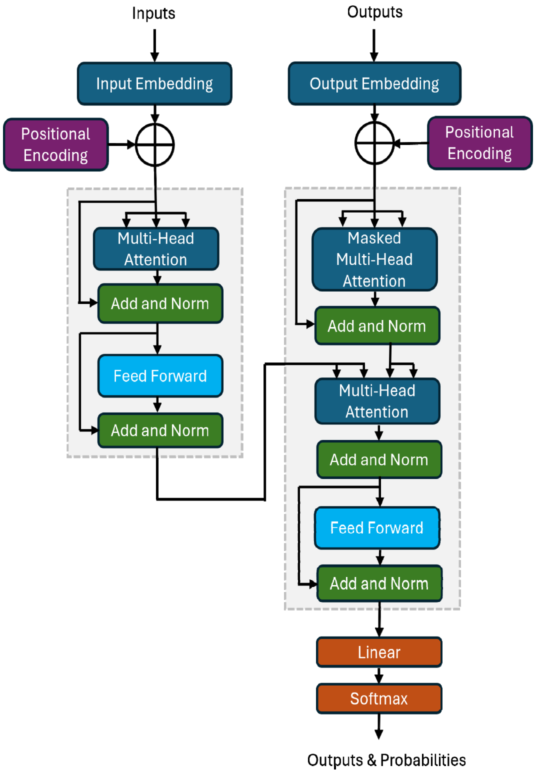


Fig 1. The Transformer Architecture (derived from [12]). The encoder portion of the model is on the left and the decoder portion is on the right.

BERT, introduced by Devlin et al. in [13], is one of the first models to apply bidirectional training to Transformer models, meaning it considers both the left and right context of a word simultaneously. This allows BERT to understand the meaning of words in context more effectively than previous models, which typically processed text in a unidirectional manner.

RoBERTa, introduced by Lin et al. in [14], is an optimized version of BERT. It builds on BERT's strengths by modifying the pretraining process to make the model even more powerful. RoBERTa is trained longer than BERT on more data, which improves its performance on downstream tasks. RoBERTa also uses larger batch sizes and higher learning rates during training, contributing to better generalization and performance. Last, While BERT uses a static masking strategy (where the same words are masked every time a sequence is processed), RoBERTa uses dynamic masking, meaning the masked words change during different iterations, leading to better learning.

DistilBERT, introduced by Sanh et al. in [15], is a smaller, faster, and more efficient version of BERT. It performs similarly to BERT while being more lightweight, making it suitable for applications with limited computational resources. DistilBERT is created using a technique called knowledge distillation, where a smaller model is trained to replicate the behavior of a larger model. In other words, DistilBERT is trained to mimic BERT's behavior. According to [15]. DistilBERT has 40% fewer parameters and is 60% faster than BERT, making it ideal for real-time applications and deployment on devices with limited processing power.

We explore the following research questions in this paper. Which of the three Transformer models performs best in a transfer learning task? Also, how important is domain similarity in transfer learning to detect opinion spam?

The structure of this paper is as follows: In the next section, we introduce transformer models and explain their effectiveness in addressing NLP challenges, such as detecting opinion spam. Section III covers the Yelp dataset, while Section IV reviews related work. Section V details our experiments with single-scale and multi-scale transformer models, followed by a discussion of our findings and their implications in Section VI. Finally, we conclude the paper in Section VII.

III. Related Work

Detecting opinion spam has emerged as a significant topic in natural language processing (NLP), with considerable research focused on identifying fake reviews and deceptive content. The pioneering work in this area was conducted by Jindal and Liu, as presented in their seminal paper [16]. They were the first to rigorously investigate the detection of opinion spam, employing supervised learning techniques that utilized a variety of features. These features were review-centric, such as unigrams and review length, and reviewer/product-centric, including metrics like sales rank and growth. This foundational study laid the groundwork for the ongoing exploration and development of methods for detecting opinion spam.

Machine learning techniques have proven to be highly effective in detecting opinion spam. These approaches encompass both traditional statistical methods—such as Naive Bayes, Random Forest, Support Vector Machines, and Ensemble models—and more recent advances in deep learning, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs). Mohawesh et al. [17] thoroughly overview these machine learning methods. Generally, these techniques can be categorized into three main groups: review-centric, reviewer-centric, or a hybrid approach that combines elements of both, as highlighted by Crawford et al. [18].

Review-centric techniques analyze individual reviews, examine linguistic patterns, part-of-speech (POS) tagging, N-grams, sentiment analysis, and quality indicators. However, these methods face challenges, such as vulnerability to rewritten spam and the dependency on labeled datasets. On the other hand, reviewer-centric approaches concentrate on the behavioral characteristics of reviewers, including review frequency, polarity, and text length. These methods allow for the identification of thresholds that can effectively distinguish between fake and genuine reviews.

While several studies have examined the linguistic characteristics of reviews, only a few have provided detailed methodologies that are beneficial for other researchers. For instance, Harris [19] evaluated the Kullback-Leibler (K-L) divergence on twelve linguistic features to assess the likelihood of reviews belonging to a fake or genuine dataset, specifically using YelpZip restaurant reviews. Heydari et al. [20] focused on temporal patterns in reviewer behavior, such as the gaps between reviews and "bursty" activity, while Hussain et al. [21] expanded their evaluation to include thirteen behavioral features combined with linguistic analysis on an Amazon.com dataset.

More recently, transformer models, which have remarkable capability in capturing language patterns, have become increasingly effective in detecting fake reviews. Minaee et al. provide an extensive review of text classification techniques in [22], while Duma et al. offer a comprehensive overview of various fake review detection methods in [23].

Among those exploring advanced encoding techniques, Refaeli and Hajek [24] experimented with different aspects of the BERT model for detecting opinion spam in YelpZip reviews. They achieved the best results by not freezing the model's weights and using 2-4 epochs during training, which helped to avoid overfitting. Their approach resulted in an accuracy of 0.73.

Hyder et al. [25] addressed the challenges associated with BERT, such as long training times, high computational resource requirements, and memory constraints. They proposed a model that leverages contextual representations to enhance the precision of deceptive review identification. Their model achieved an accuracy of 0.8108 on the YelpZip dataset and 0.8479 on a dataset of 359,000 reviews from YelpNYC.

Catelli et al. [26] also utilized BERT to detect opinion spam, focusing specifically on incorporating sentiment analysis as an input feature in the YelpNYC dataset. Their optimal model employed 12 transformer blocks (the hidden layers of the transformer encoder) and 12 attention heads. The maximum sequence length parameter, which defines the maximum size of the input vector, was set to 512 in their best-performing model. By using 90% of the data for training, they achieved an accuracy of 0.9375.

We could not find any previous studies in the literature that examined transfer learning on BERT models in the examination of detecting opinion spam.

IV. Experimental Methods

In this section, we describe the datasets we used to train our Transformer models, the dataset we used to evaluate opinion spam in hotel reviews, how we tuned our hyperparameters, and the metrics we used to evaluate our models.

A. Datasets Used

Ahmed et al. introduced a new dataset for fake news detection called ISOT [27]. This dataset was collected entirely from real-world sources and contains 44,898 news articles, with 21,417 (47.7%) articles collected from the Reuters news website and 23,481 (52.3%) fake news articles gathered from unreliable sources and flagged as dubious by fact-checking websites like Politifact and BuzzFeed.

Salminen et al. [28] produced a dataset of Amazon Reviews. This corpus contains 40,000 reviews, with half (20,000) legitimate reviews taken from Amazon for products in various categories and half computer-generated (i.e., fake).

The YelpChi dataset [29] contains 67,392 reviews from 200 restaurants in the Chicago area by 38,063 reviewers. This dataset contains 13.23% filtered (i.e., fake) reviews by 7,737 spammers, and is the least balanced dataset. YelpChi contains annotations generated based on Yelp's filtering algorithms, which flag reviews suspected to be fake or manipulated. The labels indicate whether a review was filtered (suspected fake) or unfiltered (genuine).

The Deceptive Online Spam Dataset (DOSC), developed by Ott et al. in [30], contains 400 truthful positive hotel reviews from TripAdvisor, 400 deceptive positive hotel reviews from Mechanical Turk, 400 truthful negative hotel reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp, and 400 deceptive negative hotel reviews from Mechanical Turk. These reviews are based on 20 hotels in the Chicago area.

Our objective is to test the ability to train three Transformer models (BERT, RoBERTa, DistilBERT) with the four datasets (ISOT news, Amazon Reviews, YelpChi, and DOSC), which arguably represent training datasets that are most divergent to most distant to our target dataset, which we call YelpSFO. YelpSFO contains 50,583 reviews from 233 hotels in San Francisco by 23,899 reviewers scraped in July 2024. It comprises 8.36% filtered (fake) reviews by 4,229 spammers. We determine a fake review as one flagged by the Yelp filter. This YelpSFO dataset is available at https://www.kaggle.com/datasets/chrisgharris/yelpsfo.

B. *Parameter Tuning*

For each of our four source datasets, we used similar parameters with BERT, RoBERTa, and DistilBERT: 12 Transformer blocks, 12 self-attention heads, and a hidden size of 768. We follow the optimization of RoBERTa and use AdamW [28] with β1 = 0.9, β2 = 0.98, $\varepsilon = 1e^{-6}$, weight decay of 0.01, dropout of 0.1, and attention dropout of 0.1. We used cased tokenization so that "RUDE SERVICE!" would be evaluated as different tokens than "rude service," allowing us to capture this difference through our self-attention heads. Following the approach in [24], we applied a classifier model without freezing the Transformer model layers. Although freezing the lower layers of the Transformer and training only the added classification layers can help retain the pre-trained knowledge, it can lead to overfitting; our preliminary testing showed good results without the need to freeze the lower layers.

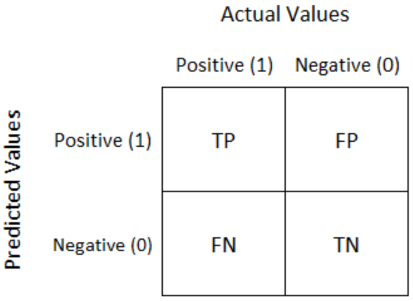
For the BERT, RoBERTa and DistilBERT Transformer models, we used the 'BertForSequenceClassification,' 'RobertaForSequenceClassification,' and 'DistilBertForSequence Classification' models, respectively, from Hugging Face's Transformers library (which adds a dropout and a 1-layer NN projecting 768 nodes directly to 2 on top of each model). We obtain the initial embeddings and logits for the hotel reviews.

C. *Metrics*

We wish to examine the capability of each model to discriminate between classes across all thresholds. AUC is the preferred choice as it offers distinct advantages over other metrics, such as accuracy and the F1 score.

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve plots the true positive rate (recall) against the false positive rate, providing a measure of the model's ability to discriminate between the positive and negative classes across all possible classification thresholds. AUC provides a comprehensive measure of model performance by considering both true positive and false positive rates across all thresholds. This makes it a more holistic measure, particularly in scenarios where the costs of false positives and false negatives vary, or where the balance between precision and recall is crucial. The AUC score is particularly useful for imbalanced datasets such as ours because it evaluates the model's performance over a range of thresholds, offering a more comprehensive view of its ability to distinguish between classes.

The ROC curve is a plot of the true positive rate (TPR, or sensitivity) against the false positive rate (FPR, or 1-specificity) at various threshold settings. Fig 2 provides a confusion matrix from which the TPR and FPR are calculated.



Predicted Values		Actual Values	
		Positive (1)	Negative (0)
	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig 2. Confusion Matrix

True Positive Rate (TPR): Also known as sensitivity or recall, it is calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

False Positive Rate (FPR): It is calculated as:

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

As the classification threshold is varied from 0 to 1, the TPR and FPR are recalculated, and these values are plotted to form the ROC curve.

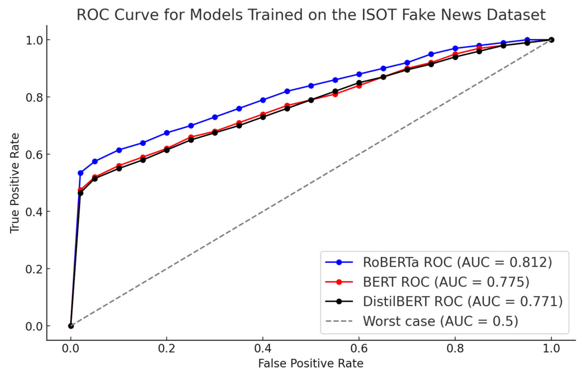


Fig 3. ROC Curve for ISOT News Dataset for our three Transformer Models

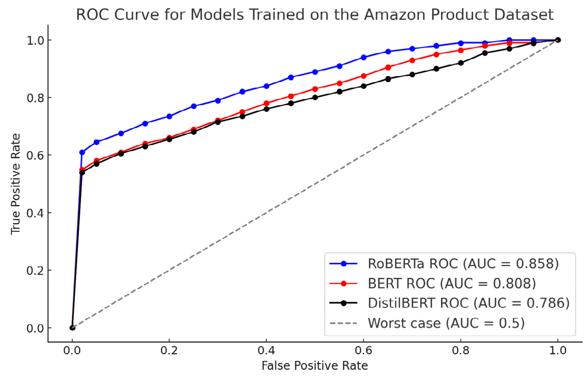


Fig 4. ROC Curve for Amazon Review Dataset for our three Transformer Models

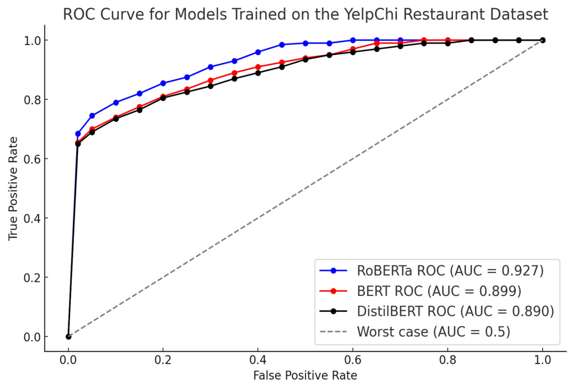


Fig 5. ROC Curve for YelpChi Dataset for our three Transformer

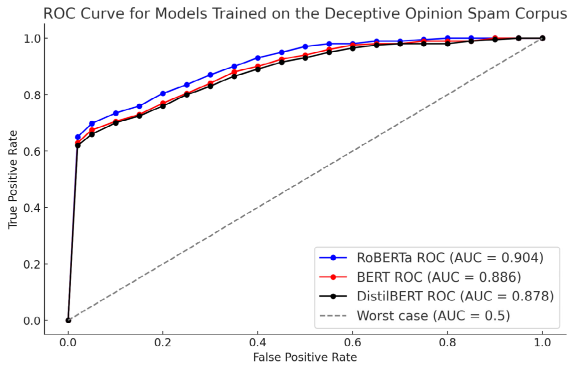


Fig 6. ROC Curve for the Deceptive Opinion Spam Corpus Dataset for our three Transformer Models

V. Results and Discussion

We trained each model using 8 NVIDIA V100 GPUs with 32 GB of memory. Larger models required more training time, taking from 15 minutes for training DistilBERT with DOSC to just under five hours to train BERT with ISOT. As with the authors' claims, DistilBERT was the fastest to train, and BERT was the slowest, but the training time was not significantly different between Transformer models. We describe the results between models trained on each dataset below.

A. *ISOT Dataset*

The result showing the AUC for the three Transformer models is shown in Fig 3. We can observe that the best-performing model was RoBERTa, with an AUC of 0.812. There was only a slight difference between BERT and DistilBERT.

B. *Amazon Review Dataset*

The result showing the AUC for the three Transformer models trained on the Amazon Review dataset is shown in Fig 4. Once again, we can observe that the best-performing model was RoBERTa, with an AUC of 0.858. As with the ISOT News dataset, there was only a slight difference between BERT and DistilBERT.

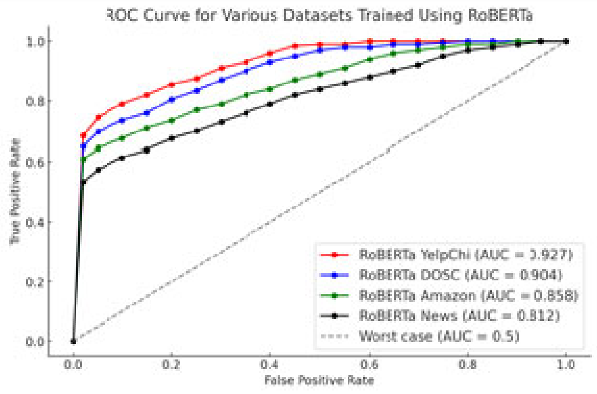


Fig 7. ROC Curve for various datasets using the RoBERTa

C. *YelpChi Dataset*

The result showing the AUC for the three Transformer models trained on the YelpChi dataset is shown in Fig 5. RoBERTa provided the best results with YelpChi, with an AUC of 0.927. BERT and DistilBERT provided a slightly lower AUC.

D. *Deceptive Opinion Spam Corpus Dataset*

The result showing the AUC for the three Transformer models trained on the Deceptive Opinion Spam Corpus

(DOSC) dataset is shown in Fig 6. As with the previous training datasets, RoBERTa performed best, with an AUC of 0.904. As with the previous training datasets, BERT and DistilBERT performed slightly below RoBERTa.

E. Comparing All Datasets Using RoBERTa

The result showing the AUC for all of the training datasets for one of the Transformer models (RoBERTa) is shown in Fig 7. YelpChi was the training dataset that provided the best AUC, and the ISOT News dataset provided the lowest AUC.

F. Discussion on Transformer Models

Examining Figs 3-6, we can observe that BERT and DistilBERT trail RoBERTa in the ability to discriminate between classes across all training sets, primarily due to its optimized training approach and improved handling of contextual information.

Unlike BERT and DistilBERT, RoBERTa uses dynamic masking during pretraining - the masking pattern changes with every input sequence. This results in a more robust understanding of the context because the model sees different masking patterns for the same text across different epochs, leading to a better grasp of how different words contribute to the meaning of a review and a stronger ability to discern a fake review. In the context of detecting opinion spam, where writing patterns differ between real and fake reviews, the masked words in each type of review would have noticeably different probabilities of being predicted. RoBERTa’s dynamic masking can take advantage of this difference.

Additionally, RoBERTa was initially trained on a much larger dataset than BERT (and DistilBERT), covering more diverse text sources. This broader training allows RoBERTa to generalize better across different domains, including opinion spam detection.

RoBERTa optimizes several key hyperparameters, including the learning rate, batch size, and training steps, to enhance the pretraining process. By using larger batch sizes, RoBERTa allows for more efficient gradient updates, which helps the model learn from more data per iteration. Additionally, by adjusting the learning rate to an optimal range, RoBERTa ensures smoother and more stable training, preventing issues like overfitting or underfitting. These optimizations also include extending the training time and removing the Next Sentence Prediction (NSP) task, allowing the model to focus on deeper context understanding, focusing solely on the masked language modeling (MLM) task. This omission allows RoBERTa to allocate more capacity to understanding single sentences in-depth, which is crucial for detecting opinion spam where the context within a single review is often sufficient to determine its authenticity.

RoBERTa can adapt more quickly and effectively to new tasks during fine-tuning. This makes it particularly valuable for transfer learning scenarios, where the model needs to be fine-tuned on specific tasks like detecting opinion spam with limited domain-specific data.

While DistilBERT is faster and more lightweight, it sacrifices some accuracy for efficiency. DistilBERT compresses BERT, leading to a model that is about 40% smaller but at the cost of some performance. For tasks like opinion spam detection, where subtle nuances matter, RoBERTa's enhanced training and full model capacity likely provide better detection capabilities than DistilBERT.

G. Discussion on Training Datasets

Our findings align with previous literature, showing that training datasets more closely related to the target dataset typically yield better model performance. However, an unexpected outcome emerged when comparing the YelpChi and DOSC datasets. We initially anticipated that the DOSC dataset, being more balanced between positive and negative classes, would produce a superior model. Contrary to our expectations, the YelpChi dataset outperformed DOSC despite being the only imbalanced dataset among the four used for training. This imbalance could have led to overfitting due to fewer negative class (opinion spam) examples, but it did not.

We believe several factors contribute to this surprising result. First, Transformer models, which are pre-trained on extensive language data, seem to mitigate overfitting by design, even when trained on imbalanced datasets. This can be done in BERT-based models through several techniques, such as class weights adjustment, data resampling, threshold adjustment, using focal loss, or ensemble learning methods. For the YelpChi dataset, we used the second technique and applied SMOTE [32], a well-known method to generate synthetic examples for the underrepresented class. Second, despite being a collection of restaurant reviews, the YelpChi dataset is inherently more aligned with our test data, YelpSFO, as both rely on Yelp's filtering system for ground truth, ensuring closer contextual relevance. Third, the DOSC dataset, created by crowdworkers, may not accurately reflect real-world spam generation techniques, which typically involve more sophisticated methods such as those seen in YelpChi, where spam is generated by human spammers and deep learning (DL) models. Finally, the larger YelpChi dataset provides a more comprehensive semantic understanding of fake and genuine reviews, further enhancing the model's learning capabilities. In future research, we plan to investigate whether these findings hold true across other domains and datasets.

VI. Conclusion

In this paper, we explored strategies to enhance the accuracy of detecting opinion spam by evaluating several Transformer models and analyzing how domain adaptation impacts the transfer learning process. Our research addresses a pressing issue in the digital landscape: the rise of opinion spam, which not only misleads consumers but also disrupts fair competition among businesses and undermines trust in online platforms.

We tested three popular Transformer models—BERT, RoBERTa, and DistilBERT—on the task of opinion spam detection and found that while all performed well, RoBERTa consistently outperformed the others. This superior performance can be attributed to several refinements in RoBERTa's architecture, including using a larger training dataset, optimized hyperparameters, dynamic masking during pretraining, and an enhanced capacity for fine-tuning in transfer learning scenarios.

To further understand the impact of domain adaptation, we trained these models on four distinct datasets: fake news detection, fake Amazon reviews, fake Yelp restaurant reviews, and crowdsourced hotel reviews. Interestingly, while one might expect the hotel review dataset to yield the best results for detecting hotel-related opinion spam, our models achieved better performance with Yelp restaurant reviews. This unexpected outcome is likely due to several factors: the similar filtering mechanisms employed by both the YelpChi dataset and our test set, the larger size of the YelpChi dataset, and the possibility that YelpChi reviews were generated using more advanced techniques, such as deep learning models, rather than relying solely on human crowdsourcing. These elements contributed to developing stronger and more accurate opinion spam detection models.

Our approach not only helps businesses protect the authenticity of their online presence but also shields consumers from misleading information. By detecting and filtering fraudulent reviews, we contribute to preserving the credibility of online marketplaces and review platforms. This, in turn, fosters a more transparent and trustworthy digital environment, enabling consumers to make informed decisions based on genuine feedback while promoting fair competition in the marketplace.

REFERENCES

[1] C. Anderson, "The impact of social media on lodging performance," Cornell Hospitality Report, vol. 12, no. 15, pp. 6-11, 2024. Web: http://scholarship.sha.cornell.edu/chrpubs/5.

[2] "WiserNotify," Web: https://wisernotify.com/blog/fake-review-stats/.

[3] D. MacRae, "Study underlines consumers' concern over fake reviews,", Jan. 10, 2024. Web: https://www.marketingtechnews.net/news/2024/jan/10/study-underlines-consumers-concern-over-fake-reviews/.

[4] Wu J AI-generated fake reviews: a new challenge for online trust. Medium. Web: https://bootcamp.uxdesign.cc/ai-generated-fake-reviews-a-new-challenge-for-online-trust-87e6ed825a80.

[5] L. Zhou, "Online Review Statistics: The Definitive List (2024 Data)," Marketing, April 13, 2024. Web: https://luisazhou.com/blog/online-review-statistics.

[6] A. Q. Mir, F. Y. Khan, and M. A. Chishti, "Online fake review detection using supervised machine learning and BERT model," arXiv preprint arXiv:2301.03225, 2023.

[7] M. Puttarattanamanee, L. Boongasame, and K. Thammarak, "A Comparative Study of Sentiment Analysis Methods for Detecting Fake Reviews in E-Commerce," HighTech and Innovation Journal, vol. 4, no. 2, pp. 349-363, 2023.

[8] R. Mohawesh, H. B. Salameh, Y. Jararweh, M. Alkhalaileh, and S. Maqsood, "Fake review detection using transformer-based enhanced LSTM and RoBERTa," International Journal of Cognitive Computing in Engineering, vol. 5, pp. 250-258, 2024.

[9] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.

[10] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-boxes, and Blenders: Domain adaptation for sentiment classification," in Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), 2007, pp. 440-447.

[11] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al., "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems* (NIPS), 2007, pp. 137-144.

[12] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[14] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019. Web: https://arxiv.org/abs/1907.11692

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS, 2019.

[16] N. Jindal and B. Liu, "Opinion Spam and Analysis," in Proceedings of the 2007 International Conference on Web Search and Data Mining (WSDM '07), 2007, pp. 219-230.

[17] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," IEEE Access, vol. 9, pp. 65771-65802, 2021.

[18] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," Journal of Big Data, vol. 2, pp. 1-24, 2015.

[19] C. G. Harris, "Detecting fraudulent online Yelp reviews using K-L divergence and linguistic features," in iSCSi 2022 Conference Proceedings, 2022.

[20] A. Heydari, M. Tavakoli, and N. Salim, "Detection of fake opinions using time series," Expert Syst. Appl., vol. 58, pp. 83–92, 2016.

[21] N. Hussain, H. T. Mirza, I. Hussain, F. Iqbal, and I. Memon, "Spam review detection using the linguistic and spammer behavioral methods," IEEE Access, vol. 8, pp. 53801-53816, 2020.

[22] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: a comprehensive review," ACM Computing Surveys (CSUR), vol. 54, no. 3, pp. 1-40, 2021.

[23] R. A. Duma, Z. Niu, A. S. Nyamawe, et al., "Fake review detection techniques, issues, and future research directions: a literature review," Knowl. Inf. Syst., 2024. [Online]. Available: https://doi.org/10.1007/s10115-024-02118-2

[24] D. Refaeli and P. Hajek, "Detecting fake online reviews using fine-tuned BERT," in Proc of the 2021 5th Intl Conf on E-Business and Internet, 2021, pp. 76-80.

[25] S. B. Hyder, N. Tariq, S. A. Moqurrab, M. Ashraf, J. Yoo, and G. Srivastava, "BERT-Based Deceptive Review Detection in Social Media: Introducing DeceptiveBERT," IEEE Transact on Comp Social Systems, 2024.

[26] R. Catelli, H. Fujita, G. De Pietro, and M. Esposito, "Deceptive reviews and sentiment polarity: effective link by exploiting BERT," Expert Systems with Applications, vol. 209, p. 118290, 2022..

[27] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First Intl Conf, ISDDC 2017, Vancouver, BC, Canada, Proc 1, Springer 2017.

[28] J. Salminen, C. Kandpal, A. M. Kamel, S. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102771, 2022. doi: 10.1016/j.jretconser.2021.102771.

[29] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp fake review filter might be doing?" in Proceedings of the International Conference on Web and Social Media (ICWSM), 2013.

[30] M. Ott, C. Cardie, and J.T. Hancock. 2013. Negative Deceptive Opinion Spam. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv preprint arXiv:1711.05101, 2017.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.