Robust Fake Review Detection using Uncertainty Aware NLP Classification Models

Sarah Zabeen*
School of Data and Sciences
Brac University
Dhaka, Bangladesh
sarah.zabeen@g.bracu.ac.bd

Alina Hasan*
School of Data and Sciences
Brac University
Dhaka, Bangladesh
alina.hasan@g.bracu.ac.bd

Md. Farhadul Islam
School of Data and Sciences
Brac University
Dhaka, Bangladesh
md.farhadul.islam@g.bracu.ac.bd

Md Sabbir Hossain
School of Data and Sciences
Brac University
Dhaka, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

Annajiat Alim Rasel
School of Data and Sciences
Brac University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—In a web-based world driven by e-commerce, customers are quick to turn to online shopping services. However, the products available for purchase cannot be personally inspected so buyers turn to on online product reviews. Potential consumers trust these reviews and are likely to spend more at stores with good evaluations. Sellers are well aware of this phenomenon and are not averse to using such unethical methods to boost the reputation of their own products or plunge that of a rival. In other words, anyone is capable of faking an opinion, which can heavily affect consumer purchasing decisions and business profits. It is not uncommon for companies to hire people to post fake reviews online. Scammers and competitors may even deploy bots to flood review sections with spam and mislead consumers into buying unreliable products. In order to counter this phenomenon of fake reviews, many fake review detection models have been extensively explored in the last decade yet there is still a lack of surveys that analyze and summarize these approaches.**

***Index Terms*—Monte Carlo Dropout, Fake Review Classification, NLP, LSTM, BERT, Reliability**

I. Introduction

Customers can now express their opinions and reviews on many websites in the era of the internet. These reviews can benefit both organizations and potential customers by providing insight into products or services before making a purchase. There has been a considerable increase in the number of consumer evaluations recently. Such assessments have a substantial impact on the choices of potential buyers. Essentially, customer reviews on social media platforms significantly impact their decision to purchase or not, making such reviews a valuable service for individuals.

Reviews authored by individuals inexperienced with the product or service are known as fake reviews. Consequently, an individual who posts these fake reviews is a spammer. A group of collaborating spammers with a common objective is a collective of spammers.

*These authors contributed equally to this work.

Many algorithms have been explored for fake review detection, Monte Carlo Dropout (MCD) is one such algorithm employed for uncertainty analysis in risk-free review evaluation. MCD trains a neural network with regular dropout during testing, allowing multiple predictions to be generated for each instance. For classification tasks, the softmax output can be averaged for each class.

II. Literature Review

Fake reviews are typically recognized using NLP approaches that prioritize textual data and lexical features, like keywords, n-grams, and linguistic style indicators [1]–[3]. Relevant non-textual features, like the user ID and their location, the number of reviews, and suspicious behaviors, are also considered [4]. Features can comprise both textual and non-textual features as discussed by [5]. Their combination generally improves detection, as evidenced in classification tasks [6]. In contrast to heuristic and behavior-based approaches, the methodology proposes an alternate strategy for detecting fake reviews.

Fake news can be classified using social context features retrieved from Twitter data using specific terms or based on news content [7]. In addition, users retweeting, the time difference between retweets, the retweet rate, and user comments provide important social context and text content features. A research study had several fake news classification models that rely on news content including the text-CNN, HAN, RST, and LIWC compared. Some models only depend on social context including the HPA-BLSTM, while others which include the CSI and TCNN-URG utilize the news content alongside. The study proposed a better-performing model called dEFEND, which includes a prediction component, a co-attention layer, and multiple encoders. [8] found that their proposed model, dEFEND, performed better than other models with an accuracy of 80.8% with an F1 score of 0.755 and an accuracy of 90.4% with an F1 score of 0.928 on the GossipCop and

PolitiFact datasets respectively. The authors also noted that the removal of either the user comments or the co-attention for news content resulted in a drop in accuracy, indicating that user comments are essential in guiding fake news detection in dEFEND. [9] proposed the use of SVM as the classifier unit and both RNN and GRU for user comments, sentence, and word encoding. The model produced an accuracy of 80.2% with an F1 score of 0.762, and an accuracy of 91.2% with an F1 score of 0.932 on the GossipCop and PolitiFact datasets respectively. However, it also relied on user comments.

BERT has gained significant attention for fake news classification. For improved learning, one method combined the BERT model with three parallel 1d-CNN blocks with changing kernel-size convolutional layers. This framework outperformed previously established deep learning models, demonstrating the importance of output features from BERT [10]. Another proposed using three BERT models for metadata, justifications, and statements on the LIAR PLUS dataset, achieving an accuracy of 74%. Similarly, a double-BERT model achieved an accuracy of 72% on the LIAR dataset [11]. In a related study involving fake review detection, sentence embeddings were generated by combining word embeddings to improve the classification models' understanding of word relationships, resulting in an increase in accuracy from 80% to 87% using the SVM classifier [12].

Additionally, in a sarcasm detection study, BERT was compared to deep learning models using GloVe embeddings, and BERT outperformed, demonstrating its superior ability to learn contextual aspects from data [13].

A study compared the accuracy of multiple machine learning algorithms which attempted to detect fake reviews from the Yelp dataset on restaurant reviews. Authors Elmogy et al. experimented with KNN (K=7), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression, and Random forest architectures incorporated with bi-gram and tri-gram language models. It was found that the F1 score of the KNN architecture outperformed the rest, at 82.3% without behavioural analysis input and 86.3% with it. Therefore the study concluded that feature engineering based on the reviewers' profiles would greatly enhance the performance of the system. [14]

III. Research Methodology

A. Classification Model

1) LSTM: LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images. This finds application in speech recognition, machine translation, etc. LSTM is a special kind of RNN, which shows outstanding performance on a large variety of problems.

2) BERT: The Bidirectional Encoder Representations from Transformers or, BERT is a powerful pre-trained language model developed by Google that employs a transformer-based bidirectional network architecture. It only uses the encoder part of the transformer to read input and understand the context of words in a text. However, before processing, BERT requires additional metadata specified in the input [12]. Once it receives the input, the model deconstructs it into individual tokens by transforming them into vectors which are subsequently fed to a neural network. The transformer is designed to map input vector sequences to output vector sequences, where each output vector represents a contextualized word embedding in the input text.

B. Uncertainty Analysis Method: Monte Carlo Dropout

Gal and Ghahramani [15] were the first to suggest Monte Carlo Dropout. They used it in evaluate deep Gaussian procedures by employing probabilistic Bayesian models. MCD can produce a series of predictions that depict uncertainty estimations of the experiment at hand. The MCD technique involves completing a number of stochastic forward passes in a Neural Network while also employing active dropout throughout the test stage.

The process of training a neural network model with the dropout f_{nn} can possibly estimate uncertainty for a given sample, x. This is done by accumulating all the forecasts of T interpretations which utilize numerous dropout masks. In particular, we observe a model with a dropout mask, d_i represented by $f_{nn}^{d_i}$. The subsequent equations demonstrate how the model produces results for a specific sample x:

$$f_{nn}^{d_0}(x), ..., f_{nn}^{d_T}(x) \tag{1}$$

An ensemble prediction can be obtained by calculating the mean and standard deviation. Here, the prediction is the sample average of the posterior probability distribution of the model, which estimates the model's uncertainty regarding x.

$$Predictive\, Posterior\, Mean,\, p = \frac{1}{T}\sum_{i=0}^{T} f_{nn}^{d_i}(x) \tag{2}$$

$$Uncertainty,\, c = \frac{1}{T}\sum_{i=0}^{T}[f_{nn}^{d_i}(x) - p]^2 \tag{3}$$

No changes are made to the dropout NN but the results of the stochastic forward passes are recorded. The predictive mean and model uncertainty is assessed with this method so that the information can be used with pre-existing dropout-trained NN models. This is done for overall uncertainty estimation and finding the list of the most uncertain samples.

IV. Experimental Analysis

A. Dataset

The 'Yelp Labelled Dataset: Spam Reviews for New York City' is a valuable resource consisting of 10,000 reviews that were extracted using the official Yelp API, a contribution of

Abid Meraj [16]. This dataset contains hotel reviews from North America.

Each review is labeled with one of five possible star ratings ranging from one to five stars, which is a significant feature. Another distinguishing aspect is the inclusion of a label value of 1 when a review is genuine and -1 when it is fake. This addition provides a more comprehensive view of customer feedback on Yelp and can be particularly useful for analyzing fake reviews and creating detection algorithms.

B. Experimental Setup

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue.

C. Performance Analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue.

D. Uncertainty Analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue.

V. Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue.

VI. Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue.

References

[1] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ser. ACLShort '09. USA: Association for Computational Linguistics, 2009, p. 309–312.

[2] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 219–230. [Online]. Available: https://doi.org/10.1145/1341531.1341560

[3] V. Sandulescu and M. Ester, "Detecting singleton review spammers using semantic similarity," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery, 2015, p. 971–976. [Online]. Available: https://doi.org/10.1145/2740908.2742570

[4] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, no. 1, 2021, pp. 409–418. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14389

[5] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, and et al., "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, vol. 2, p. 23, 2015.

[6] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," 2011.

[7] M. Mittal, I. Kaur, S. Chandra Pandey, A. Verma, and L. Mohan Goyal, "Opinion mining for the tweets in healthcare sector using fuzzy association rule," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 4, no. 16, p. e2, Oct. 2018. [Online]. Available: https://publications.eai.eu/index.php/phat/article/view/1280

[8] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 395–405. [Online]. Available: https://doi.org/10.1145/3292500.3330935

[9] M. Albahar, "A hybrid model for fake news detection: Leveraging news content and user comments in fake news," *IET Information Security*, vol. 15, no. 2, pp. 169–177, 2021. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ise2.12021

[10] R. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.

[11] D. Mehta, A. Dwivedi, A. Patra, and et al., "A transformer-based architecture for fake news classification," *Social Network Analysis and Mining*, vol. 11, p. 39, 2021.

[12] 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.

[13] C. I. Eke, A. A. Norman, and L. Shuib, "Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model," *IEEE Access*, vol. 9, pp. 48 501–48 518, 2021.

[14] A. M. Elmogy, U. Tariq, A. Mohammed, and A. Ibrahim, "Fake reviews detection using supervised machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120169

[15] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, p. 1050–1059.

[16] A. Meraj, "Yelp labelled dataset," https://www.kaggle.com/abidmeera/yelp-labelled-dataset, 2018.