Journal of Technology Innovations and Energy
https://doi.org/10.56556/jtie.v4i1.1163
ISSN: 2957-8809

RESEARCH ARTICLE

Detection of Deceptive Reviews Using Long Short-Term Memory (LSTM) and Deep Neural Network (DNN)

Md. Haidar Ali [1], Md. Mahbub Alam[1], Md. Zonaid Bin Ferdous[2], Nayan Kumar Datta[1], Md. Mahbub Alam[1], Subrata Saha[3], Osman Goni[1]

[1]Institute of Computer Science, Atomic Energy Research Establishment, Bangladesh Atomic Energy Commission
[2]ICT Division, Rajbari, Ministry of Information and Communication Technology, Bangladesh
[3]Institute of Electronics, Atomic Energy Research Establishment, Bangladesh Atomic Energy Commission

Corresponding Author: Md. Haidar Ali, haiderdiu@gmail.com
Received: 30 December, 2024, Accepted: 20 March, 2025, Published: 21 May, 2025

Abstract
Online reviews serve as social proof for potential customers, promoting confidence in businesses. The online marketplace is expanding rapidly on a global scale, with consumers increasingly relying on internet reviews. This influence is especially significant in the digital age, where customers often rely on the opinions of others to guide their purchasing decisions. As these reviews play a crucial role in shaping purchase decisions, some unethical companies are motivated to fabricate and distribute misleading evaluations. Deceptive reviews are fabricated evaluations produced with the intention of appearing real and misleading the consumers. Those deceptive reviews can be detected manually based on their patterns which are seen in their linguistic and psychological aspects. However, the deep learning techniques proposed outperform all conventional approaches and offer higher self-adaptability to extract the desired features implicitly. For the purpose of detecting false reviews, we have suggested a Deep Neural Network (DNN) based Deceptive Review Detection Model (DRDM) method.

Keywords: Deceptive reviews; Deep learning; Convolutional neural network; Word embedding; Unsupervised Learning; Long Short-Term Memory (LSTM)

Introduction

Online user-generated reviews for a wide range of goods and services have dramatically increased over the past few years across numerous websites. These evaluations include numerous in-depth information as well as the users' individual opinions. Before making decisions, we frequently refer to various user reviews, including where to eat, what to purchase, where to stay, and other decisions. Since these reviews play a vital role in online businesses, their importance is also increased. Untrustworthy businesses have a motivation and opportunity to create and post bogus reviews, either in support of themselves or in dislike of competitive competitors. As instances of deceptive reviews increase, the need for detection of those deceptive reviews increases. There are many research works which have been monitoring these types of fake reviews and creating models to identify those. It has been shown that nearly 95% of

consumers decided to purchase after reading online reviews and the product which has at least five reviews has a 270% greater possibility to be purchased than the commodities with no reviews. The lack of the gold standard dataset, however, is the primary issue with employing any classifier Alberto *et al.* (2015). To the best of our knowledge, there are very few publicly accessible gold standard datasets Islam *et al.* (2018), Rastogi *et al.* (2017). Other studies have employed synthetic data gathered from crowdsourcing websites or hand labeling Ali *et al.* (2023). However, as numerous studies have noted, it is a tedious and time-consuming effort to manually identify any genuine or fraudulent review to a reasonable degree Zhang *et al.* (2018). In a different labeling strategy, crowdsourcing websites like Amazon Mechanical Turk (AMT) are used to create synthetic data and its label. They don't, however, accurately reflect spam reviews. Furthermore, Turkers' work was not commendable due to a lack of effort and topic knowledge Saha *et al.* (2023). Additionally, probabilistic techniques like Unsupervised Bayesian approach Ren *et al.* (2017) and Hidden Markov models Alam *et al.* (2021) were used to address the labeled dataset problem. The goal of this project is to use Long Short-Term Memory (LSTM) and Deep Neural Network (DNN) networks to provide an efficient method for identifying fraudulent reviews on internet marketplaces. The study intends to improve the accuracy and dependability of differentiating phony evaluations from real ones by utilizing LSTM's capacity to extract contextual information and long-term dependencies in sequential text data. It handles issues like dataset imbalance and domain adaptation, assesses the LSTM model's performance against conventional machine learning techniques, and investigates important aspects like semantic, syntactic, and contextual patterns. The ultimate goal of the research is to address the ethical implications of such systems while offering a reliable and scalable solution for practical applications.

The rest of the paper is organized as follows. The Literature Review is discussed in Section 2. In Section 3, we introduce the Research Objectives. The methodology of the proposed model for spam review detection is discussed in section 4. Section 5 presents the findings and Data Analysis. Finally, Section 6 concludes the whole paper.

Literature Review

Researchers have put out a number of methods for detecting deceptive or dishonest reviews in recent years. The identification of spam in emails Jindal *et al.* (2008) and web texts Feng *et al.* (2012) has been extensively studied in the past. By comparing the language structures of truthful and dishonest reviews using the deception theory Akoglu *et al.* (2013) illustrated how difficult it is to identify dishonest evaluations based on their structural characteristics, such as lexical complexity. A supervised classifier (Logistic Regression) employing features based on review text, reviewer profiles, and product descriptions was proposed by Jindal *et al.* (2010). The majority of these studies concentrated on extracting the more detailed textual elements in order to enhance deception detection ability. However, the challenges of producing human-labelled data and the incapacity of hand-crafted features to capture non-local semantic information over a conversation prompted the development of a number of other approaches, such as those utilizing user behavioral elements and semi-supervised learning. By utilizing syntactic features from context-free-grammar parse trees, Feng *et al.* (2012) improved performance while using the syntactic stylometry technique. Jindal *et al.* (2008) proposed an unsupervised framework to

detect the spammers and spam reviews. They exploited the network effect between the products and its reviewers, for example, the spammers are mostly linked to a bad product with positive reviews and vice-versa. The network-based architecture assigned a score for review and reviewer upon which it was labelled as spam or real. Li *et al.* (2017) proposed a probability-based language modelling and Kull-back–Leibler (KL) divergence technique for fake review detection. They used the syntactical, lexical and stylistic features for model evaluation, extracted from the dataset of Amazon. They found that around 2% of the total consumer reviews in their dataset are fake. Wang *et al.* (2016) proposed an unsupervised approach "GSLDA" for group spamming detection in online review data. It works in two phases, first, it clusters the closely related group spammers into a small-sized reviewer cluster by adapting latent Dirichlet allocation (LDA) to the product review context and secondly, from each small-sized cluster, it extracts the high-suspicious spammer groups.

Research Method

Data preparation

To evaluate and educate an AI system, we need to work on datasets. Any model will take inputs from datasets and give output. Those outputs will be observed and correct outputs will be stored in the system. After preceding numerous data; system will be more accurate. Hence, a system will be developed.

Deceptive opinion dataset

It contains 400 real and 400 fake reviews of both positive and negative sentiments respectively of twenty individual hotels in Chicago Li *et al.* (2017).

YelpZip Dataset

The YelpZip dataset Fusilier *et al.* (2015) is made up of real-world reviews of hotels and restaurants that were taken as samples from Yelp and combined with almost accurate information provided by the Yelp review filter. YelpZip shows reviews of 5044 hotels from 260,277 users in different neighborhoods of New York.

Neural Network Models

There are numerous neural network models and architectures, each designed for specific tasks and applications. Here are some of the most well-known and widely used neural network models: Feedforward Neural Networks (FNN): Also known as Multi-layer Perceptron (MLP). Composed of an input layer, one or more hidden layers, and an output layer. Commonly used for tasks like classification and regression.
Convolutional Neural Networks (CNN): Primarily used for image-related tasks. Utilizes convolutional layers to automatically learn spatial hierarchies of features. Well-suited for tasks like image classification, object detection, and image generation. Recurrent Neural Networks (RNN): Designed to work with sequential data. Utilizes recurrent connections to maintain memory of past inputs. Often used in natural language processing (NLP) and time series analysis.

Word Embedding

A neural network is used in the popular natural language processing method known as Word2vec to learn word embeddings which are distributed representations of words. It is used in text analysis, language translation, text classification, and information retrieval. Similar words are clustered together in the vector space because these embeddings accurately represent a word's semantics. Continuous bag-of-words (CBOW) and skip-gram are the two primary model architectures used by Word2vec. While skip-gram is an unsupervised learning technique which predicts the surrounding words based on the context of the current word, CBOW predicts the current word based on the context of the surrounding words.

LSTM Model

Input Layer: The input to the LSTM model is a sequence of data points or tokens, such as words in a sentence or time steps in a time series. Each data point is represented as a feature vector. LSTM Units (Cells): The core building blocks of an LSTM model are the LSTM units, also known as cells. These cells maintain an internal state and are responsible for learning and remembering information over long sequences. An LSTM cell typically consists of three gates:

Forget Gate: It decides what information from the previous cell state should be thrown away or kept.

Input Gate: It decides what new information should be stored in the cell state.

Output Gate: It decides what information should be exposed as the output of the cell.

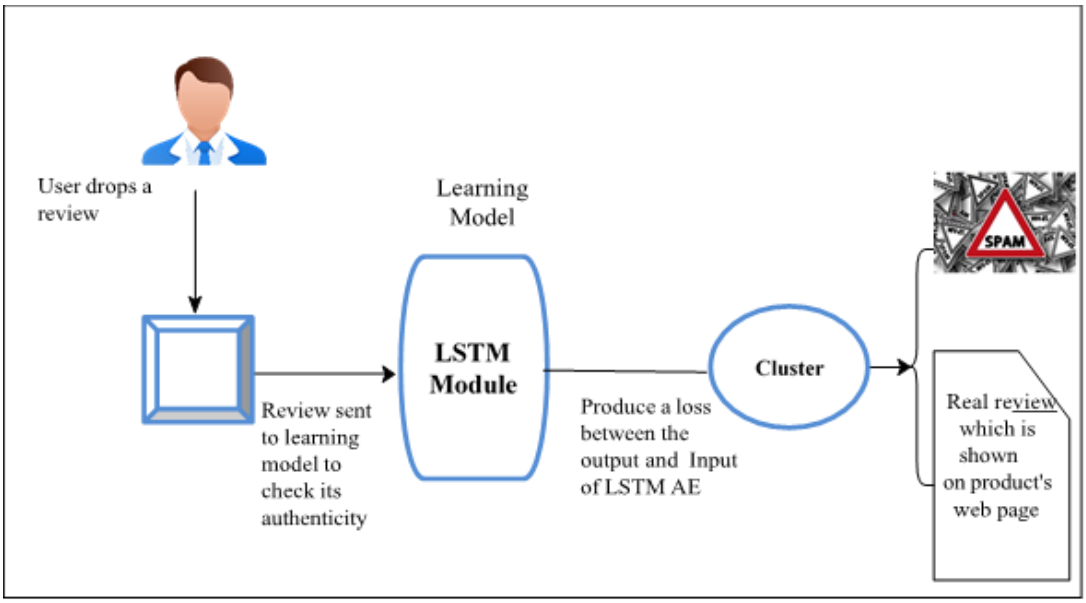

Figure 1: LSTM Architecture

Proposed Model

First Data Preparation is needed. Deceptive opinion dataset is chosen from sample dataset. Operation of preprocess the text data by tokenizing the reviews, removing stop words, and performing other text cleaning operations should be performed. Here, Word2Vec is chosen as a pre-trained word embedding model which can capture semantic information of words. Then loading of the pre-trained word embedding's into the model takes place. Conversion

of each review into a fixed-length representation using word embedding's should be done. This can be done by LSTM. After training the model with the proposed dataset; evaluation of the model's performance on the test set using metrics such as accuracy, precision, recall, F1-score is taken into account for preparing it for the real-world environment.

This diagram illustrates the flow of data through the neural network, where word embedding's are passed through an LSTM layer to capture the sequence information, followed by max pooling. Remember that in practice, one may need to fine-tune the architecture, hyper parameters, and data preprocessing steps to achieve the best results for your specific dataset and task. Additionally, techniques like data augmentation, assembling, and hyper parameter tuning to improve model performance can be done.

To understand the network structure we built in this article, a one layered and a multi-layered stacked LSTM network have been shown in Fig. 1 and 2 respectively. A one-layer architecture of LSTM model is incorporated by a single hidden LSTM layer (LSTM1) followed by an output layer whereas, the stacked architecture of LSTM model is incorporated by multiple hid- den LSTM layers (LSTM1 and LSTM2) (2 layers in this case). Similarly, a three- layered LSTM network can be formed by stacking LSTM1, LSTM2, and LSTM3. The stacking of layers adds levels of abstraction of input observations over time. LSTM1 generates a sequence output at each time step instead of single output at the final step. These sequential outputs per input time steps act as inputs for the LSTM2.

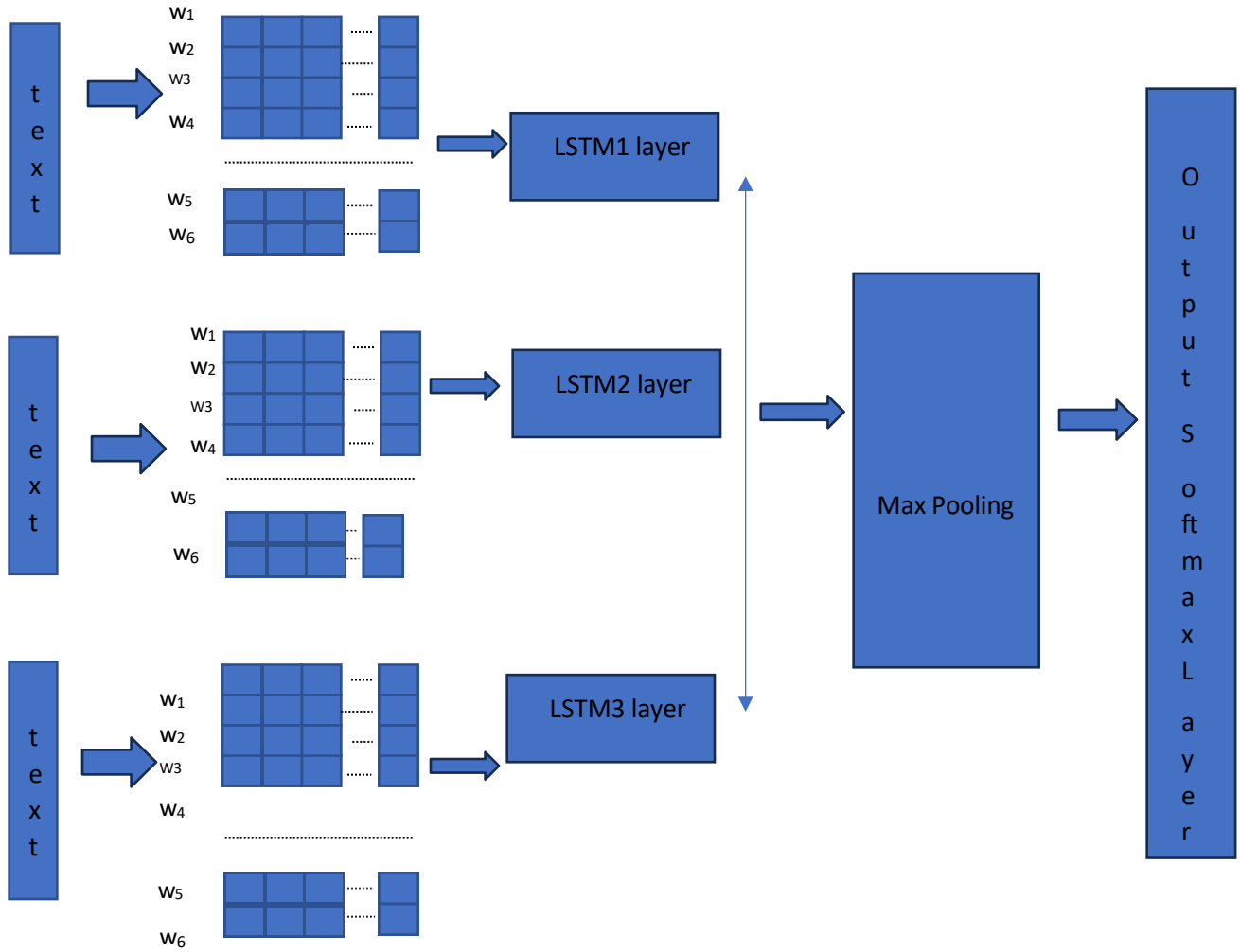

Figure 2: Deceptive Review Detection Model (DRDM)

Data Analysis

Evaluation of accuracy

Model	Accuracy	Precision	Recall	F1-Score
Baseline	0.85	0.90	0.80	0.85
DRDM	0.92	0.94	0.91	0.92
CNN-B	0.91	0.93	0.90	0.91
LSTM- C	0.89	0.92	0.88	0.90

Table 1: Accuracy Count

Here, we can see DRDM is slightly better than CNN-B and LSTM-C IN F1-score, accuracy, precision and recall.

Conclusion

Here, experiment is done to obtain an AI based model based on LSTM for determining deceptive reviews. Accuracy has been tested with various model with two datasets. This model will improve itself with the improvement of computer hardware. The text sequence will pass word embedding then LSTM layer. It will be modified with max pooling and it will be in output Soft Max layer. The proposed model will help to distinguish between deceptive reviews and the real ones. Detecting spam reviews (or comments) is a difficult undertaking. The lack of the labelled dataset is the primary problem in this particular area of study. To close this gap, we suggested an unsupervised method that predicts the review's class (spam or authentic) without the need for label information. To accomplish the goal, the current model employs an auto encoder and clustering. The duplicate review is not recognized as spam by the suggested model. This implies that the system does not consider a review to be spam if it is posted for many goods. A pre-filter is needed to find these reviews. In the future, a system that integrates the existing model in a pipeline with the duplicate review filter might be created. In order for a review to be deleted if it is discovered to be duplicate, it must first pass through a filter for duplicity identification; if not, it will go through the current system.

Declaration

Acknowledgment: This research is supported by the Institute of Computer Science, Atomic Energy Research Establishment, Bangladesh Atomic Energy Commission.

Funding: Not Applicable**.**

Conflict of interest: There are no conflicts of interest relevant to this work.

Ethics approval/declaration: Not Required.

Consent to participate: All the authors consent to participate in this research in accordance with ethical guidelines.

Consent for publication: All the authors have read and approved the final version of the manuscript and consent to its publication in [Journal of Technology Innovations and Energy].

Data availability: Open Access.

Authors contribution: All the Authors contributed equally to this work, including study design, data analysis, and manuscript preparation.

References

Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230, February 2008

Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Paper, vol. 2, pp. 171–175, July 2012.

Akoglu, L., Chandy, R., & Faloutsos, C. (2013). Opinion fraud detection in online reviews by network effects. *ICWSM*, *13*, 2–11.

Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015). Tubespam: Comment spam filtering on youtube. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)* (pp. 138–143). IEEE.

Jindal, N., Liu, B., & Lim, E. -P. (2010). Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1549–1552). ACM.

Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A., & Shao, J. (2017). Bimodal distribu- tion and co-bursting in review spam detection. In *Proceedings of the 26th international confer- ence on world wide web* (pp. 1063–1072). International World Wide Web Conferences Steering Committee.

Islam, Shamimul, Haidar Ali, Ahsan Habib, Nur Nobi, Mahbub Alam, and Dulal Hossain. "Threat minimization by design and deployment of secured networking model." *International Journal of Electronics and Information Engineering* 8, no. 2 (2018): 135-144.

Goni, Osman, Md Haidar Ali, Md Mahbub Alam Showrov, and Md Abu Shameem. "The basic concept of cyber crime." *Journal of Technology Innovations and Energy* 1, no. 2 (2022): 16-24.

Saha, Subrata, Md Shamimul Islam, Md Mahbub Alam, Md Motinur Rahman, Md Ziaul Hasan Majumder, Md Shah Alam, and M. Khalid Hossain. "Bengali Cyberbullying Detection in Social Media Using Machine Learning Algorithms." In *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pp. 1-6. IEEE, 2023.

Islam, Md Shamimul, Subrata Saha, Md Mahbub Alam, Nayan Kumar Datta, Md Haidar Ali, Md Dulal Hossain, and Md Golam Moazzam. "Natural Language Processing and Machine Learning Approaches to Detect Bangla Hate Speech on Social Media." In *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pp. 1-6. IEEE, 2023.

Saha, Subrata, Md Motinur Rahman, Tahmid Tamrin Suki, Md Mahbub Alam, Md Shah Alam, and Mohammod Abu Sayid Haque. "Heart Disease Prediction Using Machine Learning Algorithms: Performance Analysis." In *2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pp. 1-6. IEEE, 2024.

Alam, Mahbub, Osman Goni, Abu Shameem, Shamimul Islam, Nayan Kumar Datta, Shakil Ahmed, and Golam Moazzam. "An Approach for the Normalization of Short Message Service to Detect Shorter Form of Words and Find out Actual Word." *International Journal of Electronics and Information Engineering* 13, no. 3 (2021): 111-118.

Li, L., Qin, B., Ren, W., & Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. *Neurocomputing*, *254*, 33–41.

Rastogi, A., & Mehrotra, M. (2017). Opinion spam detection in online reviews. *Journal of Informa- tion & Knowledge Management*, *16*(04), 1750036.

Wang, Z., Hou, T., Song, D., Li, Z., & Kong, T. (2016). Detecting review spammer groups via bipartite graph projection. *The Computer Journal*, *59*(6), 861–874.

Zhang, W., Du, Y., Yoshida, T., & Wang, Q. (2018). DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Manage- ment*, *54*(4), 576–592.

Fusilier, D.H., Montes-y-Gómez, M., Rosso, P., Cabrera, R.G.: "Detecting positive and negative deceptive opinions using PU-learning." Inf. Process. Manag. 51, 433–443 (2015).

Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S.: Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In: Proceedings of the 2018 International Conference on Digital Health, pp. 121–125, April 2018.

Fontanarava, J., Pasi, G., Viviani, M.: Feature analysis for fake review detection through supervised classification. In: IEEE International Conference on Data Science and Advanced Analytics, pp. 658–666, October 2017.

Ren, Y., Ji, D.: Neural networks for deceptive opinion spam detection: an empirical study. Inf. Sci. 385, 213–224 (2017).

Kolhar, M. (2018). E-commerce review system to detect false reviews. *Science and Engineering Ethics*, *24*, 1577–1588. https://doi.org/10.1007/s11948-017-9959-2.

Ali, Haidar, Mohammad Zonaid Bin Ferdous, Imran Hossain Showrov, Osman Goni, Mahbub Alam, and Abu Shameem. "Cyber Security: Challenges, Threats and Protective Measures of an Organization." *International Journal of Electronics and Information Engineering* 15, no. 1 (2023): 1-11.

Saha, Subrata, Md Motinur Rahman, and Md Mahbub Alam. "Machine Learning Approach to Classify Twitter Hate Speech." *Machine Learning* 7, no. 5 (2023).

Rayana, S., Akoglu, L.: Collective opinion spam detection: bridging review networks and metadata. In: Proceedings of the 21 thacmsigkdd International Conference on Knowledge Discovery and Data Mining, pp. 985–994, August 2015