International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deep Learning Based Fake Review Detection System

Aymon Ahmed[1], B M Akhilesh[2], Prajwal S Prakash[3], Pranav R[4], Dr. K. B. Shivakumar[5]

[1234]Department of Computer Science and Engineering, Don Bosco Institute of Technology, Visvesvaraya Technological University, Bengaluru, India
[5]Professor and Head of Department, Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India

---

Abstract : Online reviews strongly affect purchase decisions, which is making them a target for fake and tricky reviews that are misleading consumers and warp product ratings, to a significant degree; the trained model is delivered through a website and a browser extension that are checking reviews as people are reading them, broadly speaking. Published detection studies use diverse datasets, attributes and evaluation protocols, so their reported precisions are hard to compare; this paper benchmarks four classic classifiers (Naïve Bayes, Logistic Regression, Linear SVM and Random Forest) on TF-IDF bigram traits, a Bidirectional LSTM (Bi-LSTM), Graph Convolutional Networks (GCN) and a fine-tuned RoBERTa-base model under one controlled protocol on two corpora, like the Cornell hotel tricky opinion spam corpus (1,600 reviews) and a balanced Amazon fake reviews dataset (about 40,432 reviews). On the Cornell corpus, TF-IDF with a Linear SVM is reaching 87.50% accuracy, close to the 87.81% reported for a BERT-embedding SVM [1]; on the Amazon dataset, Linear SVM is reaching 91.34%, Bi-LSTM 91.52% and a 2-layer GCN 91.21%. On Amazon the Bi-LSTM and GCN differ from the best standard model by less than 0.2 percentage points, so a tuned linear model remains a strong baseline there; indeed, before, fine-tuning RoBERTa-base with 5-fold stratified cross-validation is reaching 93.06% ± 0.72%.

Index Terms, Fake review detection, misleading opinion spam, TF-IDF, Bi-LSTM, graph convolutional network, RoBERTa, cross-validation, in real-world scenarios.

I. INTRODUCTION

Shoppers are, broadly speaking, counting on online reviews and ratings when choosing products and services; indeed, fake reviews can, in practice, therefore mislead buyers, distort product ratings and reduce trust in online marketplaces, in many respects. Rule-based approaches struggle with the complicated linguistic patterns employed in modern spam reviews, which motivates automated, data-driven detection, on the whole.

Three families of methods are widespread. Reviewer-centric methods use behavioural signals such as review count, rating pattern and timing, and are usable only when the dataset carries reviewer metadata; these methods are feeding BERT representations to standard classifiers, as verified out by Mir et al. [1], and content-based methods train supervised classifiers on the language of tricky reviews employing attributes such as TF-IDF and n-grams, in actual application.

The rest of this paper is organised as follows; so, Section III is outlining the datasets, traits, models, validation protocol and rollout, on the whole, and Section II reviews related work. Section IV reports findings, Section V is examining them and their limits, and Section VI concludes, on the whole.

A. Problem Statement

Existing studies use distinct datasets, attributes and evaluation protocols, so their reported precisions can't be checked directly, to a significant degree. There's a call for a framework that benchmarks classifiers fairly, under one protocol and reports false positives, since flagging a genuine review as fake unfairly penalises a real user, to a significant degree.

B. Objectives

Under one controlled protocol, 1) Benchmark Naïve Bayes, Logistic Regression, Linear SVM and Random Forest, employing the same stratified split and preprocessing for every model, in actual application.

2) Build a Bi-LSTM, graph-based models and a fine-tuned RoBERTa-base model, and gauge them with stratified k-fold cross-validation, broadly speaking.

3) Compare the findings with published work on the Cornell corpus [1], [2], stating where datasets and methods differ, in many respects.

4) Deliver the trained model as a website and a browser extension that return a forecasted class with a confidence score, in actual application.

C. Scope and Contribution

The contributions are: (i) a controlled benchmark of lexical, sequential, graph-based and fine-tuned transformer models on two corpora; (ii) a comparison with the published BERT-based [1] and Naïve Bayes [2] results, which shows how sensitive reported accuracy is to implementation choices; and (iii) a working website and browser extension built on the trained model. RoBERTa-base was checked on the Cornell corpus only.

II. RELATED WORK

The section first is reviewing four sources on fake review detection and closely related modelling [1], [4]: two detection studies [1], [2], a survey [3] and a hybrid sentiment model [4]; it then is covering the foundations, datasets and model families that this work is often building on [5], [11], broadly speaking.

A. Reviewed Studies

Mir et al. [1] are combining BERT embeddings with supervised classifiers (SVM, Random Forest and Naïve Bayes), in many respects; their insight is generally better semantic understanding and detection accuracy. The shortcoming they carry is currently high computational cost; indeed, on the Cornell corpus, the values employed for comparison in this paper are 87.81% for BERT with SVM and 78.43% for BERT with Naïve Bayes, in actual application.

TF-IDF, Bag-of-Words and n-gram attributes with SVM, Naïve Bayes and Logistic Regression is being used by dilsha and Lijo [2], and show that textual and opinion traits detect fake reviews basically, in real-world scenarios; their shortcoming is weak contextual understanding. The Naïve Bayes accuracy of 89.50% from this work is the comparison point employed here, broadly speaking.

The survey doesn't implement. Mohawesh et al. [3] survey machine learning, NLP and deep learning approaches to fake

review detection and give an overview of schemes and datasets.

Tan et al. [4] are suggesting RoBERTa-LSTM, a hybrid of a transformer encoder and a recurrent network, for sentiment analysis, and report high precision, and isn't a fake-review detector, in real-world scenarios; in fact, though, it's pertinent as a transformer-plus-recurrent design. Its check is restricted to particular datasets.

TABLE I
SUMMARY OF REVIEWED SOURCES

Ref.	Method	Contribution	Limitation
[1]	BERT embeddings + SVM, RF, NB	Better semantic understanding	High computational cost
[2]	TF-IDF, BoW, n-grams + SVM, NB, LR	Text and sentiment features are detecting fakes	Restricted contextual understanding
[3]	Survey of ML, NLP, DL	Overview of methods and datasets	No implementation
[4]	RoBERTa + LSTM hybrid	High exactness (sentiment task)	Limited to particular datasets

B. Foundations and Datasets

Jindal and Liu [5] found opinion spam is widespread and diverse in nature from e-mail or Web spam, on the whole; ott et al. [6] inspected tricky opinion spam, meaning fictitious opinions written to sound authentic, and created the first large-scale public gold-standard dataset for it, with truthful and deliberately misleading hotel reviews, broadly speaking. While, human judges ran at roughly chance level, their best classifier was nearly 90% precise. Ott et al. [7] expanded this corpus with 400 misleading negative reviews of Chicago hotels, so that the corpus is covering both positive and negative feeling; the 1,600-review corpus employed here (800 truthful, 800 tricky) is of this kind.

A second kind of fake review is machine-generated. Salminen et al. [8] employed GPT-2 to fakely review from an Amazon e-commerce dataset and built a detection dataset from them, in actual application; in fact, they report that a machine classifier detects them near-perfectly, and meanwhile, much lower precision and agreement are shown by human raters, in real-world scenarios. Two distinct threats, like human-written deception [6], [7] and machine-generated text [8] are therefore embodied by the two corpora used in this paper.

C. Model Families Used in This Work

RoBERTa [9] is a BERT type with a revised pre-training procedure and is the base model that's fine-tuned here, in real-world scenarios. My point is, graph convolutional networks (GCN) [10] are mainly multilayer networks that operate directly on a graph and compute node embeddings from the features of each node's neighbourhood, on the whole. Text GCN [11] is employing this to text category by building a single graph for a whole corpus from document-word relations and word co-occurrence, with words and documents as nodes, so that classifying documents becomes node grouping, in real-world scenarios. None of the four sources in Section II-A applies graph models to fake review detection, in real-world scenarios.

D. Common Themes

Let's get going up up, review text is the primary signal in both detection studies [1], [2], on the whole; in fact, coming up, SVM and Naïve Bayes recur as the classifiers of choice [1], [2], in many respects. Moving on, representations are, of course, moving from lexical features [2] toward the contextual embeddings [1] and hybrid transformer-recurrent models [4], a trajectory also covered by the survey [3], and fourth, feeling is employed as an backup attribute [2].

E. Debates and Contradictions

The main debate concerns representation richness versus cost. Contextual embeddings improve semantic understanding but are computationally costly [1], while lexical features are cheap but lack context [2]. The reported numbers don't settle this. In the meantime, sVM with BERT embeddings is reaching 87.81% [1], and as, well, the same classifier family differently is resulting, including Naïve Bayes is usually 78.43% in the BERT setting of [1] and 89.50% in [2], in real-world scenarios. Seeing as the sources differ in traits and rules, and the survey is offering no implementation [3], these figures can't be read as a ranking of methods, to a significant degree. A second tension concerns the machine-generated reviews: [8] reports near-perfect machine detection; in turn, even so, the obstacle of the task clearly rests on the model and protocol, on the whole, but the best exactness reported here on that kind of data is 91.52%.

F. Research Gaps

1) No widespread protocol, namely none of the sources is comparing lexical, contextual, sequential and graph-based models on the same split and preprocessing, broadly speaking.
2) The cost of contextual models is typically named as a shortcoming [1], so though, the accuracy-versus-cost trade-off isn't quantified in these sources, in real-world scenarios.
3) The hybrid model [4] is checked on precise datasets, so transfer to fake review detection is normally untested here, in actual application.
4) Graph-based text classification [11] and false-positive reporting aren't tackled by the sources on fake review detection reviewed here.
5) Delivery of a trained detector to end users isn't covered: [3] is giving no implementation, in many respects.

G. Strongest Ideas

The ideas that this work is building on most directly are the use of contextual embeddings with standard classifiers [1], the TF-IDF with n-gram and sentiment pipeline [2], the survey taxonomy of ML, NLP and deep learning approaches [3], the combination of a transformer with a recurrent network [4], and the gold-standard deceptive-review corpus of Ott et al. [6], [7] that's making comparison possible, in real-world scenarios.

III. METHOD

A. System Architecture

The input module takes a labelled review dataset and, where available, rating, reviewer ID and timestamp; it's running preprocessing and feeds four model families, including TF-IDF classifiers, a Bi-LSTM, GCNs and a fine-tuned RoBERTa model, all validated with stratified cross-validation; this system returns the projected class (fake or genuine) with a confidence score, together with metrics and confusion matrices, on the whole. Three modules (Fig. 1) are had by the framework.

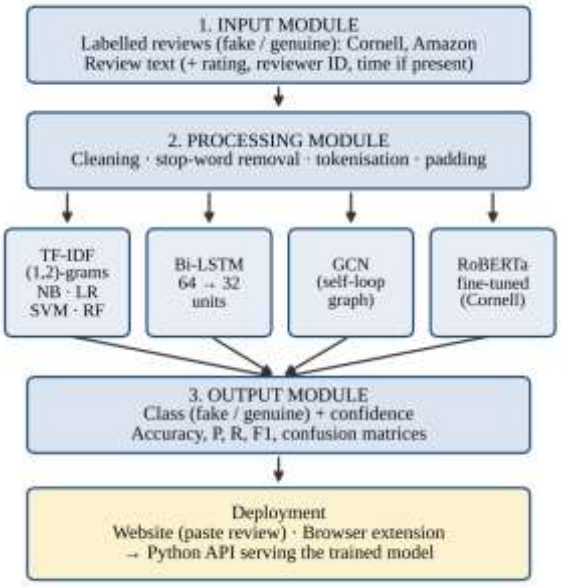


Fig. 1. Architecture of the fake review detection system.

B. Datasets

The Cornell Hotel Deceptive Opinion Spam Corpus [6], [7] has 1,600 balanced reviews (800 genuine and 800 deceptive) of Chicago hotels, in many respects. The two corpora define "fake" differently, as misleading text in Cornell and machine-generated text in Amazon, which matters when checking scores, across the them, broadly speaking. The Amazon Fake Reviews Dataset, introduced with the study of Salminen et al. [8], has about 40,432 balanced reviews (20,216 computer-generated fake reviews and 20,216 original customer reviews).

C. Preprocessing and Feature Extraction

Text is lower-cased, special characters, digits and punctuation are removed with the regular expression [^a-z\s], English stop-words are removed employing the NLTK stop-word list, and words shorter than three characters are dropped. The cleaned text is vectorised with TF-IDF over unigrams and bigrams (range (1, 2)), capped at 50,000 traits, with sublinear term-frequency scaling:

$$w(t, d) = (1 + \ln tf(t, d)) \times idf(t) \quad (1)$$

D. Classical Classifiers

Naïve Bayes, Logistic Regression, a Linear Support Vector Classifier and Random Forest are trained on the same stratified split with a fixed random seed, on the whole; indeed, fake is the positive class. TF-IDF is fitted on training data only, the test set is employed once; in fact, what's more, hyperparameters are tuned by cross-validation inside the training data. Precision, precision, recall, F1-score and confusion matrices are reported, with F1 set out as

$$F1 = 2PR / (P + R) \quad (2)$$

E. Bi-LSTM

The Bi-LSTM is reading each review forwards and backwards, so each word is seen with its left and right setting, in real-world scenarios; indeed, what's more, it learns word order that TF-IDF ignores. The layers are a 64-dimensional embedding, a Bi-LSTM with 64 units, a second Bi-LSTM with 32 units, a thick layer of 64 ReLU units with dropout 0.5, and a single sigmoid output (0 genuine, 1 fake), broadly speaking; and thus, the top 50,000 words are kept by the input, padded to 200 tokens.

F. Graph Convolutional Networks

Two versions are trained, like a standard 2-layer GCN and a deeper 3-layer GCN with 64 and 32 hidden dimensions, in actual application; indeed, GCNs [10] implemented in PyTorch Geometric operate on a word-document graph, in the spirit of Text GCN [11], so they're using structural relations between nodes rather than word order, in real-world scenarios.

G. RoBERTa Fine-Tuning

A pre-trained RoBERTa-base model [9] is fine-tuned end to end, so the whole encoder is changing to the vocabulary of misleading reviews, unlike the frozen BERT embeddings employed with classic classifiers in [1], and is found on the Cornell corpus (1,600 reviews) with 5-fold stratified cross-validation, on the whole, and precision and weighted F1-score are reported per fold, with their mean and standard departure, on the whole. The use of a transformer is, to be sure, motivated by the transformer-recurrent hybrid in [4].

H. Cross-Validation

Stratified k-fold cross-validation is used (5 folds for RoBERTa and for the Cornell Naïve Bayes comparison, and 5-fold and 6-fold runs for the other models) so that each fold is keeping the fake and genuine proportions, in real-world scenarios; indeed, per fold the tokenizer is fitted on the training folds only, the model is trained, in many respects. And it's tested on the held-out fold. Tuning is, as a practical matter, done inside the training folds so that it doesn't leak into the reported score; in fact, performance is summarised by the mean and standard departure across the folds; where a standard departure is reported below, it's the population standard departure over the folds.

I. Deployment

A Python backend service loads the trained model and returns the forecasted class and confidence, broadly speaking; indeed, a content script is being employed by the browser extension to read the reviews on a product page, sends their text to the API, and shows the number of fake reviews found (Fig. 2), to a significant degree. The website lets a user paste a review and shows the probability that it's fake.



(a) Website (b) Browser extension

Fig. 2. Rolled out framework: (a) website for pasted reviews and (b) browser extension reporting the number of fake reviews found on a page, in many respects.

IV. RESULTS

A. Cornell Hotel Corpus

Table II is listing the findings on the Cornell corpus with the published values employed for comparison, in real-world scenarios. Five-fold cross-validation of Naïve Bayes showed stable results across the splits; Linear SVM on TF-IDF (87.50%) is 0.31 points below the BERT-embedding SVM of [1] (87.81%), in many respects. Naïve Bayes on TF-IDF (86.56%) is 8.13 points above the BERT-embedding Naïve Bayes of [1] (78.43%) and 2.94 points below the Naïve Bayes of [2] (89.50%); and thus, Fine-tuned RoBERTa-base is clearly the strongest model on this corpus, at 93.06% ± 0.72% accuracy (Table III and Fig. 3), which is 5.56 points above our Linear SVM and 5.25 points above the BERT-embedding SVM of [1].

TABLE II
RESULTS ON THE CORNELL HOTEL CORPUS (%)

Model	Source	Accuracy	F1-score
Linear SVM (TF-IDF)	Ours	87.50	87.48
Logistic Regression	Ours	87.19	87.17
Naïve Bayes	Ours	86.56	86.47
Random Forest	Ours	82.50	82.42
RoBERTa-base (fine-tuned, 5-fold)	Ours	93.06 ± 0.72	93.04 ± 0.73
BERT + SVM	[1]	87.81	–
BERT + Naïve Bayes	[1]	78.43	–
Naïve Bayes	[2]	89.50	–

TABLE III
ROBERTA-BASE PER-FOLD RESULTS, CORNELL (%)

Fold	Accuracy	Weighted F1
1	93.44	93.42
2	92.81	92.80
3	94.06	94.05
4	91.88	91.85
5	93.12	93.10
Mean ± SD	93.06 ± 0.72	93.04 ± 0.73

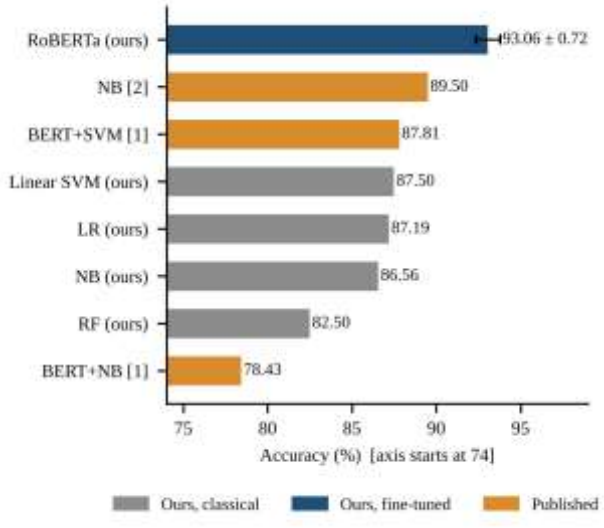


Fig. 3. Precision on the Cornell corpus. Error bar, such as standard departure over five folds (RoBERTa only).

B. Amazon Fake Reviews Dataset

Table IV and Fig. 4 find; so, the Bi-LSTM is reaching 91.52%, the 2-layer GCN 91.21% and the 3-layer GCN 91.00%, in real-world scenarios, and the standard ordering is the same as on Cornell. RoBERTa-base wasn't run on this dataset. Linear SVM is the best classic model (91.34%), ahead of Logistic Regression (90.10%), Naïve Bayes (89.89%) and Random Forest (86.90%), broadly speaking, and more, every standard model scores 2.9 to 4.4 points higher on Amazon than on Cornell.

TABLE IV
RESULTS ON THE AMAZON DATASET (%)

Model	Accuracy	F1-score
Linear SVM (TF-IDF)	91.34	91.34
Logistic Regression	90.10	90.09
Naïve Bayes	89.89	89.88
Random Forest	86.90	86.90
Bi-LSTM	91.52	–
GCN (2-layer)	91.21	–
GCN (3-layer, 64/32)	91.00	–

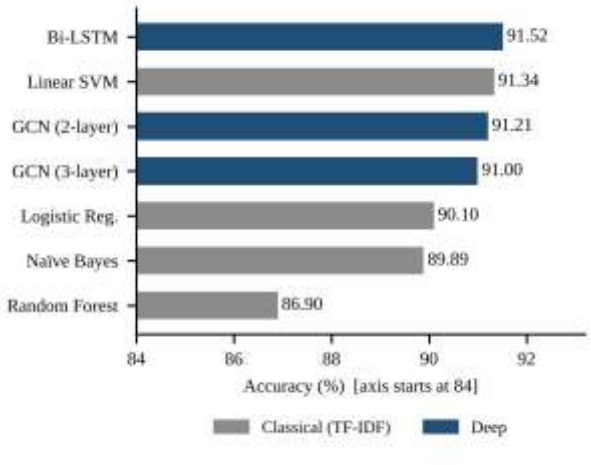


Fig. 4. Exactness on the Amazon dataset.

C. Deployment

The website and extension run end to end - the content script collects page reviews, the API returns a class and confidence, broadly speaking; no latency measurements were taken. What's more, the extension reports the count of fake reviews found.

V. DISCUSSION

A. Representation: Lexical, Embedded or Fine-Tuned

TF-IDF bigrams with a Linear SVM match a BERT-embedding SVM on Cornell, within 0.31 points [1], and on Amazon they're within 0.2 points of the Bi-LSTM and GCN, in actual application. Fine-tuning changes the picture, including RoBERTa-base is reaching 93.06% on Cornell, 5.25 points above the frozen-embedding SVM of [1]; this is proposing that the gain comes from fitting the whole encoder, not from contextual embeddings alone, to a significant degree; in fact, the uniform classic ordering on both corpora indicates thin high-dimensional attributes favour linear margin-based classifiers. The higher Amazon scores may reflect dataset size, but also the diverse notion of "fake" (machine-generated text), so they shouldn't be credited to size alone.

B. Sensitivity to Implementation

Naïve Bayes on the Cornell corpus ranges from 78.43% [1] to 86.56% (ours) to 89.50% [2], an 11.07-point spread for one classifier family on one corpus; this is supporting the problem statement: differences in traits, preprocessing and protocol can outweigh differences between algorithms. And published precisions aren't directly comparable.

C. Do Deep Models Help?

On Amazon, the Bi-LSTM directs Linear SVM by 0.18 points, so what's more, the 2-layer GCN trails it by 0.13; these gaps are small and no significance test was run, so these findings don't show that such models are better there, to a significant degree; On Cornell, the RoBERTa gain of 5.56 points is much larger than its 0.72-point fold-to-fold standard departure, broadly speaking. Only the same, it wasn't tested for significance, and the deeper GCN (91.00%) scored below the 2-layer GCN (91.21%), so extra depth didn't help. What's more, as well, what's more, its computational cost wasn't measured.

D. Limitations

Across the datasets, models weren't tested (train on one corpus, test on the other), broadly speaking. No significance tests were run, and fold-wise standard departures are reported only for RoBERTa, in real-world scenarios. F1-scores for the deep models aren't reported. Confusion matrices exist for the standard models, but false-positive and false-negative rates aren't analysed in this paper. RoBERTa-base was checked on Cornell only, with cross-validation, so its score isn't directly comparable to a single-split score, on the whole. The comparison with [1] and [2] employs their published figures; in turn, and the rules behind them weren't reproduced exactly, in actual application.

VI. CONCLUSION

This paper benchmarked standard, sequential, and graph-based models for fake review detection on a hotel corpus and a larger Amazon dataset, and delivered the trained model, through a website and a browser extension, in many respects; On Amazon, Bi-LSTM (91.52%) and GCN (91.21%) perform at the same level, within 0.2 points, in real-world scenarios. A TF-IDF bigram Linear SVM is a fast and exact baseline (87.50% on Cornell, 91.34% on Amazon), and the broad spread in published Naïve Bayes results is suggesting that comparisons need a typical protocol. Fine-tuned RoBERTa-base is typically reaching 93.06% ± 0.72% on Cornell, the best finding there.

Future work includes checking RoBERTa-base on the Amazon dataset, cross-dataset check, significance testing, listed false-positive assessment, and cost and latency measurements for the rolled out service, broadly speaking.

ACKNOWLEDGMENT

The authors thank Dr. K. B. Shivakumar and Professor of the Department of Computer Science and Engineering, for his guidance and support, and Don Bosco Institute of Technology for the facilities employed in this work, on the whole.

REFERENCES

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

[2] D. Dilsha and V. P. Lijo, “Opinion spam detection in online reviews using machine learning techniques,” *Int. J. Eng Res Technol. (IJERT)*, vol. 9, no. 6, 2020.

[3] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, and Y. Jararweh, “Fake reviews detection: A survey,” *IEEE Access*, vol. 9, pp. 65753-65769, 2021.

[4] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, “RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network,” *IEEE Access*, vol. 10, pp. 21517-21525, 2022.

[5] Y. Liu *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, 2019.