JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# IDENTIFYING FAKE ONLINE REVIEWS THROUGH SUPERVISED AND SEMI-SUPERVISED LEARNING

**[1]Prof.MOHAMMED ILYAS AHMED**
[1]Assistant Professor, [1]Dept. of ECE
[1]KCT Engineering College, Gulbarga, Karnataka

**[2]Prof. SHAIKH MOHAMMAD SAMIR**
[2]Assistant Professor, [2]Dept.of Computer Engineering
[2]JSPM's JSCOE, Hadapsar, Pune, Maharashtra, India

**[3]Prof. NILIMA MAHESH GHOLAP**
[3]Assistant Professor, [3]Dept.of Computer Engineering
[3]JSPM's JSCOE, Hadapsar, Pune, Maharashtra, India

***Abstract:*** Fake online reviews can mislead consumers and damage business reputations. This preliminary benchmark study examined whether adding unlabelled reviews through semi-supervised self-training improves the detection of deceptive hotel reviews when labelled data are scarce. Experiments used the Deceptive Opinion Spam Corpus v1.4, comprising 1,600 reviews from 20 Chicago hotels. A single test fold contained 320 reviews from four hotels excluded from training. Multinomial Naïve Bayes, linear support vector machine, and logistic regression served as supervised baselines. The primary comparison used the same initial labelled reviews and term frequency–inverse document frequency (TF–IDF) features for supervised logistic regression and its self-training counterpart. Four label budgets of 64, 128, 256, and 320 reviews were evaluated across ten draws per budget. At a confidence threshold of 0.65, self-training yielded lower mean macro-F1 at every label budget. With 320 labelled reviews, supervised logistic regression achieved 80.57% mean macro-F1, compared with 75.37% for self-training. At that budget, self-training increased deceptive-class recall from 85.69% to 88.50%, while precision declined from 77.87% to 70.88%. These results identify a trade-off under the tested configuration. Self-training detected a larger share of deceptive reviews, but a smaller share of its flagged reviews were deceptive.

***IndexTerms*** - Fake online reviews, deceptive opinion spam, supervised learning, semi-supervised learning, self-training, TF–IDF.

## I. INTRODUCTION

Online reviews help consumers compare hotels, restaurants, and other services before making a purchase or booking. In hospitality, prospective guests use reviews to assess aspects of a stay that are difficult to judge from a hotel's own description, such as room cleanliness, service quality, and value for money. The usefulness of this feedback depends on whether reviews reflect genuine experiences. Fabricated reviews can distort perceptions of a hotel by exaggerating its strengths or inventing negative experiences. Detecting such content is challenging because a fabricated account can use the same vocabulary, tone, and level of detail as a genuine review.

Fake review detection is commonly formulated as a text classification problem in which a model assigns each review a truthful or deceptive label. Early research by Jindal and Liu (2008) examined opinion spam in product reviews. Ott et al. (2011, 2013) developed benchmark datasets containing truthful and deceptive hotel reviews across positive and negative sentiment. These datasets support controlled comparisons of text classifiers. Their collection procedures also affect how results should be interpreted. The deceptive reviews were commissioned from crowdworkers, while truthful reviews were collected from review websites. A classifier may therefore learn language patterns associated with how the reviews were collected as well as patterns associated with deception.

Supervised models rely on labelled examples to learn these patterns. When reliable labels are scarce, semi-supervised learning offers a way to use the larger pool of unlabelled reviews. In confidence-based self-training, a classifier first learns from a labelled subset, then assigns provisional labels to unlabelled reviews. Reviews with sufficiently confident predictions are added to the training set, and the classifier is retrained. This process can expand the training set with informative examples, but it can also reinforce early errors if the classifier assigns incorrect labels with high confidence. The research question is therefore whether access to unlabelled reviews improves detection when the initial labelled examples and evaluation conditions are held constant.

This study addresses that question using the Deceptive Opinion Spam Corpus v1.4, which contains 1,600 reviews from 20 Chicago hotels. Multinomial Naïve Bayes, linear support vector machine, and logistic regression serve as supervised baselines. The primary comparison pairs supervised logistic regression with self-training based on the same classifier and term frequency–inverse document frequency (TF–IDF) features. Four label budgets, ranging from 64 to 320 reviews, are evaluated across ten draws of the labelled training subset. All models are tested on the same 320 reviews from four hotels excluded from training.

Macro-F1 is the primary metric, while deceptive-class precision and recall show how the methods trade off missed deceptive reviews against genuine reviews incorrectly flagged as deceptive.

### 1.1 Research Objectives

i. To establish supervised baselines for detecting deceptive hotel reviews using Multinomial Naïve Bayes, linear support vector machine, and logistic regression with TF–IDF features.
ii. To determine whether confidence-based self-training improves detection by comparing it with supervised logistic regression using identical initial labelled reviews and the same held-out test set.
iii. To assess how the availability of labelled data affects model performance by evaluating four label budgets of 64, 128, 256, and 320 reviews across ten draws of the labelled training subset.
iv. To evaluate performance on reviews from hotels excluded from training using macro-F1 as the primary metric, and to examine the trade-off between missed deceptive reviews and truthful reviews incorrectly flagged using precision, recall, and confusion matrices.

### 1.2 Significance of the Research

This study examines whether unlabelled reviews can improve deceptive review detection when reliable labels are scarce. The comparison uses the same initial labelled examples, TF–IDF features, and test reviews for supervised logistic regression and its self-training counterpart. Evaluating four label budgets across ten draws shows how sensitive each approach is to the amount and selection of labelled data. Holding out entire hotels also provides evidence about how the models perform on reviews of properties they have not encountered during training.

The findings show why review detection should be evaluated using both precision and recall. At the largest label budget, self-training identified a greater share of deceptive reviews, while a smaller share of the reviews it flagged were deceptive. Its mean macro-F1 was lower than that of supervised logistic regression at all four label budgets. In a moderation workflow, this trade-off matters because flagging more reviews can increase the number of truthful reviews sent for manual inspection. The findings provide a reproducible benchmark for studying that trade-off under limited-label conditions. Their scope is the hotel-review corpus, test split, and self-training configuration evaluated in this study.

### 1.3 Problem Statement

Deceptive hotel reviews can mislead prospective guests and affect the reputation of hotels. Detecting these reviews from text is difficult because fabricated accounts can resemble truthful descriptions of a stay. Supervised classifiers require reliable examples of both truthful and deceptive reviews, but obtaining such labels can be difficult when the reviewer's actual experience cannot be verified. Semi-supervised self-training can use unlabelled reviews by assigning them provisional labels. If those labels are incorrect, however, retraining can reinforce the classifier's initial errors and cause truthful reviews to be flagged.

The problem addressed in this study is whether confidence-based self-training improves the detection of deceptive hotel reviews when only a limited number of labelled examples are available. The study compares self-training with supervised logistic regression using the same initial labelled examples and TF–IDF features across four label budgets. Both methods are evaluated on the same reviews from hotels excluded from training. Macro-F1 measures overall classification performance, while deceptive-class precision and recall show whether detecting more deceptive reviews comes at the cost of flagging more truthful ones.

### 1.4 Motivation for the Research

An online review platform may collect thousands of reviews, yet verifying which ones describe genuine experiences requires evidence that is rarely available from the text alone. This creates a practical motivation for semi-supervised learning. If a classifier can learn from unlabelled reviews, it could improve detection without requiring every review to be manually annotated. Confidence-based self-training offers a way to test this possibility by adding reviews to the training set when the model is sufficiently confident in its predicted labels.

The value of this approach depends on whether those predicted labels are reliable, especially when the initial training set is small. An early error can be reinforced during retraining, and incorrectly flagging truthful reviews can create unnecessary work for moderators. These concerns motivate a comparison across several label budgets using reviews from hotels excluded from training. By examining macro-F1 alongside precision and recall for deceptive reviews, the study evaluates whether self-training improves overall detection and how it affects the balance between missed deceptive reviews and truthful reviews flagged for inspection.

## 2. REVIEW OF RELATED WORKS

### 2.1 Supervised detection of deceptive reviews

Jindal and Liu (2008) established opinion spam as a problem in product-review analysis. Their study used duplicate and near-duplicate reviews to investigate potentially untrustworthy content and considered information about reviews, reviewers, and products. Duplication provides a measurable signal, but individually written fake reviews can evade detection based on text overlap. Ott et al. (2011) addressed this problem with a benchmark of truthful and commissioned deceptive hotel reviews. Their supervised text classifiers achieved nearly 90% accuracy on the positive-review benchmark, demonstrating that linguistic features could distinguish the two groups under controlled conditions.

Ott et al. (2013) extended the benchmark to negative reviews and examined whether classifiers trained on one sentiment could detect deception in the other. The n-gram SVM achieved 89.3% accuracy when trained and evaluated on positive reviews through cross-validation, while the positive-to-negative evaluation yielded 75.1% accuracy. Li et al. (2014) found a related limitation across domains. An SVM trained with hotel-review unigrams achieved approximately 55% accuracy when applied to doctor reviews in one transfer setting. These studies show that sentiment and domain can influence which textual cues a classifier learns. They also clarify the scope of the present evaluation. Holding out hotels tests generalisation to unseen properties within the hospitality domain.

Mukherjee et al. (2013) examined Yelp reviews using the platform's filter decisions as proxy labels. Reviewer behaviour features outperformed linguistic n-grams in identifying filtered reviews. This result highlights the value of behavioural information when it is available. It also illustrates how the source of the labels changes the classification task. Detecting commissioned deceptive text and identifying reviews flagged by a platform filter involve different evidence about review authenticity.

### 2.2 Semi-supervised detection with limited labels

Ren et al. (2014) formulated deceptive review detection as a positive–unlabelled learning problem. Their approach used topic modelling to identify representative examples and assigned similarity weights to ambiguous reviews before incorporating them into an SVM. The method outperformed the baselines used in their benchmark evaluation. Its treatment of uncertain examples is relevant to the present study because confidence-based self-training also relies on inferred labels, but adds selected reviews to the training set using the classifier's predicted confidence.

Arazo et al. (2020) showed in deep semi-supervised learning experiments that incorrect pseudo-labels can reinforce a model's initial errors, a phenomenon known as *confirmation bias*. Amini et al. (2025) reviewed self-training methods and discussed the role of threshold selection and pseudo-label quality. For the present study, these findings motivate comparing self-training against its own supervised base classifier at several label budgets. They also support examining precision and recall separately, since adding pseudo-labelled reviews can change which kinds of errors the model makes.

Table 1: Comparison of related studies on deceptive review detection and self-training

| Study | Data and method | Finding relevant to the present study |
|---|---|---|
| Jindal and Liu (2008) | Product reviews and duplicate-based opinion spam analysis | Review duplication provides a signal for identifying potentially untrustworthy conten. |
| Ott et al. (2011) | Truthful and commissioned deceptive positive hotel reviews, evaluated with supervised text classifiers | Linguistic features supported high accuracy under controlled benchmark conditions. |
| Ott et al. (2013) | Positive and negative hotel reviews, evaluated with an n-gram SVM | Performance declined when training and testing used reviews of different sentiments. |
| Mukherjee et al (2013) | Yelp filter decisions and linguistic and behavioural features | Behavioural features outperformed linguistic n-grams for identifying filtered reviews. |
| Li et al. (2014) | Hotel, restaurant, and doctor reviews, with cross-domain evaluation | Some features learned from hotel reviews transferred poorly to doctor reviews. |
| Ren et al. (2014) | Deceptive review benchmark and positive–unlabelled learning with a weighted SVM | Incorporating unlabelled reviews with explicit weighting improved performance over the study's baselines. |
| Arazo et al. (2020) | Deep semi-supervised classification with pseudo-labelling | Incorrect pseudo-labels can reinforce model errors during retraining. |

### 2.3 Research gap and position of the present study

The reviewed studies establish the usefulness of supervised text classifiers and show that some semi-supervised methods can benefit from unlabelled reviews. The question addressed here is narrower. *Does adding unlabelled hotel reviews through confidence-based self-training improve detection when the initial labelled examples, feature representation, and test set are held constant?* The present study isolates the effect of self-training by pairing it with the same logistic regression model trained solely on the initial labelled examples. It compares four label budgets across ten draws of the labelled subset, using the same 320 test reviews from four held-out hotels. Macro-F1 measures performance across both classes, while deceptive-class precision and recall show whether any change in detection comes with more truthful reviews being flagged.

## 3. METHODOLOGY

### 3.1 Research design

This preliminary benchmark study examined whether confidence-based self-training improves deceptive hotel review detection when labelled training data are limited. The primary comparison paired supervised logistic regression with semi-supervised self-training using the same logistic regression classifier. Both approaches used identical initial labelled reviews, the same text features, and the same test set. Self-training additionally used the remaining training reviews with their labels masked. Multinomial Naïve Bayes and linear support vector machine (SVM) served as additional supervised baselines.

### 3.2 Dataset and partitioning

The Deceptive Opinion Spam Corpus v1.4 contains 1,600 English-language reviews from 20 Chicago hotels (Ott et al., 2011, 2013). It includes 800 truthful and 800 deceptive reviews, with 400 positive and 400 negative reviews in each class. Truthful reviews were collected from travel websites, and deceptive reviews were commissioned from crowdworkers.

The corpus's predefined fold 1 was used as the single fixed test fold. It contained 320 reviews from four hotels, with 160 reviews per class. The remaining 1,280 reviews from 16 hotels formed the training pool. A check found no hotel overlap or exact duplicate review text across the training and test partitions after normalising letter case and whitespace.

### 3.3 Label budgets and text representation

Ten labelled subsets were drawn at each budget using random seeds 0–9. Sampling was stratified by the joint combination of deception label and sentiment polarity. Reviews outside the labelled subset remained in the training pool with their labels masked for self-training.

Table 2: Distribution of Labelled, Unlabelled, and Test Reviews across Label Budgets

| Label budget (% of training pool) | Labelled training reviews | Training reviews with masked labels | Test reviews |
|---|---|---|---|
| 5% | 64 | 1,216 | 320 |
| 10% | 128 | 1,152 | 320 |
| 20% | 256 | 1,024 | 320 |
| 25% | 320 | 960 | 320 |

Review text was converted into term frequency–inverse document frequency (TF–IDF) vectors using word unigrams, word bigrams, and sublinear term-frequency scaling. The vocabulary and IDF weights were fitted only on the initial labelled subset in each run. The fitted vectoriser then transformed the remaining training reviews and the held-out test reviews. All models within a run used the same representation. Review source, hotel identifier, and sentiment polarity were excluded from the model inputs.

### 3.4 Model training

The supervised models were fitted on the initial labelled subset. Self-training began by fitting logistic regression on that same subset and predicting class probabilities for the reviews with masked labels.

Table 3: Supervised and Semi-Supervised Models and Training Configurations

| Approach | Model | Configuration |
|---|---|---|
| Supervised | Multinomial Naïve Bayes | Additive smoothing with $\alpha = 1$ |
| Supervised | Linear SVM | Regularisation parameter $C = 1$ |
| Supervised | Logistic regression | liblinear solver, $C = 10$, maximum 2,000 solver iterations |
| Semi-supervised | Logistic regression with self-training | Same logistic regression settings, confidence threshold of 0.65, maximum ten self-training iterations |

A review was assigned a pseudo-label if its highest predicted class probability exceeded 0.65. The classifier was then refitted using the original labelled reviews and the accepted pseudo-labelled reviews. This process continued for up to ten iterations, ending earlier if no additional reviews met the threshold or the unlabelled pool was exhausted. Model settings were fixed without using the test results for selection.

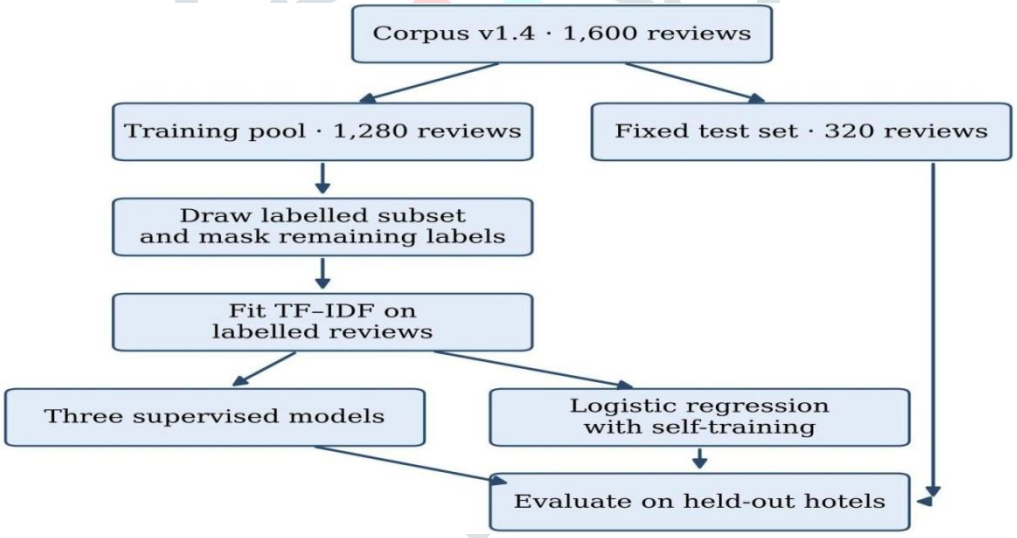



Figure 1: Experimental workflow for supervised and semi-supervised detection of deceptive hotel reviews.

Figure 1 shows how the study isolates the effect of adding unlabelled reviews. Supervised logistic regression and self-training start with the same labelled reviews and TF–IDF features. Only self-training uses the remaining training reviews with their labels masked. Both methods are evaluated on the same 320 reviews from four hotels excluded from training. This design supports a direct comparison under the same training and testing conditions.

### 3.5 Performance evaluation

Macro-F1 was the primary metric. It gives equal weight to the F1 scores of the truthful and deceptive classes. Accuracy, deceptive-class precision and recall, and confusion matrices were also recorded. Treating deceptive reviews as the positive class made it possible to measure both how many deceptive reviews were missed and how many truthful reviews were incorrectly flagged.

Performance was summarised as the mean and standard deviation across the ten labelled-subset draws at each budget. For the primary comparison, the difference in macro-F1 between self-training and supervised logistic regression was calculated within each matched run. Since all runs used the same test fold, the standard deviations describe sensitivity to the selection of labelled training reviews on that holdout.

## 4. RESULTS AND DISCUSSION

### 4.1 Performance across label budgets

Table 4 presents the mean macro-F1 scores across ten draws of the labelled training subset. All models were evaluated on the same 320 reviews from four hotels excluded from training.

Table 4: Mean test macro-F1 across four label budgets

| Labelled training reviews | Multinomial Naïve Bayes | Linear SVM | Logistic regression | Self-training | Self-training change vs. logistic regression |
|---|---|---|---|---|---|
| 64 | 66.32% | 70.36% | 70.29% | 46.61% | −23.68 pp |
| 128 | 70.50% | 74.30% | 74.24% | 62.61% | −11.63 pp |
| 256 | 75.03% | 79.04% | 79.01% | 73.54% | −5.47 pp |
| 320 | 75.81% | 80.48% | 80.57% | 75.37% | −5.19 pp |

*pp = percentage points. Changes were calculated from the unrounded scores for each matched run before averaging.*

Mean macro-F1 increased with the number of labelled reviews for all four models. Linear SVM and logistic regression produced similar scores across the four budgets. With 320 labelled reviews, logistic regression achieved the highest mean macro-F1 of 80.57%**,** followed by linear SVM at 80.48%**.**

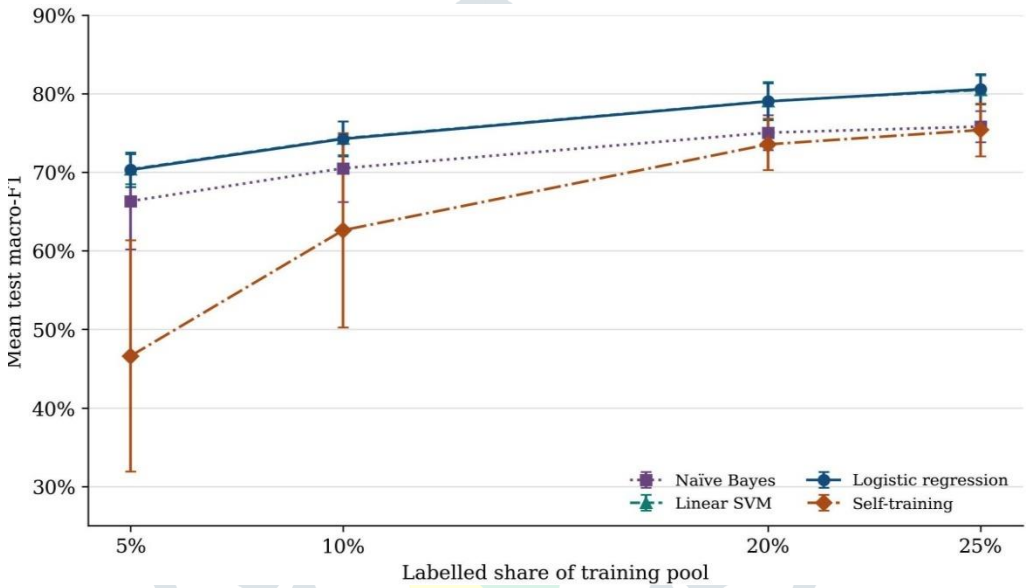



Figure 2 : Mean test macro-F1 by label budget. Error bars show one standard deviation across ten labelled-subset draws.

Figure 2 illustrates the Mean macro-F1 increased for all four models as the number of labelled reviews rose from 64 to 320. Linear SVM and supervised logistic regression performed similarly across the tested budgets. Self-training showed the greatest variability at the smallest budget, with a standard deviation of 14.72 percentage points, compared with 2.19 for supervised logistic regression. The wider error bars indicate that self-training performance was more sensitive to which reviews were labelled when only 64 labels were available.

### 4.2 Effect of adding unlabelled reviews

Confidence-based self-training yielded lower mean macro-F1 than its supervised logistic regression counterpart at every tested label budget. The largest reduction occurred with 64 labelled reviews, where mean macro-F1 fell by 23.68 percentage points. At 320 labelled reviews, the average reduction narrowed to 5.19 percentage points. Self-training was also more sensitive to the choice of labelled reviews at the smallest budget. Its macro-F1 standard deviation was 14.72 percentage points, compared with 2.19 percentage points for supervised logistic regression.

At the smallest budget, the self-training model accepted an average of approximately 1,176 of the 1,216 reviews with masked labels. Adding so many pseudo-labelled examples to a model initially trained on 64 labelled reviews could amplify early classification errors. The experiment did not isolate that mechanism, so the large number of accepted pseudo-labels provides a possible explanation for the performance decline.

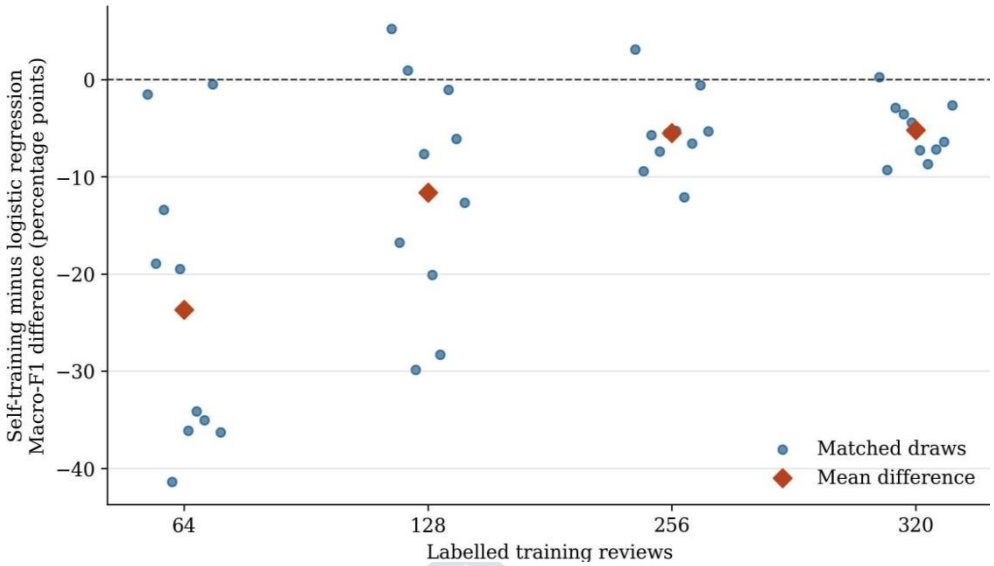



Figure 3: Paired macro-F1 differences between self-training and supervised logistic regression. Each dot represents one matched draw, and each diamond represents the mean difference at that label budget.

In Figure 3, each point represents the difference between self-training and supervised logistic regression for the same labelled subset. The mean difference was negative at all four budgets, narrowing from −23.68 percentage points with 64 labelled reviews to −5.19 percentage points with 320. Some individual draws showed a small improvement, but gains were inconsistent. On the fixed test fold, the tested self-training configuration lowered mean macro-F1 at every budget.

### 4.3 Precision and recall for deceptive reviews

At the largest label budget, self-training increased mean deceptive-class recall from 85.69% to 88.50%, a gain of 2.81 percentage points. Mean precision fell from 77.87% to 70.88%, a decline of 6.99 percentage points. Self-training therefore detected a larger share of deceptive reviews, while a larger share of the reviews it flagged were truthful.

Table 5: Mean test performance with 320 labelled training reviews

| Model | Accuracy | Macro-F1 | Deceptive class precision | Deceptive class recall |
|---|---|---|---|---|
| Multinomial Naïve Bayes | 76.53% | 75.81% | 69.84% | 93.56% |
| Linear SVM | 80.53% | 80.48% | 77.82% | 85.50% |
| Logistic regression | 80.62% | 80.57% | 77.87% | 85.69% |
| Self-training | 75.84% | 75.37% | 70.88% | 88.50% |

Multinomial Naïve Bayes achieved the highest recall at this label budget, but its lower precision indicates that more of its alerts were false alarms. For a review moderation workflow, this trade-off affects how many genuine reviews are sent for manual inspection. Logistic regression achieved the highest mean macro-F1 on the held-out fold, while Naïve Bayes retrieved the largest share of deceptive reviews

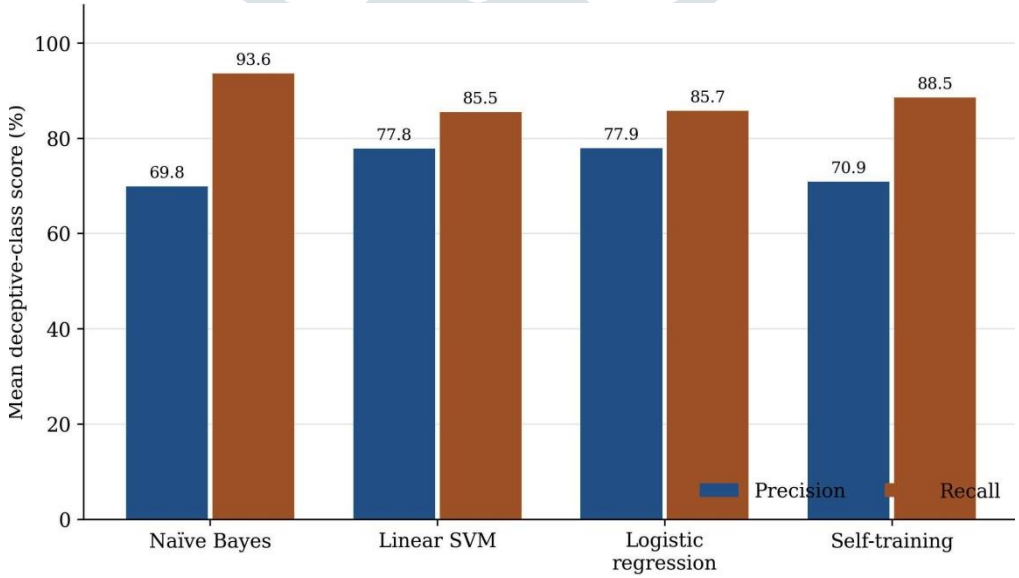



Figure 4: Mean precision and recall for deceptive reviews with 320 labelled training examples.

In Figure 4 with 320 labelled reviews, self-training increased deceptive-class recall from 85.69% to 88.50% while reducing precision from 77.87% to 70.88%. It detected a larger share of deceptive reviews, and a larger share of its flagged reviews were truthful. Multinomial Naïve Bayes achieved the highest recall (93.56%), while supervised logistic regression achieved the highest precision (77.87%). For review moderation, this trade-off matters because higher recall identifies more deceptive reviews, while higher precision reduces the share of flagged reviews that require unnecessary inspection.

### 4.4 Scope of the findings

The results answer the study's primary question for the evaluated configuration. Adding unlabelled reviews through self-training at a confidence threshold of 0.65 did not improve mean macro-F1**.** The ten repetitions capture variation in the selection of labelled

training reviews on one fixed test fold. Evaluation across the remaining hotel folds and validation-based threshold selection would help establish whether the pattern persists. The benchmark uses commissioned deceptive reviews, so applying these findings to naturally occurring platform fraud or AI-generated reviews would require separate validation.

### 4.5 Discussion

The experiment examined whether confidence-based self-training improves deceptive hotel review detection when labelled reviews are scarce. With the initial labelled reviews, TF–IDF representation, and test set held constant, self-training achieved lower mean macro-F1 than supervised logistic regression at every label budget. The gap narrowed from 23.68 percentage points with 64 labelled reviews to 5.19 percentage points with 320 labelled reviews. This pattern suggests that the self-training configuration was particularly sensitive to the amount of labelled data available at the start.

At the smallest label budget, self-training accepted an average of 1,176 of the 1,216 reviews with masked labels. If the initial classifier assigned some of these reviews incorrect labels with high confidence, retraining on those predictions could have reinforced its early errors. This explanation is consistent with the confirmation bias discussed by Arazo et al. (2020). The experiment did not directly measure pseudo-label accuracy, so the number of accepted reviews alone cannot establish the cause of the decline.

The class-level results show a practical trade-off. With 320 labelled reviews, self-training increased deceptive-class recall from 85.69% to 88.50%, while precision fell from 77.87% to 70.88%. In a moderation workflow, higher recall can help identify more deceptive reviews, while lower precision means a larger share of flagged reviews would be truthful. Logistic regression and linear SVM achieved similar mean macro-F1 scores at this budget, 80.57% and 80.48%, respectively. Multinomial Naïve Bayes achieved the highest deceptive-class recall, 93.56%, with precision of 69.84%. Model suitability therefore depends on the relative cost of missing a deceptive review and sending a truthful review for inspection.

The findings apply to the tested self-training threshold of 0.65 and a single test fold containing reviews from four held-out hotels. The ten draws capture how performance varies with the selection of labelled training reviews on that fold. Evaluating the remaining hotel folds, selecting the confidence threshold using training-side validation, and measuring pseudo-label accuracy would help determine whether the decline persists. Because the benchmark contains commissioned deceptive reviews, separate validation is needed to assess performance on naturally occurring platform fraud.

## 5. CONCLUSION AND FUTURE WORKS

This study evaluated whether confidence-based self-training improves deceptive hotel review detection when labelled examples are scarce. Multinomial Naïve Bayes, linear support vector machine, supervised logistic regression, and logistic regression with self-training were evaluated on 320 reviews from four hotels excluded from training. The comparison between supervised logistic regression and self-training used the same initial labelled reviews, TF–IDF features, and test set at each of the four label budgets.

Self-training achieved lower mean macro-F1 than supervised logistic regression at every label budget tested. The difference narrowed from 23.68 percentage points with 64 labelled reviews to 5.19 percentage points with 320 labelled reviews. At the largest budget, supervised logistic regression achieved a mean macro-F1 of 80.57%, compared with 75.37% for self-training. Self-training increased deceptive-class recall from 85.69% to 88.50%, while precision decreased from 77.87% to 70.88%. This trade-off matters for moderation because detecting more deceptive reviews can increase the share of truthful reviews among those flagged for inspection.

Under the evaluated conditions, supervised logistic regression provided the best mean balance between the two review classes at the largest label budget. The conclusion applies to the confidence threshold of 0.65 and the single hotel holdout used in the experiment. The ten repetitions measure sensitivity to the selection of labelled training reviews on that holdout.

## 6. FUTURE WORKS

The identified future research works are as follows

i. **Evaluate all five predefined hotel folds.** Rotate the held-out hotels and repeat the matched comparisons across label budgets. This would show whether the observed performance differences persist across properties and labelled subsets.

ii. **Tune the confidence threshold using training-side validation.** Compare fixed and class-specific thresholds while keeping test reviews separate. Record the number, class distribution, and accuracy of accepted pseudo-labels at each iteration. In the benchmark, the masked labels can be used for this diagnostic after training.

iii. **Compare alternative semi-supervised methods.** Evaluate label spreading, graph-based approaches, or other pseudo-labelling strategies using the same labelled subsets and test folds. This would help determine whether the observed decline is specific to the self-training configuration.

iv. **Test on more representative review data.** Extend the evaluation to other platforms and review domains using independently verified examples where available. Evaluate commissioned deceptive reviews, naturally occurring fraud, and AI-generated deceptive reviews separately because they represent different sources of evidence.

v. **Evaluate performance for moderation.** Report precision recall curves and compare the number of deceptive reviews detected at a fixed manual-review capacity. This would show how each model performs when moderators can inspect only a limited number of flagged reviews.

## REFERENCES

[1] Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, É., & Maximov, Y. (2025). Self-training: A survey. *Neurocomputing, 616*, Article 128904. https://doi.org/10.1016/j.neucom.2024.128904

[2] Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness, K. (2020). Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks* pp. 1–8, IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207304

[3] Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 219–230). Association for Computing Machinery. https://doi.org/10.1145/1341531.1341560

[4] Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1566–1576). Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1147

[5] Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What Yelp fake review filter might be doing? *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 409–418. https://doi.org/10.1609/icwsm.v7i1.14389

[6] Ott, M., Cardie, C., & Hancock, J. T. (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* (pp. 497–501). Association for Computational Linguistics. https://aclanthology.org/N13-1053

[7] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 309–319). Association for Computational Linguistics. https://aclanthology.org/P11-1032

[8] Ren, Y., Ji, D., & Zhang, H. (2014). Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 488–498). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1055

[9] Dataset used in the experiments**:** Deceptive Opinion Spam Corpus v1.4. Cite Ott et al. (2011, 2013) when describing its positive and negative hotel reviews.