A Comparative Study of Machine Learning Methods for Detection of Fake Online Consumer Reviews

Petr Hajek
Faculty of Economics and Administration
University of Pardubice
Studentska 84, 53210 Pardubice
Czech Republic
+420 466 036 147
petr.hajek@upce.cz

Aliaksandr Barushka
Faculty of Economics and Administration
University of Pardubice
Studentska 84, 53210 Pardubice
Czech Republic
+420 466 036 147
st47591@student.upce.cz

ABSTRACT
Online product reviews provide valuable information for consumer decision making. Customers increasingly rely on the reviews and consider them a trusted source of information. For businesses, it is therefore tempting to purchase fake reviews because competitive advantage can be easily achieved by producing positive or negative fake reviews. Machine learning methods have become a critical tool to automatically identify fake reviews. Recently, deep neural networks have shown promising detection accuracy. However, there have been no studies which compare the performance of state-of-the-art deep learning approaches with traditional machine learning methods, such as Naïve Bayes, support vector machines or decision trees. The aim of this study is to examine the performance of several machine learning methods used for the detection of positive and negative fake consumer reviews. Here we show that deep neural networks, including convolutional neural networks and long short term memory, significantly outperform the traditional machine learning methods in terms of accuracy while preserving desirable time performance.

CCS Concepts
• Machine learning approaches→Neural networks • Artificial intelligence→Natural language processing • Applied computing→Document management and text processing

Keywords
Fake, reviews, machine learning, deep learning, classification

1. INTRODUCTION
With many online purchase options, positive or negative product reviews may be highly important for consumer decision making during the purchase process. A meta-analysis of more than twenty empirical studies found that both review volume and review valence are significant determinants of retail sales [1]. This holds particularly for high-involvement products that can only be evaluated upon consumption. An important assumption is that the online review is based on consumer's experience of product/service use. The issue of trust must therefore be addressed. As presented in a recent consumer review survey [2], more than eighty percent of consumers trust online reviews as much as they trust personal recommendations. In highly competitive business environments, it is obviously tempting to generate good (bad) reviews for our (competition) products. In fact, it is now easy to find freelance writers that are able to produce a large number of positive or negative reviews. Most marketplaces like Amazon give priority to well-evaluated products (the so-called snowball effect), thus potentially rewarding businesses paying for fake reviews. Therefore, there has been an increasing interest in automatic detection of fake online consumer reviews.

To detect fake reviews, machine learning methods have extensively been used in recent years, see [3] for a survey. These methods utilize features extracted from the texts of reviews. A common approach is to extract a bag of words consisting of the list of words or phrases. Word categories can also be extracted, such as sentiment features or part-of-speech tagging. Reviewer information can be used as the additional source of features. The machine learning methods then use those features to classify the reviews into fake / legitimate class. Most studies in this domain have only been carried out using one or few machine learning methods. The performance of traditional machine learning methods, such as Naïve Bayes (NB) and support vector machine (SVM), have been compared in previous studies [4]. However, deep neural networks have demonstrated remarkable performance in detecting fake reviews in recent years [5-7]. Moreover, no research has been found that surveyed and compared the performance of the deep neural network methods with that of traditional machine learning methods. The purpose of this paper is to perform such a comparative analysis. A well-known benchmark dataset of hotel reviews is used in this study. Both positive and negative reviews are considered in the comparison.

The remainder of this paper is organized as follows. Section 2 reviews related literature on machine learning methods applied to detecting fake online consumer reviews. Section 3 presents the datasets and methods used for the comparative analysis. Section 4 shows the results of the comparison and section 5 concludes the paper.

2. RELATED LITERATURE
A considerable amount of literature has been published on the intelligent detection of fake online consumer reviews in the last decade. The automatic intelligent detection methods, such as those using machine learning, have been identified as being faster, cheaper and more accurate than human judgment [8]. The machine learning methods aim to identify fake (spam) reviews

ICEBI '2019, November 9–11, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-7170-4/19/11…$15.00
https://doi.org/10.1145/3383902.3383909

based on their content or reviewer's behavior. In other words, the reviews are classified into two categories (fake / legitimate) using the methods trained on the dataset of annotated (labelled) reviews.

Various approaches have been put forward to classify fake online consumer reviews, see Table 1. First, logistic regression (LR) has been investigated as the traditional machine learning method [9]. Reviews from amazon.com were used due to its large coverage, and LR was employed owing to its capacity to produce the probability estimate reflecting the likelihood that a review is a fake review. However, the LR method, similarly as *k*-NN (*k*-nearest neighbor), suffers from several shortcomings [10]. First, these methods cannot handle high dimensional data effectively. This is an important problem as the fake review data usually involve a large number (hundreds or thousands) of word features generated in the bag-of-words fashion. Another problem is the limited ability of those methods to deal with data sparsity. This issue should be effectively addressed because the reviews usually contain only a small number of words or phrases. These limitations caused decrease in detection accuracy. Machine learning methods, such as NB [11] or SVM [12,13], address those problems with the review data more effectively. To overcome the problem of overfitting, ensemble learning approaches have been utilized [14,15]. A detailed overview of the traditional machine learning methods used in fake review detection can be found in [3].

Table 1. List of previous studies on detecting fake online consumer reviews

Study	Data	Method	Performance
[9]	Amazon	LR	AUC=0.78
[11]	Epinions	NB, Co-training	*F*-score=0.63
[17]	iOS App Store	DT	Acc=0.59
[8]	Hotels	SVM	Acc=0.86
[12]	Yelp	SVM	Acc=0.86
[18]	Hotels and doctors	SAGE	Acc=0.65
[13]	Restaurants	SVM	Acc=0.85
[16]	Hotels, restaurant and doctors	CNN, SWNN	Acc=0.84
[5]	Hotels, restaurant and doctors	CNN, GRNN	Acc=0.84
[4]	Movies	*k*-NN, NB, DT, SVM	Acc=0.82
[14]	Hotels	*k*-NN, RF	Acc=0.77
[7]	Hotels	DFFNN	Acc=0.89
[15]	Hotels, restaurant and doctors	LSTM	Acc=0.85
[6]	Consumer electronics	AdaBoost	*F*-score=0.82

Notes: Acc is accuracy, AUC is area under receiver operating characteristic curve, DT is decision tree, RF is random forest, SAGE is sparse additive generative model, and SWNN is sentence weighted neural network.

As it turns out, more complex features can be extracted from the high-dimensional data using deep neural networks. Therefore, machine learning methods, such as convolutional neural network (CNN) [16], general regression neural network (GRNN) [5], deep feed-forward neural network (DFFNN) [7] and long short term memory (LSTM) [6], have been gaining much attention in recent years.

Although extensive research has been carried out on fake review detection using machine learning methods, no single study exists which adequately covers the comparative analysis of the above-mentioned methods.

3. DATA AND METHODS

3.1 Datasets

In this comparative study, two datasets from Cornell University[1] were selected for benchmarking. These datasets were used because they are considered a gold-standard fake review dataset [8], with fake reviews generated by unique Turkers pretending to be customers. The reviews were required to sound realistic and positive or negative. Positive and negative reviews were sorted, with one dataset has only positive reviews, while the other one has only negative reviews. Both datasets contain 800 reviews with different polarity. Each dataset includes 400 legitimate reviews and 400 fake reviews. Each dataset comprises reviews on 20 hotels from TripAdvisor and there are 20 legitimate and 20 fake reviews for every hotel. Both datasets contain only reviews with certain evaluation (number of stars), namely 1 and 2 star reviews for the negative dataset, while only 5 star reviews for the positive dataset. Each review in the dataset has legitimate/fake label, hotel information, travel agency name, polarity and review content. The reviews contained 116 words on average.

To preprocess the content of all reviews, we adopted the approach proposed in [7]. Specifically, (1) the Rainbow stopwords list was used to remove stopwords; (2) punctuation and special symbols were stripped off; (3) lowercase letters were used for all words; and (4) top 2000 words and phrases of length 1, 2 and 3 (unigrams, bigrams and trigrams) were extracted from the text according to their frequency. To calculate their weights, we used the *tf.idf* weighting scheme defined by the following formula:

$$v_{ij} = (1 + \log(tf_{ij})) \times \log(N/df_i), \quad (1)$$

where v_{ij} denotes the weight of the *i*-th word (phrase) in the *j*-th review, tf_{ij} and df_i respectively represent term and document frequencies, and N is the number of reviews in the dataset. Note that this is a common text preprocessing methodology, considering the frequency of word occurrences, the length of the reviews and the rareness of the words.

3.2 Machine Learning Methods

To compare the performance on the benchmark datasets, we selected the methods used in earlier related research, namely: (1) *k*-NN [4,14], (2) NB [4,11], (3) decision tree (DT) [17], (4) random forest (RF) [14], (5) SVM [8,12,13], (6) CNN [5,16], (7) DFFNN [7], and (8) LSTM [15]. In this subsection, the methods are briefly introduced together with the setting of their learning parameters.

The *k*-NN algorithm does not require the training phase as the classification is based on the comparison with training data only. Specifically, the decision on class depends on the *k* most-similar instances (reviews). The Euclidean distance is typically used to

[1] http://myleott.com/op-spam.html

measure the similarity. In our experiments, we set *k* to a standard value 3.

The NB method represents a probability-based approach. More precisely, the posterior probability is calculated that a review is fake (legitimate) given the words or phrases that occur in the review.

As a common representative of DT, we used the J48 algorithm that generates a binary tree based on features selected using the gain ratio measure. The advantage of this algorithm is that pruned trees can be generated. This is done to avoid model overfitting. To perform pruning, we set the confidence factor for pruning to 0.25. Another parameter was the minimum number of instances per leaf. We set its value to 2.

From the category of ensemble machine learning methods, we opted to use RF because it combines single trees so that the generalization error of RF is limited. This is also attributed to the random selection of features used to split nodes in individual trees. Thus, noise robustness is achieved. In the experiments, we did not limit the maximum depth of the trees, and the number of randomly selected features was set using the heuristic $\log_2(\#\text{predictors})+1$.

In SVM, the optimal separating hyperplane is based on the maximum margin between fake and legitimate classes. For this, only a subset of reviews (support vectors) can be applied. In this study, we used the sequential minimal optimization (SMO) algorithm to learn SVM parameters. Polynomial kernel functions were used to handle the non-linearity in the data. To find the optimum complexity, different values of parameter *C* were examined in our experiments, $C=\{2^0, 2^1, \ldots, 2^6\}$.

CNN is a deep neural network model consisting of layers with convolving filters. Complex features can then be extracted by applying the filters to form a feature map from the local features in adjacent layers. Multiple feature maps are stored in each hidden layer. Max pooling is then applied to extract the most important features from the convolutional layers. More precisely, the used CNN comprised two convolutional layers (5×5 and 2×2) with 20 feature maps. Droupout was used to avoid overfitting and rectified linear units ensured fast convergence. The mini-batch gradient descent algorithm was employed to train the CNN model.

The fully connected deep feed-forward neural network (DFFNN) used in this study was trained using the mini-batch gradient descent algorithm two hidden layers and number of neurons in the hidden layers in the range {10, 20, 50, 100}. The optimal DFFNN structure was found using a grid search procedure. Similarly as for CNN, dropout and rectified linear units were utilized.

The LSTM model is a recurrent neural network (RNN) that overcomes the vanishing gradients problem using memory cell. Gating mechanism is introduced to control the information in the memory cell. In the LSTM model used here, LSTM layer and RNN output layer were constructed with $\{2^4, 2^5, 2^6, 2^7\}$ and 2 neurons, respectively. TanH activation function was used in the LSTM layer, and the stochastic gradient descent algorithm with Adam updater was employed to train this neural network.

4. EXPERIMENTAL RESULTS

To evaluate the results of experiments, we used accuracy Acc (% correctly classified reviews). Furthermore, we evaluated the performance of the methods for each class (fake and legitimate) using true positive (TP) and true negative (TN) rates. TP (TN) rate is the percentage of fake (legitimate) reviews classified correctly. In addition, area under ROC (receiver operating characteristic) curve (AUC) was used to report the trade-off between TP rate and FP (false positive) rate at various threshold values. For the four evaluation measures, we report the results obtained over stratified 10-fold cross-validation. The experiments were carried out in Weka 3.8 environment (traditional machine learning methods) and Deeplearning4j environment (deep neural networks) on an Intel Core CPU (six i5-8400 cores) with 16 GB RAM. To compare the performance statistically, Wilcoxon signed rank test was used in this comparative study.

Table 2 summarizes the results of the compared methods in terms of Acc and AUC. As Table 2 shows, there is a significant difference in the performance between deep neural networks and other machine learning methods except RF for both datasets. The DFFNN model performed best for the positive reviews with Acc=89.00%, while LSTM dominated for the negative review dataset with Acc=89.13%. Hence, the results suggest that similar accuracy can be obtain regardless of the data polarity (positive / negative). These results can be explained by the complex features extracted by using hidden layers in the deep neural networks. The high values of AUC suggest that the methods performed well for both classes, fake and legitimate.

Table 2. Performance of compared methods in terms of Accuracy and AUC

	Positive reviews		**Negative reviews**	
Method	**Acc**	**AUC**	**Acc**	**AUC**
k-NN	55.88	0.61	60.88	0.65
NB	83.13	0.87	82.13	0.85
DT	72.50	0.73	69.25	0.72
RF	87.13*	0.94*	85.25	0.93*
SVM	85.38	0.85	85.38	0.85
CNN	88.00*	0.95*	88.38*	0.95*
DFFNN	89.00*	0.96*	89.13*	0.95*
LSTM	88.13*	0.95*	89.63*	0.96*

* significantly similar as the best classifier at $P<0.05$

To further examine the performance of the machine learning methods for each class, Table 3 shows the results in terms of TP rate and TN rate, respectively. On one hand, TP rate can be seen as a more important evaluation measure as it reports the percentage of fake reviews identified correctly. On the other hand, legitimate reviews represent a valuable source of information for the customers. A high TP rate accompanied with low TN rate would result in unwanted excessive removal of legitimate reviews. A balanced performance is therefore desirable. Table 3 shows that the deep neural networks performed best with respect to the balanced performance on both classes.

Finally, we report the time performance of the compared method in terms of training and testing time. Training time represents the time needed to train the classifier, while testing time is the time needed to perform detection of newly submitted reviews. The results in Table 4 show that the DFFNN model was the least time efficient of all the methods, while LSTM was the most time efficient deep neural network model. Overall, all the results can be considered acceptable for real-time fake review detection as the testing time is less than 2 seconds [19].

Table 3. Performance of compared methods in terms of TP rate and TN rate

	Positive reviews		Negative reviews	
Method	**TP rate**	**TN rate**	**TP rate**	**TN rate**
k-NN	23.75	88.00*	68.25	53.50
NB	83.75	82.50	82.00	82.25
DT	70.75	74.25	66.50	72.00
RF	86.50	87.75*	86.75	83.75
SVM	85.25	85.50	85.75	85.00
CNN	88.25*	87.75*	87.50*	89.25*
DFFNN	89.25*	88.75*	89.50*	88.75
LSTM	89.25*	87.00*	88.00*	91.25*

* significantly similar as the best classifier at $P<0.05$

Table 4. Training and testing time of compared methods

	Positive reviews		Negative reviews	
Method	**Training time [s]**	**Testing time [s]**	**Training time [s]**	**Testing time [s]**
k-NN	0.01	0.49	0.00	0.32
NB	0.92	0.32	0.47	0.19
DT	11.11	0.00	7.13	0.00
RF	8.02	0.11	5.70	0.04
SVM	1.45	0.02	0.75	0.01
CNN	106.04	0.07	101.00	0.07
DFFNN	750.01	1.64	543.07	0.50
LSTM	19.37	0.02	21.34	0.03

5. CONCLUSIONS

In conclusion, the evidence from this study suggests that deep neural networks significantly outperform traditional machine learning methods in terms of accuracy on both fake and legitimate review classes. This finding corroborates those presented in recent literature [16,19]. Remarkably, these models performed well irrespective of review polarity. Interestingly, statistically similar classification performance was obtained for all the three tested models of deep neural networks, namely CNN, DFFNN and LSTM. This suggests that additional useful hidden features can be extracted from reviews using multiple hidden layers. Although more training time is required to build these models, this learning process is usually done offline and only once. Therefore, testing time is more important for real-time fake review detection. Regarding this time performance measure, all the tested methods performed well.

A number of important limitations need to be considered. First, the experiments were performed only on the datasets preprocessed using bag of words approach. Recent evidence suggests that additional features should be extracted from the text of the reviews, such as context information [7]. Moreover, reviewer information can be considered in future comparative studies. Second, cost-sensitive performance measures have recently been tested for spam detection tasks [19]. Different misclassification cost for legitimate and fake reviews should therefore be considered in future research. Third, the datasets were related to hotel reviews only, thus neglecting other consumer domains [20]. A cross-domain comparative study is therefore another recommended research direction. Finally, our study was limited to machine learning methods based on supervised learning because all the reviews in the datasets were labelled with classes. However, several previous studies also utilized unlabelled reviews and employed methods with unsupervised or semi-supervised learning (see [21] for a survey). A comparative analysis of those methods would also be interesting direction for future research.

6. ACKNOWLEDGMENTS

This article was supported by the by the scientific research project of the Czech Sciences Foundation Grant No: 19-15498S and grant No. SGS_2019_17 of the Student Grant Competition.

7. REFERENCES

[1] Floyd, K., Freling, R., Alhoqail, S., Cho, H.Y., and Freling, T. 2014. How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing* 90(2), 217-232. DOI= 10.1016/j.jretai.2014.04.004.

[2] BrightLocal. 2018. *Local consumer review survey 2018*. Available at: https://www.brightlocal.com/research/local-consumer-review-survey/.

[3] Crawford, M., Khoshgoftaar, T.M., Prusa, J.D., Richter, A.N., and Al Najada, H. 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data* 2(1), 1-23. DOI=10.1186/s40537-015-0029-9.

[4] Elmurngi, E. and Gherbi, A. 2017. An empirical study on detecting fake reviews using machine learning techniques. In *7th Int. Conf. on Innovative Computing Technology (INTECH)*, IEEE, 107-114. DOI=10.1109/INTECH.2017. 8102442.

[5] Ren, Y. and Ji, D. 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences* 385, 213-224. DOI=10.1016/j.ins.2017.01.015.

[6] Zeng, Z.Y., Lin, J.J., Chen, M.S., Chen, M.H., Lan, Y.Q., and Liu, J.L. 2019. A review structure based ensemble model for deceptive review spam. *Information* 10(7), 243. DOI=10.3390/info10070243.

[7] Barushka, A. and Hajek, P. 2019. Review spam detection using word embeddings and deep neural networks. In *IFIP Int. Conf. on Artificial Intelligence Applications and Innovations*, Springer, Cham, 340-350. DOI=10.1007/978-3-030-19823-7_28.

[8] Ott, M., Cardie, C., and Hancock, J.T. 2013. Negative deceptive opinion spam. In *2013 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies*, 497-501.

[9] Jindal, N. and Liu, B. 2007. Analyzing and detecting review spam. In *7th IEEE Int. Conf. on Data Mining (ICDM 2007)*, IEEE, 547-552. DOI=10.1109/ICDM.2007.68.

[10] Barushka, A. and Hajek, P. 2018. Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks. *Applied Intelligence* 48(10), 3538-3556. DOI=10.1007/s10489-018-1161-y.

[11] Li, F., Huang, M., Yang, Y., and Zhu, X. 2011. Learning to identify review spam. In *Int. Joint Conf. on Artificial Intelligence (IJCAI 2011)*, 2488-2493.

[12] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. 2013. What yelp fake review filter might be doing?. In *7th Int. AAAI Conf. on Weblogs and Social Media*, 409-418.

[13] Li, H., Chen, Z., Mukherjee, A., Liu, B., and Shao, J. 2015. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *9th Int. AAAI Conf. on Web and Social Media (ICWSM 2015)*, AAAI, 634-637.

[14] Rout, J. K., Dalmia, A., Choo, K.K.R., Bakshi, S., and Jena, S.K. 2017. Revisiting semi-supervised learning for online deceptive review detection. *IEEE Access* 5, 1319-1327. DOI=10.1109/ACCESS.2017.2655032.

[15] Barbado, R., Araque, O., and Iglesias, C.A. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management* 56(4), 1234-1244. DOI=10.1016/j.ipm.2019.03.002.

[16] Li, L., Qin, B., Ren, W., and Liu, T. 2017. Document representation and feature combination for deceptive spam review detection. *Neurocomputing* 254, 33-41. DOI=10.1016/j.neucom.2016.10.080.

[17] Chandy, R. and Gu, H. 2012. Identifying spam in the iOS app store. In *Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, ACM, 56-59. DOI=10.1145/2184305.2184317.

[18] Li, J., Ott, M., Cardie, C., and Hovy, E. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 1566-1576.

[19] Barushka, A. and Hajek, P. 2019. Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. *Neural Computing and Applications* 1-19. DOI=10.1007/s00521-019-04331-5.

[20] Rout, J.K., Dash, A.K., and Ray, N.K. 2018. A framework for fake review detection: Issues and challenges. In *2018 Int. Conf. on Information Technology (ICIT)*, IEEE, 7-10. DOI=10.1109/ICIT.2018.00014.

[21] Patel, N.A. and Patel, R. 2018. A survey on fake review detection using machine learning techniques. In *2018 4th Int. Conf. on Computing Communication and Automation (ICCCA)*, IEEE, 1-6. DOI=IEEE.10.1109/CCAA.2018.8777594.