Opinion Spams Detection with Machine Learning

Xinyu Zhang, Yuhui Tang

Abstract

This project applies multiple machine learning and deep learning models to evaluate their performance on opinion spam detection using 'Gold-Standard' dataset. Different feature types (Term Frequency (TF) and Term Frequency-Inverse Data Frequency (TF-IDF)), number of maximum feature limitations and number of n-grams are evaluated in each model. Based on the modeling result, it is concluded that SVM and Logistic Regression have the best performance with larger number of n-grams. For deep learning models (CNN), unigram TF features outperforms the other models.

1 Introduction

Online Review an important source of information for customers to gain insights into the products they are planning to buy. The reviews include the quality of the product, shopping experience, service etc. The reviews will impact the potential customers positively or negatively, and therefore the profitability of the store in the future.

Usually, the reviews should reflect the true experience of each customers, which could provide a reference for the future customers. However, Fake Opinions and Spam Reviews will mislead the customers by either improve or damage the reputation of a business or a product.

Opinion spam has many forms such as fake reviews, fake comments, fake blogs, fake social network postings, etc.

In our project, we would like to use machine learning models to distinguish the true reviews and fake reviews.

2 Objective

Our goal is to distinguish fake reviews and true reviews from the hotel reviews. Six models will be applied: Support Vector Machine (SVM), Decision Tree, Logistic Regression, Xgboost, Random Forest and Convolutional Neural Network (CNN). From previous researches, SVM is widely use and usually has the best performance. Therefore, we would use SVM as our baseline model and compare its results with other machine learning and deep learning models that are not commonly used in the fake review detection area. We hypothesize that other classification models (such as tree-based models) and deep learning (CNN) will outperform the baseline model.

From the previous research, the authors mention that other content-based detection models involve semantic similarity metrics are used. Therefore, one other hypothesis is that we believe by adding some non-NLP features will help to improve the model.

In addition to those mentioned above which are modifications from previous literatures, we also try different feature number limitation to see how dimension impacts the model performance. Difference n-grams are applied (unigram, bigram, trigram and fourgram) to test on which n-gram tokenizer give the best performance.

3 Dataset

The dataset used in our project is 'Gold-standard' dataset. It is originally created by Ott et al. and widely used in Opinion Spam Detection research. This corpus consists of truthful and deceptive hotel reviews of 20 Chicago hotels.

The corpus contains four groups with equal number of reviews: 400 truthful positive reviews from TripAdvisor; 400 deceptive positive reviews from Mechanical Turk; 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp; and 400 deceptive negative reviews from Mechanical Turk.

Each of the above datasets consist of 20 reviews for each of the 20 most popular Chicago hotels.

4 Methodology

This part consists of four steps: Data Preprocessing, Exploratory Analysis, Feature Extraction and Modeling.

4.1 Data Preprocessing

The 'Gold-standard' dataset is preprocessed before further analysis. For each review in the dataset, all punctuations and non-letter characters are removed. These steps removes irrelevant information that exists in the data and reduces the size of the dataset for modeling. In addition, all the letters are set to be lowercase for consistency.

Next, we extract the stem of each words to get a standard form of words. This step is to convert the words into their original form and minimizes the number of distinct words in the dataset.

In addition, all the stop words are removed, as they are not critical to be used as features in text classification. For example, common stop words such as *a, an, are, at, as, on, that, these, those, too* are removed.

4.2 Exploratory Analysis

After initial cleaning and preprocessing the corpus, we conduct exploratory analysis to have a better understanding of the dataset.

For each group (negative deceptive reviews, negative true reviews, positive deceptive reviews, and positive true reviews), we first visualized top used single words in each group using word cloud, to compare the difference.

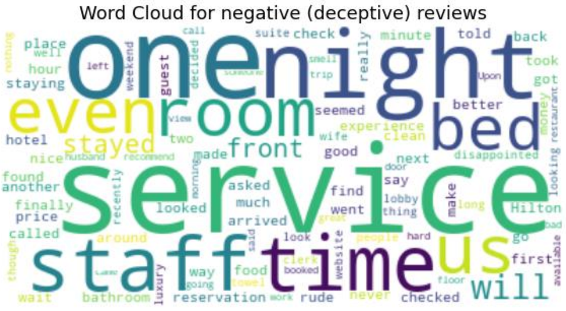


Figure 1: Word Cloud for negative (deceptive) reviews

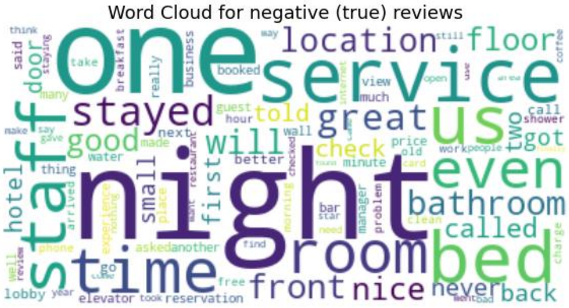


Figure 2: Word Cloud for negative (true) reviews

From the above two figures which each describes the top used words for negative deceptive reviews and negative true reviews, it is difficult to observe any differences between the uses of words. To be more specific, both groups have the most commonly used words such as *'one', 'night', 'service'*.

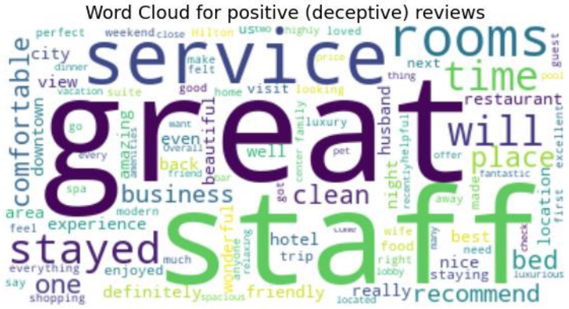


Figure 3: Word Cloud for positive (deceptive) reviews

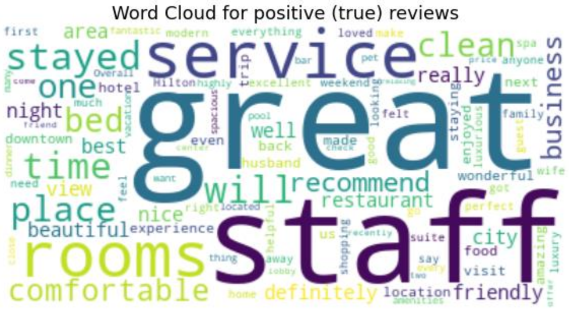


Figure 4: Word Cloud for positive (true) reviews

For the two groups with positive reviews, the commonly used words are also similar such as *'great', 'service', 'staff'*.

In addition, we also extract the top 20 bigrams for these four groups and expect to observe differences between deceptive and true reviews.

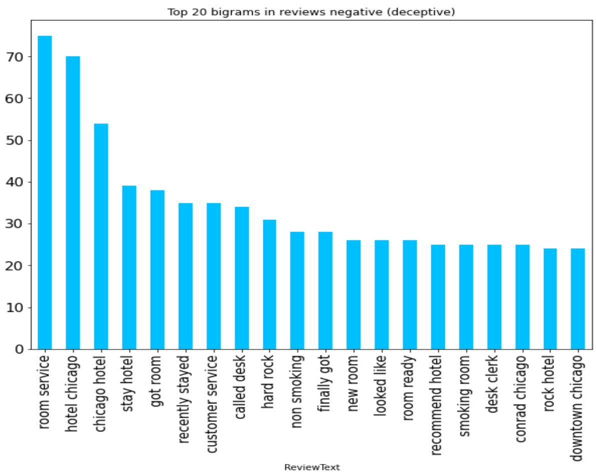


Figure 5: Top 20 bigrams in negative deceptive reviews

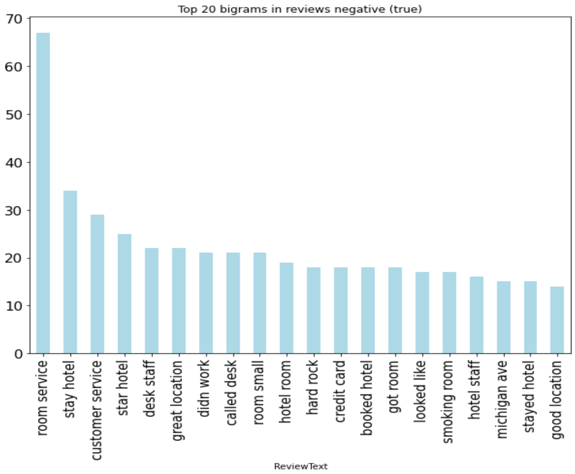


Figure 6: Top 20 bigrams in negative true reviews

By observing the bigrams of the two groups above, we can observe some differences in the selection of words, especially compared to the Word Cloud figure we discussed before. However, the evidence is not significant enough to successfully distinguish the deceptive and true reviews.

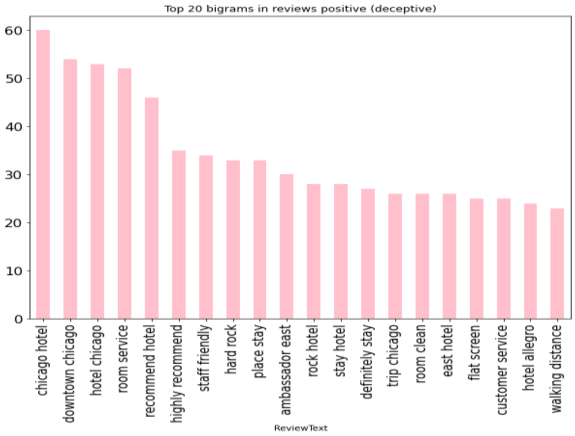


Figure 7: Top 20 bigrams in positive deceptive reviews

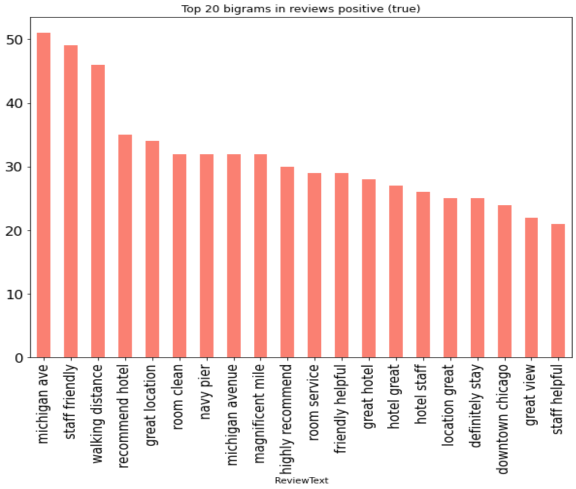


Figure 8: Top 20 bigrams in positive true reviews

For the deceptive and true reviews in the positive group, we have the similar conclusion that it is difficult to observe the differences by simply looking at the bigrams in both groups. Further analysis is required.

4.3 Feature Extraction

Before modeling, we extract NLP features and non-NLP features. For NLP features, we use Term Frequency (TF) and Term Frequency-Inverse Data Frequency (TF-IDF). TF calculates the ratio of the times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency. IDF is the log of the number of documents divided by the number of documents that contain the word w. Inverse data frequency determines the weight of rare words across all documents in the corpus.

For both TF and TFIDF, unigram, bigram, trigram and fourgram are extracted for our model, to test how the number of n-grams affect the model performance.

For non-NLP features, we use Sentiment polarity and the Length of review.

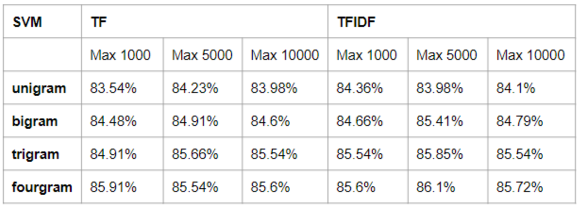
4.4 Modeling

Six classification models are applied: Support Vector Machine (SVM), Logistic Regression, Decision Tree, Xgboost, Random Forest and CNN. For each model, we compare model performance in respect of feature type and feature number limitations.

5 Result

For each model, we run with TF or TFIDF features with different feature number limitations and different number of n-grams. SVM serves as our baseline model. The metric we use to evaluate the success of the model is accuracy:

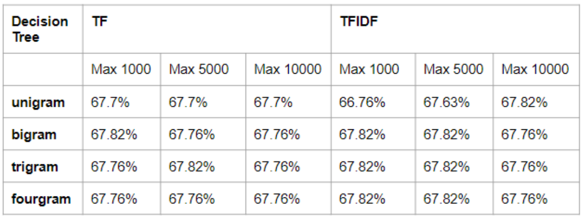
$$\text{Accuracy} = \frac{TP + TN}{Total\ number\ of\ observation}$$



SVM	TF			TFIDF		
	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	83.54%	84.23%	83.98%	84.36%	83.98%	84.1%
bigram	84.48%	84.91%	84.6%	84.66%	85.41%	84.79%
trigram	84.91%	85.66%	85.54%	85.54%	85.85%	85.54%
fourgram	85.91%	85.54%	85.6%	85.6%	86.1%	85.72%

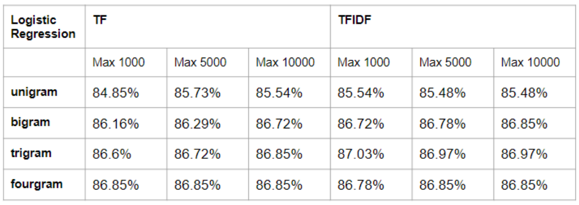
Figure 9: SVM results

Based on previous research, SVM has the best performance in distinguishing true and spam reviews. Our results with SVM model (Figure 9) have overall about 85% of accuracy. While the larger feature number limitation does not necessarily lead to better performance, the larger number of n-gram in general has stronger distinguishing power compared to smaller number of n-grams. For example, the result with fourgram features, on average, has better performance than those with unigram features.



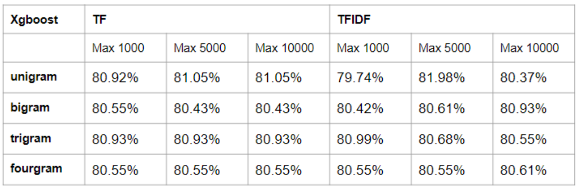
Decision Tree	TF			TFIDF		
	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	67.7%	67.7%	67.7%	66.76%	67.63%	67.82%
bigram	67.82%	67.76%	67.76%	67.82%	67.82%	67.76%
trigram	67.76%	67.82%	67.76%	67.82%	67.82%	67.76%
fourgram	67.76%	67.76%	67.76%	67.82%	67.82%	67.76%

Figure 10: Decision Tree results



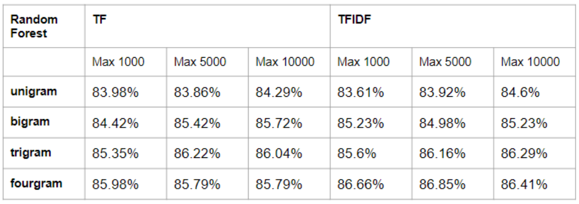
Logistic Regression	TF			TFIDF		
	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	84.85%	85.73%	85.54%	85.54%	85.48%	85.48%
bigram	86.16%	86.29%	86.72%	86.72%	86.78%	86.85%
trigram	86.6%	86.72%	86.85%	87.03%	86.97%	86.97%
fourgram	86.85%	86.85%	86.85%	86.78%	86.85%	86.85%

Figure 11: Logistic Regression results



Xgboost	TF			TFIDF		
	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	80.92%	81.05%	81.05%	79.74%	81.98%	80.37%
bigram	80.55%	80.43%	80.43%	80.42%	80.61%	80.93%
trigram	80.93%	80.93%	80.93%	80.99%	80.68%	80.55%
fourgram	80.55%	80.55%	80.55%	80.55%	80.55%	80.61%

Figure 12: Xgboost results



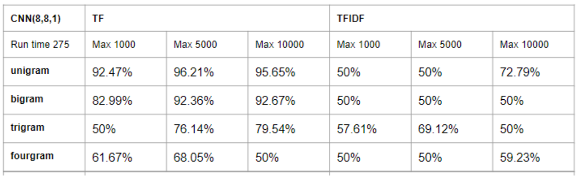
Random Forest	TF			TFIDF		
	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	83.98%	83.86%	84.29%	83.61%	83.92%	84.6%
bigram	84.42%	85.42%	85.72%	85.23%	84.98%	85.23%
trigram	85.35%	86.22%	86.04%	85.6%	86.16%	86.29%
fourgram	85.98%	85.79%	85.79%	86.66%	86.85%	86.41%

Figure 13: Random Forest results

Among the model results we use to compare with the benchmark, Logistic Regression outperforms the other models and takes least time to run. It even achieves overall better results than SVM which is regarded as the model with the best performance by previous researchers.

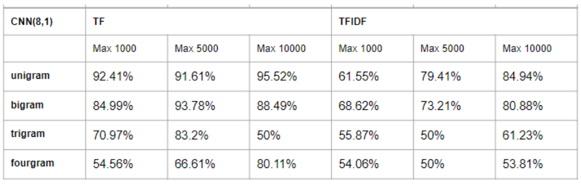
In tree-based models, Random Forest and Xgboost perform better than Decision Tree, the simplest tree-based model. However, they take much longer to run and do not have the best performance compared to SVM and Logistic Regression.

In addition to machine learning models, we also apply deep learning model (CNN) with different number of layers and neurons.



CNN(8,8,1)	TF			TFIDF		
Run time 275	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	92.47%	96.21%	95.65%	50%	50%	72.79%
bigram	82.99%	92.36%	92.67%	50%	50%	50%
trigram	50%	76.14%	79.54%	57.61%	69.12%	50%
fourgram	61.67%	68.05%	50%	50%	50%	59.23%

Figure 14: CNN with layers (8,8,1)



CNN(8,1)	TF			TFIDF		
	Max 1000	Max 5000	Max 10000	Max 1000	Max 5000	Max 10000
unigram	92.41%	91.61%	95.52%	61.55%	79.41%	84.94%
bigram	84.99%	93.78%	88.49%	68.62%	73.21%	80.88%
trigram	70.97%	83.2%	50%	55.87%	50%	61.23%
fourgram	54.56%	66.61%	80.11%	54.06%	50%	53.81%

Figure 15: CNN with layers (8,1)

From the CNN results from the above table, it is noticed that TF features significantly outperform the TFIDF features. Among different number f n-grams, unigram, surprisingly, has the best performance than the rest and achieved the highest accuracy of 96.21% with max 5,000 feature limitations with layers (8,8,1).

6 Discussion

From results in the previous, it is noticed that within machine learning models, SVM, our baseline model, and Logistic Regression have the highest overall accuracy and shortest running time. As advanced tree-based models, Random Forest and Xgboost do not outstand from other models. As they need to build multiple decision trees in the model, they also take relatively longer to run. Decision Tree, as the simplest tree-based model, has the lowest overall accuracy.

Deep learning models takes less time to perform and achieve better performance compared to all the machine learning models we performed, although not mentioned in many spam opinion detection researches, which may due to the lack of large amount of data to train on.

From the previous literature, SVM is widely used and regarded as the best model to distinguish true and spam reviews. In our project, we used tree-based models (Decision Tree, Random Forest and Xgboost) and deep learning which are not commonly mentioned by researchers in spam review detection fields. Tree based models do not have expected good performance, probably because the theory behind these models are not suitable in the context of spam detection where the features used are the frequency of n-grams in each reviews. Random Forest, for example, will only pick a subset of features out of the total features every time it build a decision tree. Therefore, some important n-grams of the review might not be considered when building the model, which lead to potential underperformance of the model.

For machine learning models, considering the different types of features and feature number limitations, there is not big difference between TF and TFIDF features. In addition, as the feature number limitations increases, the model performance does not have much variance. Models that trained by trigram and fourgram features perform better than those with unigram and fourgram. The reason might be trigram and fourgram contain more information about the review and therefore help to distinguish the true and spam reviews.

For deep learning models, the highest accuracy (96.21% with max 5,000 feature limitations) is achieved, but only limit to unigram and bigram with TF features. For TFIDF features, the accuracy is about 50%, which indicates that the model has no ability of distinguish spam and true reviews.

7 Conclusion

Deep learning models such as CNN achieved very different results with different number of feature limitations, feature types and number of n-grams. For example, it can achieve the accuracy as high as of 96.21% with TF unigram features, but under the same condition, the fourgram features only have 68.05% of accuracy.

Machine learning models have different results compared to deep learning models. SVM, as mentioned by previous literatures, and Logistic Regression have relatively better performance and less running time. Among different n-grams, it is noticed that trigram and fourgram usually outperform the unigram and bigram which contain less information about the corpus. There is no big difference between TF and TFIDF, and different feature number limitations.

For the future work, we'd like to investigate the models with top performances i.e. SVM, Logistic Regression and CNN with different layers and neurons. We will try to improve the model performance by tuning the key parameters.

In addition, more efforts on feature engineering are expected. Those features should have the ability to capture the authors' distinctive writing style and therefore help us to improve the model's strength in distinguishing spam and true reviews.

References

Ahmed H, Traore I, Saad S.Detecting opinion spams and fake news using text classification, Security and Privacy,2018;1:e9.https://doi.org/10.1001/spy2.9

Algur SP, Patil AP, Hiremath PS, Shivashankar S. Conceptual level similarity measure based review spam detection. Paper presented at: Signal and Image Processing (ICSIP), 2010 International Conference; December, 2010:416-423; Chennai, India: IEEE.

Li F, Huang M, Yang Y, Zhu X. Learning to identify review spam. Paper presented at: IJCAI Proceedings-International Joint Conference on Artificial Intelligence; July, 2011, Vol.22, No.3, p.2488, Barcelona, Spain.

Ott M, Choi Y, Cardie C, & Hancock JT. (2011,June).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 - Volume1 (pp.309– 319).Association for Computational Linguistics.

Shojaee S, MuradMA A, AzmanA B, SharefN M, Nadali S. Detecting deceptive reviews using lexical and syntactic features. Paper presented at: Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference; December, 2013:53-58; Seri Kembangan, Malaysia:IEEE.