Synthetic Review Spamming and Defense

Alex Morales, Huan Sun, and Xifeng Yan
Dept. of Computer Science
University of California, Santa Barbara
{alex_morales, huansun, xyan}@cs.ucsb.edu

ABSTRACT

Online reviews are widely adopted in many websites such as Amazon, Yelp, and TripAdvisor. Positive reviews can bring significant financial gains, while negative ones often cause sales loss. This fact, unfortunately, results in strong incentives for opinion spam to mislead readers. Instead of hiring humans to write deceptive reviews, in this work, we bring into attention an automated, low-cost process for generating fake reviews, variations of which could be easily employed by evil attackers in reality. To the best of our knowledge, we are the first to expose the potential risk of machine-generated deceptive reviews. Our simple review synthesis model uses one truthful review as a template, and replaces its sentences with those from other reviews in a repository. The fake reviews generated by this mechanism are extremely hard to detect: Both the state-of-the-art machine detectors and human readers have an error rate of 35%-48%. A novel defense method that leverages the difference of semantic flows between fake and truthful reviews is developed, reducing the detection error rate to approximately 22%. Nevertheless, it is still a challenging research task to further decrease the error rate.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; K.4.1 [**Computers and Society**]: Public Policy Issues—*abuse and crime involving computers*

Keywords

Review Spam; Spam Detection; Classification

1. AUTOMATED REVIEW GENERATION

We demonstrate an automatic review synthesis model in Figure 1.

[***Review pool***] We first collect truthful reviews from online websites like TripAdvisor with high positive scores and containing more than 150 characters. [***Base review***] A base review is randomly drawn from the pool, based on which a synthetic review will be generated. [***Synthesizer***] A synthesizer takes the base review as a template and synthesizes a new one using the reviews in the pool. In our work, we adopt one simplest strategy: The synthesizer replaces each sentence in a base review by the most similar (not exactly the same) sentence in the review pool. For the similarity metrics, we take into account both cosine similarity between two sentence vectors, and set similarity (the number of overlapped words in two sentences). A synthetic review is output after a full replacement of sentences in the base review. In practice, one might need to do location/name check in synthesized reviews whereas in this work we put little emphasis on this issue. Our to-be-proposed detection methods will not rely on any location/name information.

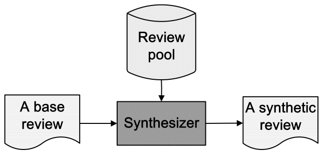

Figure 1: Review Synthesization

We tested human performance on the fake reviews generated by the above method. Ten volunteers are solicited. Figure 2 shows the average error rate is around 48%, where error rate is defined as the percentage of misclassified reviews over all the reviews. Readers are also welcome to try the synthetic review detection task via `www.cs.ucsb.edu/~alex_morales/reviewspam/`.

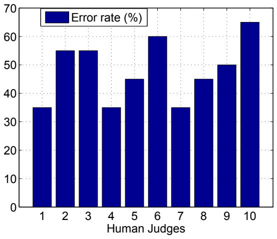

Figure 2: Error Rate of 10 Human Judges

We examined the performance of the state-of-the-art fake review detectors [4, 2, 1] on the synthesized reviews. Table 1 shows the result of three algorithms. While the computational approaches outperform human readers, their performance is not impressive.

One might think about generating a synthetic review by simply duplicating the base review. However, a duplication of an entire review could be detected more easily. In contrast, using sentence-wise replacement, one can significantly

Copyright is held by the author/owner(s).
WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

Algorithms	Error rate(%)
Ott *et al.* [4]	40.5
Liu *et al.* [2]	34.5
Harris *et al.* [1]	43.3

Table 1: Error Rate of the Existing Detectors.

increase the fake review space and detection difficulty. To pass those sentence-level duplication detectors, one could further use automatic rewriting/paraphrasing techniques , e.g., synonym replacement. Details involving sentence paraphrase and paraphrase detection will deviate too much from the current focus of this work. We stay focused on the simple sentence replacement strategy and resort to feature-based defense techniques. Nonetheless, our to-be-proposed methodology is directly applicable to paraphrased synthetic reviews.

2. SYNTHETIC REVIEW DETECTION

Synthetic reviews using sentence transplants bear subtle semantic incoherence between sentences. Based on this intuition, we advocate a general methodology for coherence analysis, which consists of two components: pairwise sentence coherence and multiple sentence coherence. Figure 3 shows the framework. Each filled circle denotes one sentence in a review. f denotes a general measure (feature) that is imposed on either a sentence pair or multiple sentences.

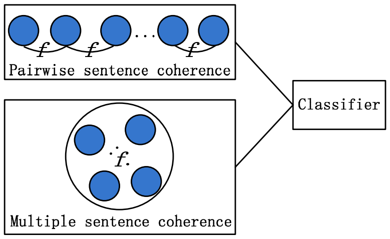

Figure 3: Illustration of Our Methodology.

Pairwise sentence coherence evaluates the information flow smoothness between two sentences:

Sentence transition: Given a word in one sentence, one could expect to observe certain words in its following sentence with some probability. The pointwise transition probability matrix records one-step transition probability from word w_i to word w_j in each element (i, j). We propose a measure, denoted as *ptp*, based on pointwise transition probabilities.

Word co-occurrence: Words generally demonstrate co-occurrence patterns (joint probability) in two consecutive sentences, which can be employed for coherence measurement. Based on this, a sentence co-occurrence score, denoted as *sco*, is proposed.

Pairwise sentence similarity: Subtly different from the transition and co-occurrence properties, pairwise similarity (SIM) measures the word/semantic overlap between two consecutive sentences. We take into account several variations for computing pairwise sentence similarity including word overlap between sentences, WordNet[3]-based similarity, and latent semantic similarity based on Latent Semantic Indexing.

Multiple sentence coherence measures the stretch and changes of topics in multiple consecutive sentences:

Semantic dispersion: Given a vectorized semantic representation of each sentence (e.g., topic distribution), we quantify how dispersed/focused the content of a review is. A truthful human-written review should be neither too diversified nor too focused. We define the semantic dispersion (*sd*) as the average distance between each sentence vector and the centroid of all the sentence vectors.

3. EXPERIMENTS

We collected 12,500 reviews of hotels located in New York City from TripAdvisor.com. 10 datasets created based on this collection are employed: Each dataset contains 500 truthful reviews and 500 fake reviews synthesized following the pipeline in Figure 1. Measures (features) proposed in Section 2 are computed for each review. Based on the features, we classify one review as truthful or synthetic. The average classification result on the 10 datasets is reported.

We tested various feature combinations under different well-known classifiers such as SVM and Naive Bayes classifier. It turns out the feature combination *ptp+sco+sd* achieves the lowest error rate **22%** under a linear SVM. To show the adaptability of our method, we train the classifier using one of the above 10 datasets, and test it using reviews from Washington D.C.. We obtain an average error rate 26.7% for *ptp+sco+sd* and over 40% for the other methods, which shows that our method is also promising in cases where training datasets are not quite relevant to testing datasets.

4. CONCLUSION

In this paper, we first bring into attention a simple yet powerful review synthesis technique. Furthermore, we propose a general framework instantiated by new coherence measures to detect such automatically synthesized reviews. Compared with the existing spam detectors, the classifier built on our new coherence features can reduce the error rate from 35%-48% to roughly 22%. While our method achieves the initial success, it is still an open research problem to further improve the detection accuracy.

5. ACKNOWLEDGEMENTS

This research was sponsored in part by the U.S. National Science Foundation under grant IIS-0917228.

6. REFERENCES

[1] C. Harris. Detecting deceptive opinion spam using human computation. In *Workshops at AAAI on Artificial Intelligence*, 2012.

[2] J. Liu, Y. Cao, C. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, pages 334–342, 2007.

[3] G. Miller and C. Fellbaum. Wordnet: An electronic lexical database, 1998.

[4] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *ACL-HLT*, pages 309–319, 2011.