**OPINION SPAM DETECTOR: DETECTING SPAM REVIEWS**

Natural Language Processing Project

Submitted by:

**FELARCA, JOHN DENVER**

**MILITAR, JESSICA JUSTINE**

**SANTOS, CLINT LENNARD**

**ZUBIRI, CEASAR JEMRICK**

BSCS 3-4

Submitted to:

**PROFESSOR RIA A. SAGUM**

1. **INTRODUCTION AND BACKGROUND**

It has become a common practice for people to read online opinions/reviews for different purposes. For example, if one wants to buy a product, one typically goes to a review site (e.g., amazon. com) to read some reviews of the product. If most reviews are positive, one is likely to buy the product. If most reviews are negative, one will almost certainly not buy it. Positive opinions can result in significant financial gains and/or fames for businesses, organizations, and individuals. This, unfortunately, gives strong incentives for opinion spamming. Opinion spam can range from annoying self-promotion of an unrelated website or blog to deliberate review *fraud.*

This project is focused on a potentially more insidious type of opinion spam: *DECEPTIVE OPINION SPAM*—fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader. For example, one of the following two hotel reviews is truthful and the other is *deceptive opinion spam* (Liu, 2008).

1. I have stayed at many hotels traveling for both business and pleasure and I can honestly say that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to bot business travelers and couples.
2. My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn’t ask for more!! We will definitely be back to Chicago and we will for sure be back to the James Chicago.

*(The answer to the question above: No. 2 is phony.)*

Typically, these deceptive opinions are neither easily ignored nor even identifiable by a human reader; consequently, there are few good sources of labeled data for this research. Indeed, in the absence of gold-standard data, related studies have been forced to utilize ad hoc procedures for evaluation.

To obtain a deeper understanding of the nature of deceptive opinion spam, we explore the relative utility of three potentially complementary framings of our problem.

We have used supervised learning, textual analysis, and relational modeling to solve the problem. Below are some main methods that we have used:

* In machine learning, ***support vector machines*** (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.
* In statistics, l**ogistic regression**, or logit regression, or logit model[1] is a regression model where the dependent variable (DV) is categorical. This article covers the case of a binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick. Cases where the dependent variable has more than two outcome categories may be analysed in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression.[2] In the terminology of economics, logistic regression is an example of a qualitative response/discrete choice model.
* In machine learning, **naive Bayes classifiers** are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s,:488 and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features.

We compare the performance of each approach on our novel dataset. Particularly, we find that machine learning classifiers trained on features traditionally employed in (a) psychological studies of deception and (b) genre identification are both outperformed at statistically significant levels by n gram–based text categorization techniques. Notably, a combined classifier with both n-gram and psychological deception features achieves nearly 90% cross-validated accuracy on this task. In contrast, we find deceptive opinion spam detection to be well beyond the capabilities of most human judges, who perform roughly at-chance—a finding that is consistent with decades of traditional deception detection research.

We also present findings that are consistent with recent work highlighting the difficulties that liars have encoding spatial information. Lastly, our study of deceptive opinion spam detection as a genre identification problem reveals relationships between deceptive opinions and imaginative writing, and between truthful opinions and informative writing.

1. **REVIEW OF RELATED LITERATURE**

The opinion spam problem was ﬁrst formulated by Jindal and Liu in the context of product reviews, Jindal and Liu [1]. By analyzing several million reviews from the popular Amazon.com, they showed how widespread the problem of fake reviews was. The existing detection methods can be split in the context of machine learning into supervised and unsupervised approaches. Second, they can be split into three categories by their features: behavioral, linguistic or those using a combination of these two. They categorized spam reviews into three categories: non-reviews, brand-only reviews and untruthful reviews. The authors ran a logistic regression classiﬁer on a model trained on duplicate or near-duplicate reviews as positive training data, i.e. fake reviews, and the rest of the reviews they used as truthful reviews. They combined reviewer behavioral features with textual features and they aimed to demonstrate that the model could be generalized to detect non-duplicate review spam. This was the ﬁrst documented research on the problem of opinion spam and thus did not beneﬁt from existing training databases. The authors had to build their own dataset, and the simplest approach was to use near-duplicate reviews as examples of deceptive reviews. Although this initial model showed good results, it is still an early investigation into this problem.

Lim et al. [2] is also an early work on detecting review spammers which proposed scoring techniques for the spamicity degree of each reviewer. The authors tested their model on Amazon reviews, which were initially taken through several data preprocessing steps. In this stage, they decided to only keep reviews from highly active users - users that had written at least 3 reviews. The detection methods are based on several predeﬁned abnormalities indicators, such as general rating deviation, early deviation - i.e. how soon after a product appears on the website does a suspicious user post a review about it or very high/low ratings clusters. The features weights were linearly combined towards a spamicity formula and computed empirically in order to maximize the value of the normalized discounted cumulative gain measure. The measure showed how well a particular ranking improves on the overall goal. The training data was constructed as mentioned earlier from Amazon reviews, which were manually labeled by human evaluators. Although an agreement measure is used to compute the inter-evaluator agreement percentage, so that a review is considered fake if all of the human evaluators agree, this method of manually labeling deceptive reviews has been proven to lead to low accuracy when testing on real-life fake review data. First, Ott et al. [3] demonstrated that it is impossible for humans to detect fake reviews simply by reading the text. Second, Mukherjee et al. [4] proved that not even fake reviews produced through crowdsourcing methods are valid training data because the models do not generalize well on real-life test data.

Xie et al. [5] observed that the vast majority of reviewers (more than 90% in their study or resellerratings.com reviews up to 2010) only wrote one review, so they have focused their research on this type of reviewers. They also claim, similarly to Feng et al. [10], that a ﬂow of fake reviews coming from a hired spammer distorts the usual distribution of ratings for the product, leaving distributional traces behind. Xie et al. observed the normal ﬂow of reviews is not correlated with the given ratings over time. Fake reviews come in bursts of either very high ratings, i.e. 5-stars, or very low ratings, i.e. 1-star, so the authors aim to detect time windows in which these abnormally correlated patterns appear. They considered the number of reviews, average ratings and the ratio of singleton reviews which stick out when looking over diﬀerent time windows.

Ott et al. [3] used a bag-of-words approach and calculated the frequency of certain words from the review text. They then classiﬁed some reviews as suspicious if the text contained a high number of predeﬁned suspicious words. This led to more subjective conclusions that spammers prefer to use more personal pronouns than genuine reviewers or they usually write reviews of more than 150 characters on average. The authors cataloged some words, e.g. “vacation” and “husband” as highly suspicious. They concluded these couple of words appeared more often in the fake reviews created through Amazon Mechanical Turk, but one can hardly say that a review containing the word “vacation” is 100% fake. An obvious aspect is that once the spammers ﬁnd out about these textual frequency traps which cause suspicion, they will simply avoid them.

1. **OBJECTIVES**
2. **General Objective**

To detect opinion spam of reviewers using bag-of-words approach and to device a software that will detect spam reviews.

1. **Specific Objectives**
2. To be able to identify the authenticity of a certain review and separate it from spams
3. To be able to extract useful information through the use of Natural Language Processing methods.
4. To further improve the accuracy, and also look into spam in other kinds of media, e.g., forums and blogs.
5. **SCOPE AND LIMITATIONS**

In this work, we study review spam. Our objective of this work is to highlight review spam in order to shed some light on the trustworthiness of on-line reviews and to detect possible spam activities. We propose to perform spam detection based on duplicate finding and classification. For classification, we regard spam detection as a 2-class classification problem, spam and non-spam. Support Vector Machine is applied to learn a predictive model.

Possible directions for future work include an extended evaluation of the methods proposed in this work to both negative opinions, as well as opinions coming from other domains. Many additional approaches to detecting deceptive opinion spam are also possible, and a focus on approaches with high deceptive precision might be useful for production environments.

1. **TERMINOLOGY**

Terms here are conceptually and operationally defined for better understanding of the reader(s).

* ***Dataset***

-is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

* ***Data Clustering***

-Data clustering is the process of making a group of data into classes of similar objects. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.

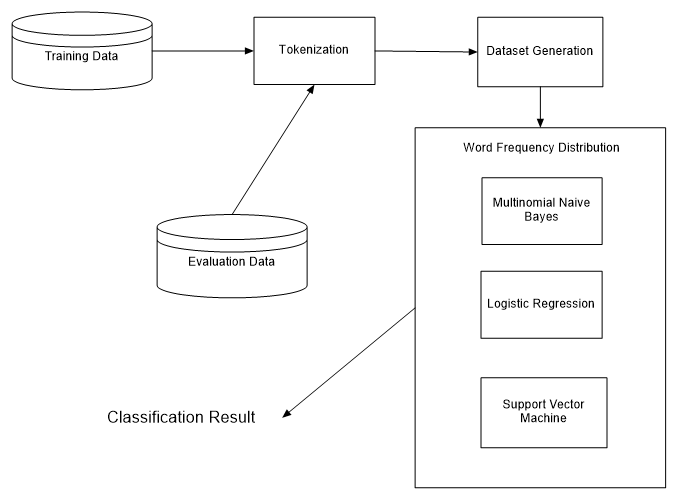
* ***Machine Learning***

-is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

* ***Support Vector Machine (SVM)***

-is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

1. **METHODS**
2. **System Architecture**

****

1. **Experimental Methodology**

We need training data to sharpen the accuracy of our system. Tokenization is the splitting of words to classify their functions in a sentence. Frequent Words Extraction counts the most frequent words used in the dataset and identify if they are frequently used in reviews or spams. Dataset Generation is the conversion of datasets from text files. Multinomial Naïve Baise Classifier, Support Vector Machine, and Logistic Regression are the classifications of datasets from the reviews. Evaluation Data evaluates data from the datasets and identifies reviews from spams. Data results outputs the results evaluated by the evaluation data.

1. **PROJECT PLAN**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task Name** | **Start** | **End** | **Days** | **Status** |
| Project Initiation | 11/22 | 11/25 | 4 | Completed |
| Requirements Gathering | 12/5 | 12/8 | 4 | Completed |
| Scheduling and Assigning Tasks | 12/13 | 12/14 | 1 | Completed |
| Analysis and Design | 12/15 | 12/22 | 7 | Completed |
| Data Preprocessing | ¼ | 1/10 | 7 | Completed |
| Training Data | 1/10 | 1/18 | 9 | In progress |
| Data Analysis | 1/18 | 1/23 | 6 | Not started |
| Testing Data | 1/24 | 1/25 | 1 | Not started |
| Checking of Level of Accuracy | 2/7 | 2/10 | 4 | Not started |
| System Development | 2/11 | 2/18 | 8 | Not started |
| Documentation | 2/24 | 2/28 | 4 | Not started |

# **REFERENCES**

[1] 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). New York, NY, USA: ACM. doi:10.1145/1341531.1341560

[2] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM ’10, pages 939–948, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871557.

[3] 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 - Volume 1 (pp. 309–319). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002512

[4] Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Spotting Opinion Spammers Using Behavioral Footprints. In Proceedings of the

[5] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. Review spam detection via time series pattern discovery. In Proceedings of the 21st International Conference Companion on World Wide Web, WWW ’12 Companion, pages 635–636, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. doi: 10.1145/2187980.2188164.

[6] Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting Fake Reviewer Groups in Consumer Reviews. In Proceedings of the 21st International Conference on World Wide Web (pp. 191–200). New York, NY, USA: ACM. doi:10.1145/2187836.2187863