Classification for the Detection of Opinion Spam

Pablo Pardos Medem - 8453586
Kasper van der Pol - 7969996
David Valero Martinez - 6548601

Utrecht University

October 2024

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce accumsan sit amet metus sit amet lacinia. Vestibulum ornare, velit non placerat aliquam, quam purus fringilla enim, et bibendum sem erat at mauris. Sed sagittis ac lectus at ornare. Praesent id lacus nibh. Donec quis lobortis dui. Maecenas luctus tellus odio. Pellentesque imperdiet, justo a pharetra aliquet, nisi orci posuere lacus, in egestas eros nulla sit amet libero.

Etiam feugiat hendrerit scelerisque. Integer eu metus id elit vulputate suscipit. Donec pulvinar pretium gravida. Aenean elementum arcu in vestibulum venenatis. Nulla quam lacus, vestibulum vitae odio ut, ultricies suscipit libero. Nam arcu erat, vulputate eu turpis vulputate, pellentesque malesuada dolor. Duis condimentum est venenatis hendrerit accumsan. Phasellus at pretium neque. Fusce sed nibh mattis, rutrum augue lobortis, convallis arcu.

Cras malesuada sagittis risus, a rhoncus diam rhoncus quis. Phasellus nulla turpis, elementum at urna feugiat, malesuada tristique tortor. Morbi sodales euismod sem.

1 Introduction

In the modern age, technolog With the advent of generative AI, generating fake reviews is becoming easier than ever. Tools like ChatGPT allows malicious agents to automate the process of writing fake reviews on a large scale, tricking costumers into purchasing low-quality or even actively harmful products [1].

It is, therefore, imperative to develop methods to accurately discern between genuine and deceptive reviews online.

2 Data Description

For this study, we employ the dataset assembled by Otis et al. [2] containing a collection of negative reviews of different hotels in the city of Chicago, Illinois. The first part of the dataset consists of genuine reviews collected from the following review websites: Expe-dia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp. The second part consists of fake reviews created by anonymous workers from Amazon's crowdsourcing platform Mechanical Turk.

In total, the dataset contains 800 reviews, 400 truthful and 400 deceptive, saved as `.txt` files. The dataset is divided into five folds, each with 160 reviews. Folds 1 to 4 are used as the training data for our model, while fold 5 is used as the test data.

3 Methodology

The selected models for this study are:

1. Multinomial naive Bayes (generative linear classifier).
2. Logistic regression with Lasso penalty (discriminative linear classifier).
3. Classification trees (non-linear classifier).
4. Random forests (ensemble of non-linear classifiers).

For each model, we prepare two implementations: one with unigram features and one with bigram features.

4 Results

As performance measures, we employ accuracy, precision, recall and the F1 score.

5 Discussion

6 Conclusion

References

[1] S. Lazzaro, "Generative AI is accelerating the spread of fake reviews and malicious apps," Fortune, Aug. 29, 2024. https://fortune.com/2024/08/29/generative-ai-fake-reviews-bad-apps/ (accessed Oct. 24, 2024).

[2] M. Ott, C. Cardie, and J.T. Hancock. 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.