


# Deceptive Opinion Spam Corpus v1.4

- [Download TAR Ball](#)
- [Download ZIP File](#)
-  [e-mail](#)

---

## Overview

This corpus consists of truthful and deceptive hotel reviews of 20 Chicago hotels. The data is described in two papers according to the sentiment of the review. In particular, we discuss positive sentiment reviews in [1] and negative sentiment reviews in [2].

While we have tried to maintain consistent data preprocessing procedures across the data, there *are* differences which are explained in more detail in the associated papers. Please see those papers for specific details.

This corpus contains:

- 400 truthful positive reviews from TripAdvisor (described in [1])
- 400 deceptive positive reviews from Mechanical Turk (described in [1])
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp (described in [2])
- 400 deceptive negative reviews from Mechanical Turk (described in [2])

Each of the above datasets consist of 20 reviews for each of the 20 most popular Chicago hotels (see [1] for more details). The files are named according to the following conventions:

- Directories prefixed with `fold` correspond to a single fold from the cross-validation experiments reported in [1] and [2].
- Files are named according to the format `%c_%h_%i.txt`, where:
  - `%c` denotes the class: (t)ruthful or (d)eceptive
  - `%h` denotes the hotel:
    - `affinia`: Affinia Chicago (now MileNorth, A Chicago Hotel)
    - `allegro`: Hotel Allegro Chicago - a Kimpton Hotel
    - `amalfi`: Amalfi Hotel Chicago
    - `ambassador`: Ambassador East Hotel (now PUBLIC Chicago)
    - `conrad`: Conrad Chicago
    - `fairmont`: Fairmont Chicago Millennium Park
    - `hardrock`: Hard Rock Hotel Chicago
    - `hilton`: Hilton Chicago
    - `homewood`: Homewood Suites by Hilton Chicago Downtown
    - `hyatt`: Hyatt Regency Chicago
    - `intercontinental`: InterContinental Chicago

- james: James Chicago
- knickerbocker: Millennium Knickerbocker Hotel Chicago
- monaco: Hotel Monaco Chicago - a Kimpton Hotel
- omni: Omni Chicago Hotel
- palmer: The Palmer House Hilton
- sheraton: Sheraton Chicago Hotel and Towers
- sofitel: Sofitel Chicago Water Tower
- swissotel: Swissotel Chicago
- talbott: The Talbott Hotel
- %i serves as a counter to make the filename unique

## References

- [1] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. 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.
- [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.

## License

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/> or send a letter to Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.

If you use any of this data in your work, please cite the appropriate associated paper (described above).

Theme by [orderedlist](#)