Case 11 - Detecting Fake Reviews for Hotels using Text Analytics

Situation

There is not a day that passes by that fake news and fake content is not mentioned in the popular press. Platforms such as Facebook and Tripadvosr generate tons of user generated content. If this content is fake, it breaks the users' trust in the platform. In the case of sites such as TripAdvisor they serve as important gatekeepers to the entire hotel industry. Lots of make hotel decisions based on reviews on sites such as TripAdvisor

Complication

It is being proclaimed that many of the reviews that one sees on TripAdvisor are fake. While the management understands the power of machine learning to say detect spam versus no-spam email (this was the first application of AI at Google), they do not know how to apply this technology to automatically detect/flag fake reviews.

Key question

Can we build a classifier to detect a truthful versus deceptive/fake review for hotels?

Data

A key challenge here is the need for labeled training data that has 'genuine fake' reviews and genuine genuine reviews!' This seems like a conundrum, right? Presumably if we knew that a certain set of reviews on platforms such as TripAdvisor were fake than we could use that labeled data and try to find out. B

A key innovation here was to generate genuine fake reviews by under-cover hiring workers on Amazon Mechanical Turk and paying them to write fake reviews for hotels in Chicago. They were told that they have to make the reviews believable and to pretend as if they were real guests that were either happy or disgusted with the stay

The resulting dataset[1] contains 400 Truthful positive , 400 Truthful negative , 400 Deceptive positive and 400 Deceptive negative reviews of the customers from 20 most popular hotels in Chicago. The authors state, “"To solicit **gold-standard** deceptive opinion spam using AMT, we create a pool of 400 Human- Intelligence Tasks (HITs) and allocate them evenly across our 20 chosen hotels. To ensure that opinions are written by unique authors, we allow only a single submission per Turker. We also restrict our task to Turkers who are located in the United States, and who maintain an approval rating of at least 90%. Turkers are allowed a maximum of 30 minutes to work on the HIT, and are paid one US dollar for an accepted submission.

Each HIT presents the Turker with the name and website of a hotel. The HIT instructions ask the Turker to assume that they work for the hotel's marketing department, and to pretend that

[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, available at https://myleott.com/op_spamACL2011.pdf

their boss wants them to write a fake review (as if they were a customer) to be posted on a travel review website; additionally, the review needs to sound realistic and portray the hotel in a positive light."

Target Variable - 'deceptive' column is the dependent variable and our task is to Predict if a review is truthful or deceptive.

Links to dataset - deceptive-opinion.csv (see folder)

Deck

Code

Hotel-Reviews-bigrams.RMD (see folder)

Notes and points for discussion

1. Notice that once we get past the data engineering task of systematically converting the ‘text’ to ‘numbers’ the same binary classification workflow that we have seen to be so effective works for us.
2. In this example I introduce the random forest model.
3. We can work a little harder to get the random forest to compute which of the many words in the reviews are more discriminating in classifying deceptive versus genuine reviews. This is called calculating feature importance. We use the permutation method here such that the values of features are randomly perturbed and then we see how much the model’s error increases. The more the increase, the more critical the feature is.
4. Can you think of other applications of this model? What other companies might you be able to sell this technology to?