The Enemy in Your Own Camp: How Well Can We Detect Statistically-Generated Fake Reviews – An Adversarial Study

Dirk Hovy
Center for Language Technology
University of Copenhagen
2300 Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Abstract

Online reviews are a growing market, but it is struggling with fake reviews. They undermine both the value of reviews to the user, and their trust in the review sites. However, fake positive reviews can boost a business, and so a small industry producing fake reviews has developed. The two sides are facing an arms race that involves more and more natural language processing (NLP). So far, NLP has been used mostly for detection, and works well on human-generated reviews. But what happens if NLP techniques are used to *generate* fake reviews as well? We investigate the question in an adversarial setup, by assessing the detectability of different fake-review generation strategies. We use generative models to produce reviews based on meta-information, and evaluate their effectiveness against deception-detection models and human judges. We find that meta-information helps detection, but that NLP-generated reviews conditioned on such information are also much harder to detect than conventional ones.

1 Introduction

Online reviews written by customers are a booming market. Several companies cater to a wide variety of audiences, supplying—among others—reviews for restaurants (Yelp), travel (TripAdvisor), businesses (Trustpilot), and specialized communities, such as beer (RateBeer). While the revenue of the providers is in the billions of dollars, the currency this industry is built on is consumer trust. The majority of consumers uses such reviews to inform themselves before buying.[1] Online review companies therefore put considerable effort into maintaining this trust, by addressing the greatest threat to consumer trust (and therefore income)—fake reviews.

Identifying fake reviews is a natural fit for NLP, since they presumably contain linguistic cues that indicate their nature. Indeed, a number of previous works have dealt with the detection of fake reviews (Jindal and Liu, 2007; Badaskar et al., 2008; Mackiewicz, 2008; Jindal et al., 2010; Ott et al., 2011; Fornaciari and Poesio, 2014). However, in those cases, human writers were producing reviews to fool a human audience, not an NLP model. The detection models were therefore able to exploit the regularities resulting from the writers' tendency to follow a pattern to minimize their effort.

Writing fake reviews has become a lucrative business (Streitfeld, 2012), and so there is now an arms race going on between producers and detectors (Roberts, 2012). What if fake review writers become aware of the ways to game a detection algorithm?[2] As NLP technology becomes more common, we should expect to also see fake reviews *generated* by NLP models. This pits technology against technology.

In this paper, we explore the impact fake review generation has on NLP models' ability to detect them, and an ethical challenge in our development of NLP technology: the fact that it can be used for both sides (Hovy and Spruit, 2016).

Our contributions We set up an adversarial evaluation approach inspired by (Smith, 2012), using graphical models to build various language models that generate fake reviews, with and without recurrence to meta-information. We then test

[1] http://www.business2community.com/infographics/impact-online-reviews-customers-buying-decisions-infographic-01280945

[2] Similarly, some members of the Mechanical Turk community have adapted to the presence of assessment tools.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 351–356,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

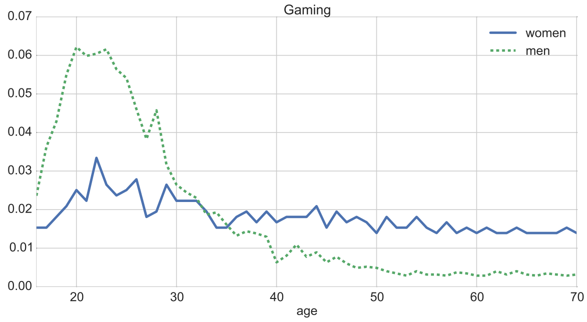

Figure 1: Age distribution of gaming reviews for men and women in the US

how well a logistic regression model can distinguish real from fake reviews from both models under two settings (the model has access to meta-information or not), and how well human judges can detect fake reviews generated by the model with meta-information.

Our results indicate that fake review generation could be a serious problem for detection mechanisms that solely rely on textual features.

2 Data

We use data extracted from the American and British versions of the review site Trustpilot.[3] It comprises reviews for a variety of online businesses, as well as information about users' age and gender.[4] We extracted all reviews that contained the full set of meta-information. We lower-case and tokenize by words, but leave reviews intact, rather than splitting them up into sentences. This results in 120,976 review instances. We reserve 10,000 for evaluation purposes, and use the rest to induce our adversarial generative models, and to derive features for the detection model (see below).

3 Review Generation Models

The basic approach is a simple Markov chain with a sufficiently large horizon to generate fluent reviews. Such an n-gram language model (LM) is a function that assigns probability to any sentence S, where S is a sequence $w_0, w_1, \cdots, w_n$, and

$$P(S) = \prod_i^N P(w_i|w_{i-n:i-1})$$

w_0 is a special (sequence of) start token(s), n is the size of the Markov horizon, and $w_{i-1:i-(n-1)}$ is the sequence of preceding words in context. The model is depicted in Figure 2 a). Since this is a generative model, it can not only be used to assign probability to observed sentences, but also generate new sentences based on the model parameters.

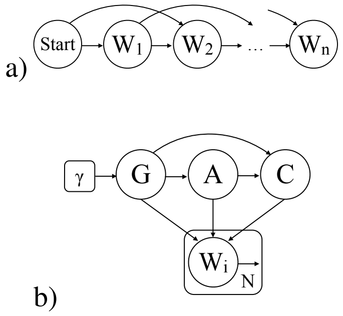

Figure 2: Regular n-gram Markov chain LM (top) and conditioned LM (bottom). γ based on empirically-observed gender distribution in data

However, extra-linguistic information, if available, can improve classification performance (Volkova et al., 2013; Hovy, 2015), and fake-review detection models often also exploit meta-information about the author and their behavior (Lim et al., 2010), looking for irregularities. We therefore use a generative story that assumes that people of different age and gender review different things, which in turn influences the type of business reviewed, and the choice of words. This assumption is borne out in the data (cf. Figure 1).

We extend this model by conditioning on latent variables age (A), gender (G), and review category (C).[5] In the generative story of this model, we first draw a user from one of the two genders in our data, select an age based on gender-specific age distributions, and choose a review category dependent on the two previous variables. We then then generate a sentence conditioned on all of these settings and the Markov horizon. Our model is depicted as plate diagram in Figure 2 b) (we omit the start token and the Markov horizon for clarity). It can formally be written as:

$$\begin{aligned} P(S|G,A,C) =& P(G) \cdot P(A|G) \cdot P(C|G,A) \\ & \cdot \prod_i^N P(w_i|w_{i-n:i-1}, C, G, A) \end{aligned}$$

[3] www.trustpilot.com

[4] For a more detailed description of the data, see Hovy et al. (2015).

[5] These factors could of course be extended to cover other information, including ratings. We do not condition on ratings here, but a commercial system for fake reviews would presumably be restricted to positive scores.

Category	Age group	Sex	Text
Hotels	30–45	F	*i ordered a new toner for our printer and after price matching the best i could find they also honoured a free delivery on top .*
Computer Accessories	45–60	M	*this is a company that takes customer care seriously . we received really impressive service when we contacted the company .*

Table 1: Generated examples from unconditional (top) and conditional (bottom) LM.

We can now use either model to generate fake reviews. In both cases, we use a Markov horizon of 6 words, i.e., a 7-gram model. For the unconditioned LM, meta-information is generated at random. This simulates a fake-review writer who is unaware of the context effects, but knows that companies might take profile information into account. Two examples are shown in Figure 1. Both examples are fluent, but the unconditioned one suffers from two problems: somewhat stream-of-consciousness-like sentences and a for human readers obvious mismatch between the category and the discussed topic.

Conditional LMs, on the other hand, suffer from a certain sparsity: the more meta-information we condition on, the sparser the n-gram counts become. They are therefore more likely to faithfully re-generate the training data. We use interpolation between genders, but this could also be addressed with a wide variety of techniques (Chen and Goodman, 1998).

Even though the classifier does not have access to the training data, we want to make the task as difficult as possible, so we remove all duplicates, as well as any generated reviews that do not end in a punctuation mark, that exceed 200 words, or that have a category not contained in our real-review test set, and select from the rest by lowest entropy.

4 Experiments

In our experiments, we pit an adversary (i.e., the two LMs we experiment with) against a judge (a classifier or human annotator). The goal of the adversary is to produce fake reviews that convince the judge.

4.1 Logistic Regression Model

As classifier, we use a logistic regression model, regularized with L_2 norm, and fit it on a data set of 10,000 true reviews and a varying amount of fake reviews. In one setting, we use 1600 fake reviews , based on current estimates of 16% (Luca and Zervas, 2015). In a second setting, we use 10,000 fake reviews, a scenario where 50% of all reviews are fake. Given the ease of generating fake reviews with the models presented here, the rate could quickly go up in the future, so this ration gives a bound on how much our detection models could decline.

The base features of the classifier are word n-grams, with n ranging from 1 to 4. Depending on the setting, we also add meta-information features, including combinations of the n-grams with each category (e.g., `category=Hotels & word="soft bed"`), and the average PMI score for the words in the sentence and each category (e.g., `PMI(Hotels, soft bed)`). That way, we hope to capture mismatches between the stated category and the review content.

We measure F1 performance over 5-fold stratified cross-validation.

Initially, we would like to establish whether conditioning LMs on demographic information has any effect on detection. For this purpose, we compare the performance of the logistic regression model on (1) a test set including fake reviews generated by an unconditioned 7-gram LM and (2) a test set whose fake reviews have been conditioned on meta-information. In both cases, the detector has only access to the base features, i.e., ignores demographic information. This is equivalent to a situation where the judge can only see the text, not the meta information.

As mentioned before, though, many companies employ meta-information in order to capture fake reviews, and if a spammer knew this, they could simply generate some meta-information. The question is: *does this meta-information have to follow a coherent generative story?* Intuitively, we expect the answer to be "*yes*": we would be surprised to see a teenager review retirement homes.

To test this assumption, we compare the performance of the classifier when having access to the base features plus meta-information under two settings: (3) with each piece of meta-information generated independently at random, and (4) with meta-information generated as part of our generative process.

4.2 Human Judges

The first experiment tests the detectability of fake reviews by statistical means. How hard is it for humans, though, to distinguish the fake reviews generated by this model from real reviews, and do they exploit meta-information?

To answer these questions, we also conduct a human judges study on Crowdflower[6]. We select 200 items at random (100 real reviews and 100 from the conditional model, half of each with meta-information), and ask annotators to rate them as real or fake. Judges were not informed about the nature of the reviews, only advised to use their best judgement. The task involved 8 test questions to bar bad annotators from entering. 76 unique judges participated, and rated the task as relatively difficult (3.5/5).

5 Results

Bear in mind that this is an adversarial setup: we are trying to improve the fake reviews to "trick" the judge into producing as many false positives as possible. A low F1-score thus means that the respective LM has managed to fool the classifier. Table 2 shows the results. Note also that the fake reviews are generated independent of the classification model, i.e., the generative LM does not take the classification model into account.

5.1 Logistic Regression Model

In order to assess whether the differences in performance are statistically significant, we conduct a bootstrap sampling test (Efron and Tibshirani, 1993) with 10,000 repetitions on the overall predictions.

The numbers are generally in a high range, which is encouraging, since it means that our models can detect fake reviews fairly reliably. However, we also see that the conditional models introduced here quickly become significantly harder to detect than the regular LM.

Adding meta-information leads to sometimes small, but always significant increases in performance. This effect is especially pronounced in the regular LMs, since the detection model is able to pick up on the mismatch between category and text content.

	16% fake		
Judge	cond. LM	reg. LM	$p < 0.01$
words only	88.27	87.89	no
+meta-info	88.92	92.42	yes
$p < 0.01$	yes	yes	
	50% fake		
words only	75.52	72.78	yes
+meta-info	77.40	88.43	yes
$p < 0.01$	yes	yes	

Table 2: Model performance (F1) with different amounts of information on reviews generated by regular or conditional model under two conditions

As the number of fake reviews grows, though, detection gets more difficult, and the rift between the two generation models becomes more apparent: at 50% fake reviews, the conditional LM is almost twice as hard to detect as the regular LM when using meta-information.

Feature Analysis Finally, we analyze the features (word-based and meta-information) to find the most predictive elements of fake reviews. For each feature, we average over all folds of our cross-validation. Features which are selected frequently, irrespective of the exact training conditions, can be assumed to be robust predictors.

Unsurprisingly, the most predictive features are `PMI(`*`gender`*`, ·)` and `PMI(`*`category`*`, ·)`, followed by gender- and age-specific words (`gender=M & word="delivery"`, `age-group=3 & word="."`), category-specific words (`category=Package Service & word="parcel"`), and individual words (`service`, `easy`, `quick`)

For the unconditional models, the PMI-category coefficients dominate other features in a power-law distribution, while for the conditional model, the PMI-age score is only slightly ahead.

5.2 Human Judges

The general tendency among human judges was to assume reviews are real: overall, 87% of the individual answers judged a review to be real. That

[6] `https://crowdflower.com`

is, only a minority of judges suspected fraud, irrespective of whether they had access to the meta-information or not. Whatever signal the logistic regression model picks up seems to be more subtle than what the average human can perceive.

This tendency plays out in the F1 scores (see Table 3): human judges have a much lower detection rate than the logistic regression model, even though the availability of meta-information improves performance here as well.

These results hold whether we treat each vote as an individual item or aggregating the five votes for each instance by an item-response model (Hovy et al., 2013). In the latter case, the performance for both conditions and the average increases, more so for the instances without meta-information, but still not reaching the same level.

ACCESS TO	RAW	AGGREGATED
words only	63.90	65.77
+meta-info	65.31	66.66
avg.	64.65	66.22

Table 3: Human performance (F1) with different amounts of information on reviews generated by conditional model

6 Related Work

Reviews are a rich source of studies for NLP, and a variety of recent papers (McAuley et al., 2012; Danescu-Niculescu-Mizil et al., 2013; Reschke et al., 2013; Jurafsky et al., 2014; Hovy et al., 2015) have explored it.

Badaskar et al. (2008) also use real and fake reviews and LMs, but in almost exactly the opposite setup: they select features that have high discriminative power in distinguishing real from fake reviews to include in their LMs. However, they use a review corpus that is more than an order of magnitude smaller, focus on tri- and quad-gram features, and do not take meta-information into account.

The work of Fornaciari and Poesio (2014) is similar in that they also deal with fake review detection. However, they do not use an adversarial setup, but focus on the use of an item-response model to detect fake-review writers. Their corpus is considerably smaller than ours, but the detection rate they report is similar to the one we find when not using meta-information.

To our knowledge, only Lappas (2012) has taken the view from the adversary's point of view, although the paper does not generate fake reviews, but assesses the presence of several defined measures of meretriciousness.

7 Conclusion

We have investigated the detectability of fake reviews generated with meta-information. We find that (1) using access to meta-information can significantly improve the detection of fake reviews, and (2) generated reviews conditioned on meta-information are considerably harder to detect than the ones generated without. We also see that statistical models fare better than human judges. Our results indicate the viability of an adversarial setup to test detection tasks, but also highlight the fact that NLP techniques can be used for either side. We should therefore be more vigilant and willing to play devil's advocate, pitting potential models as adversaries against our solutions.

Acknowledgments

The author would like to thank the members of the CoAStaL group and the anonymous reviewers for their detailed comments and suggestions, and Noah Smith for pointing out a missing reference. This research was funded under the ERC Starting Grant LOWLANDS No. 313695.

References

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Stanley Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. International World Wide Web Conferences Steering Committee.

Bradley Efron and Robert Tibshirani. 1993. *An introduction to the bootstrap*. Chapman & Hall, Boca Raton, FL.

Tomes Fornaciari and Massimo Poesio. 2014. Identifying fake amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. Beyond Privacy: The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL, Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the ACL*.

Nitin Jindal and Bing Liu. 2007. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM.

Nitin Jindal, Bing Liu, and Ee-Peng Lim. 2010. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1549–1552. ACM.

Dan Jurafsky, Victor Chahuneau, Bryan R Routledge, and Noah A Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4).

Theodoros Lappas. 2012. Fake reviews: The malicious perspective. In *Natural Language Processing and Information Systems*, pages 23–34. Springer.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM.

Michael Luca and Georgios Zervas. 2015. Fake it till you make it: Reputation, competition, and yelp review fraud. *Harvard Business School NOM Unit Working Paper*, (14-006).

Jo Mackiewicz. 2008. Reviewer motivations, bias, and credibility in online reviews. *Handbook of research on computer mediated communication*, 1:252–266.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1020–1025. IEEE.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey 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-Volume 1*, pages 309–319. Association for Computational Linguistics.

Kevin Reschke, Adam Vogel, and Dan Jurafsky. 2013. Generating recommendation dialogs by extracting information from user reviews. In *Proceedings of the 51st annual meeting of the ACL*, pages 499–504.

Jeff John Roberts. 2012. Amazon sues people who charge $5 for fake reviews. Fortune, October 19. `http://fortune.com/2015/10/19/amazon-fake-reviews/` Retrieved Feb 27, 2016.

Noah Smith. 2012. Adversarial evaluation for models of natural language. http://arxiv.org/abs/1207.0245.

David Streitfeld. 2012. The Best Book Reviews Money Can Buy. The New York Times, August 25. `http://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html` Retrieved Feb 27, 2016.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, pages 1815–1827.