Building a Better Lie Detector with BERT: A First Step to Finding the Rules of Deception

Dan Barsever
Cognitive Science
University of California, Irvine

Sameer Singh
Computer Science
University of California, Irvine

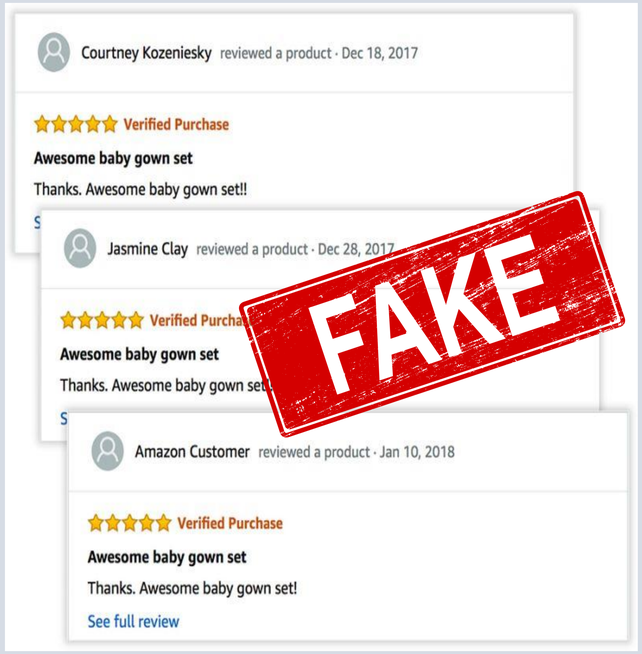
Emre Neftci
Cognitive Science
University of California, Irvine

Introduction

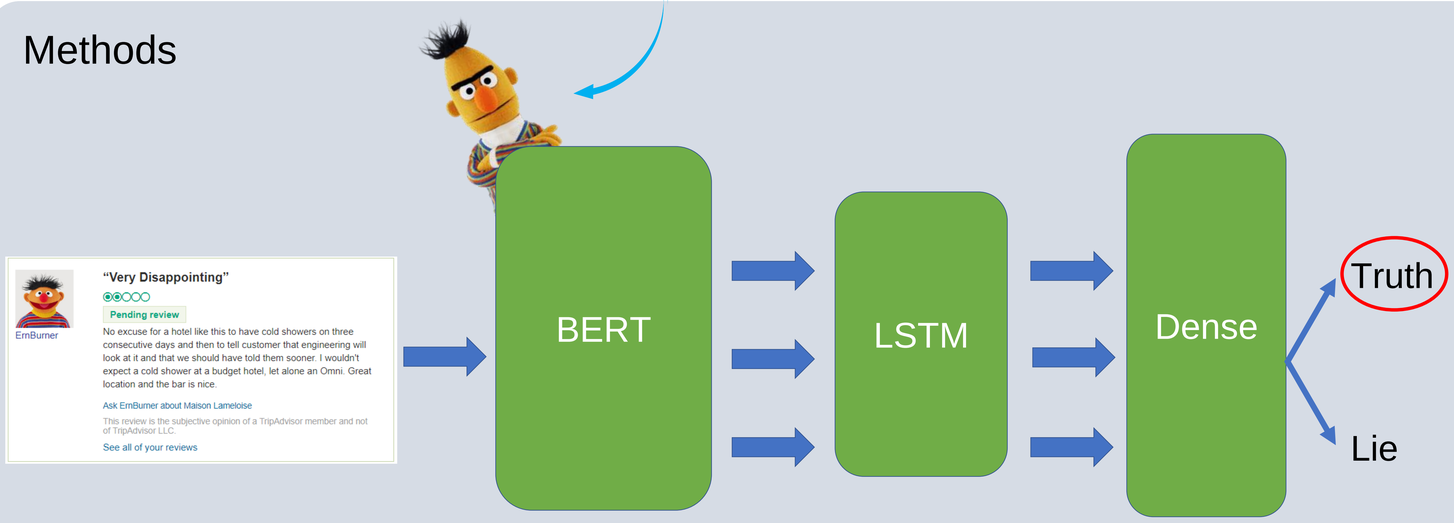
- False online reviews interfere with consumers' buying decisions, and they are difficult for humans to detect[1]
- The first step to understanding how these lies manifest in text is to build a tool that can detect fake reviews. For this we use BERT, a neural network that is good at language tasks[2]
- BERT will learn how to classify the text as truth or lie by constructing internal rules and features to differentiate the two
- BERT can then be analyzed to try and extract the rules that it creates to classify the text

Why BERT?

- BERT can be used out of the box
- BERT is able encode contextual information about a sequence



Methods

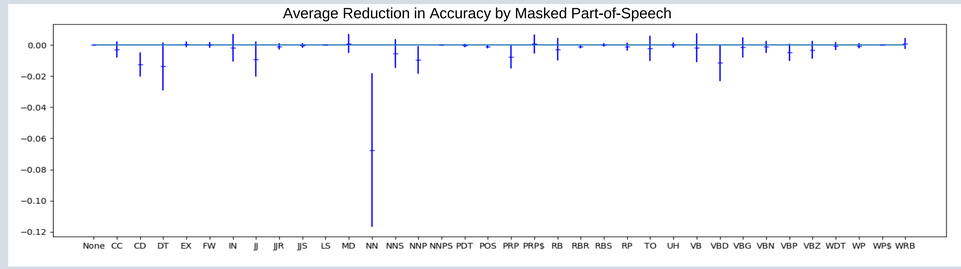


- We used the Ott Deceptive Opinion Spam of hotel reviews[3]
- We performed an ablation study on the trained network. Each part of speech was removed from the text in turn to see how its removal would affect the accuracy
- After that, we looked at what tokens the network was paying attention to for a given input to create saliency maps

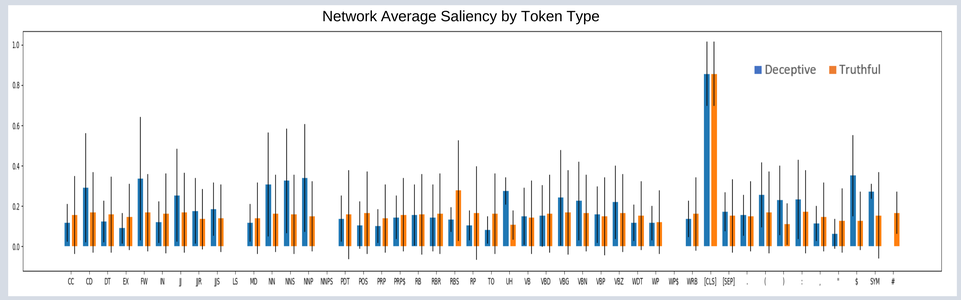
Results

BERT beat the state of the art with an accuracy of 93.6%, proving it can accurately classify deceptive text

Source	Accuracy
Ott[3]	89.8%
Vogler[4]	87.0%
Xu & Zhao[5]	91.6%
Ren & Ji[6]	85.7%
BERT	**93.6%**



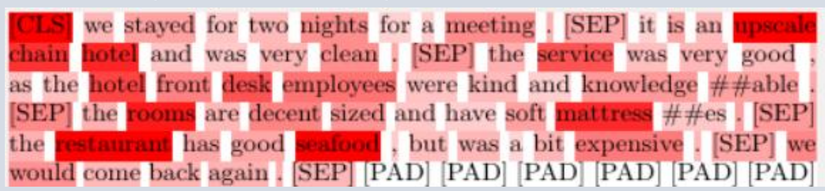
Only singular nouns (NN) had a strong effect in the ablation, perhaps because their removal hurts readability

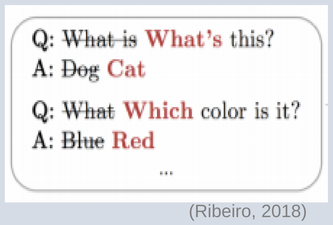


There are differences in saliency, but not enough to be informative

Future Work

Further analyze BERT's saliency maps to find patterns of attention





(Ribeiro, 2018)

Modify the input, substituting phrases that are similar in meaning but different semantically until something tips the classifier in the other direction

Test BERT on other corpora, like the Mafiascum dataset of online mafia games, in order to features in corpora of different natures[7]

To view the pdf of this poster and paper, scan this code

dbarseve@uci.edu

[1]Levine (2014). T. R. Levine and C. F. Bond. Direct and indirect measures of lie detection tell the same story: A reply to ten brinke, stimson, and carney (2014).Psychological science, 25(10):1960–1961, 2014. [2]Devlin (2018). J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. [3]Ott (2011). M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. 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, 2011. **[4]Vogler (2019)** N. Vogler and L. Pearl. Using linguistically-defined specific details to detect deception across domains. Natural Language Engineering, 1(1):1–32. **[5]Xu & Zhang (2012)** Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. Proceedings of COLING 2012: Posters, pages 1341–1350, 2012. **[6]Ren & Ji (2017)** . Ren and D. Ji. Neural networks for deceptive opinion spam detection: An empirical study.Information Sciences, 385:213–224, 2017. **[7]de Ruiter (2018)**. de Ruiter, Bob, and George Kachergis. "The Mafiascum Dataset: A Large Text Corpus for Deception Detection." *arXiv preprint arXiv:1811.07851* (2018). Ribeiro (2018). Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Semantically equivalent adversarial rules for debugging nlp models." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.