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

Dan Barsever
Department of Cognitive Sciences
University of California, Irvine

Sameer Singh
Department of Computer Sciences
University of California, Irvine

Emre Neftci
Department of Cognitive Sciences
University of California, Irvine

Abstract

Detecting lies in text is a valuable skill to have due to the prevalence of deceptive text online. This is partly because the patterns that underlie deceptive text are not known. The aim of this project is to investigate these underlying patterns of deception in text. The first step in this approach is to train an accurate classifier based on the BERT network. BERT is able to beat the state of the art in deception classification accuracy on the Ott Deceptive Opinion Spam corpus. Our preliminary results indicate that certain components of the input, such as some parts of speech, are more informative to the classifier than others.

1 Introduction

Most traditional methods of lie detection consist of analyzing a physiological response, such as sweat or heart rate. Comparatively little study has been made into detecting lies in text, where there are no physiological clues [7]. One area this affects everyday life is in false hotel reviews, or Deceptive Opinion Spam. This usually takes the form of a malicious customer posting fake negative reviews to hurt a business, or a company shill posting fake positive reviews to make the company look better. Humans are ineffective at detecting deceptive text, faring little better than chance [3, 6]. This is in stark contrast to other linguistic tasks such as sentiment analysis where humans perform extremely well [9]. Being able to more reliably detect deceptive text is beneficial for consumers and for businesses.

As a step to understanding how lies are expressed in text, we propose a machine learning tool utilizing BERT that can learn what constitutes a deceptive review. BERT (Bidirectional Encoder Representations from Transformers) is a recently developed neural network architecture that is pretrained on millions of words and is capable of forming different representations of text based on context [1]. By applying BERT to deception detection, we can use it to form a powerful classifier of deceptive text. After that, extracting the rules that BERT forms to classify the text can help us understand what patterns underlie deceptive text.

BERT has proven to be an accurate classifier, defeating the state of the art on the Ott Deceptive Opinion Spam corpus. The rules it uses to do so are still unclear, but a preliminary ablation study has indicated that certain parts of speech are more informative than others.

2 Related Work

Myle Ott [7] developed the Ott Deceptive Opinion Spam corpus, which consists of 800 true reviews from TripAdvisor and 800 deceptive reviews sourced from Amazon Mechanical Turk. He used this corpus to train Naïve Bayes and Support Vector Machine (SVM) classifiers, achieving a maximum accuracy of 89.8% with an SVM utilizing Linguistic Inquiry and Word Count (LIWC) combined with bigrams. The Ott corpus is one of the most commonly used gold-standard corpora in deception detection tasks. Other, less widespread corpora include the LIAR fake news dataset [10] and Feng et al.'s Yelp dataset [2].

Vogler and Pearl [9] used a support vector machine operating on linguistically defined features to classify the Ott corpus. They were able to achieve an accuracy of 87% using this method. Xu and Zhao [11] train a maximum entropy model on the Ott corpus and were able to achieve 91.6% accuracy. Li et al. [4] tried to find a general rule for identifying deceptive opinion spam using features like part-of-speech on several datasets including the Ott corpus, achieving 81.8% accuracy on Ott. Ren and Ji expand on Li et al. by using a recurrent neural network on the same data, improving

the accuracy to 85.7% [8].

3 Methods

The network we use for this project is based on BERT, with a bidirectional LSTM and dense linear on top of BERT as a classifier. BERT has several advantages over previous methods. First, BERT performs well in a wide variety of contextually sensitive language tasks due to being able to detect when the meaning of a sequence has changed depending on context, allowing it to detect subtle differences in phrasing [1]. BERT also requires significantly less preprocessing of data than previous methods. The primary idea behind most prior work is to extract predefined features (such as bigrams or part-of-speech counts) from a sample and classify according to those features. BERT requires no predefining of features and is free to develop its own rules. The BERT model we use is the bert-base-uncased pretrained BERT model for PyTorch provided by HuggingFace[1].

We used the Ott corpus to benchmark the network and compare it to previous approaches. 80% of the reviews form the train set, which will be used to train the network. The remaining 20% become the test set, used to evaluate the network. In each training epoch, the training set is presented to the network in random batches of 8 until the entire set has been presented. Training lasted for 100 epochs to ensure the stability of the final result.

As a preliminary investigation into which parts of the input are the most important, we also performed an ablation study on the network after training. In this study, we tagged each token of each review in the test set with its part of speech [5]. We then evaluated the accuracy of the network on the test set with each part of speech removed and replaced with a placeholder masking token. This ablation was done 10 times for each part of speech.

4 Results

BERT reached an accuracy of 93.6% (table 1), an improvement of 2% over the next best method, beating the state of the art in deception detection. This jump in accuracy is significant since, unlike other methods, BERT must learn its rules and features unsupervised. That allows BERT to find the best solution unrestricted by preconceived rules.

[1] https://github.com/huggingface/pytorch-pretrained-BERT

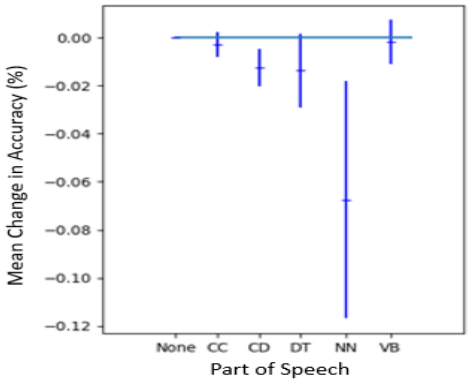

Figure 1: The mean results of the ablation study over 10 runs. The error bars are the standard deviation.

Source	Accuracy
Ott	89.8%
Vogler and Pearl	87%
Xu and Zhao	91.6%
Ren and Ji	85.7%
BERT	**93.6%**

Table 1: Comparison of accuracies on the Ott corpus.

BERT has achieved the first step for this project: being able to accurately classify deceptive text.

The ablation study (figure 1) revealed that the network is insensitive to most parts of speech being removed, with one strong exception. When the singular nouns (NN) were removed, the network accuracy dropped by 2% to 12%. This may indicate that singular nouns are a strong indicator of deception or truth; however given the prevalence of singular nouns in everyday language it is possible that removing them makes the review less comprehensible overall and harder to classify.

5 Future Work

The first step to unpacking what rules BERT is using is to build a saliency map of the inputs to the network. This will tell us which parts of the input are the most informative to the classification. We can modify the input, substituting phrases that are similar in meaning but different in language, which will allow us to see what can tip the classifier in one direction or the other. We plan to test BERT on other corpora to see if it can learn rules belonging to other text genres, as well as if the learned rules transfer from one corpora to another. Once the rules are determined, we can use them to train humans to better detect deception.

References

[1] 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.

[2] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.

[3] 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.

[4] J. Li, M. Ott, C. Cardie, and E. Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1566–1576, 2014.

[5] E. Loper and S. Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[6] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501, 2013.

[7] 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.

[8] Y. Ren and D. Ji. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385:213–224, 2017.

[9] N. Vogler and L. Pearl. Using linguistically-defined specific details to detect deception across domains. *Natural Language Engineering*, 1(1):1–32.

[10] W. Y. Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[11] Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. *Proceedings of COLING 2012: Posters*, pages 1341–1350, 2012.