The paper that I will be summarizing is *Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis*. The authors of this paper are Linyi Yang[1,2,3,4], Jiazheng Li[2], Padraig Cunningham[2], Yue Zhang[2,3], Barry Smyth[1,2], and Ruihai Dong[1,2] with the affiliations:

1) The Insight Centre for Data Analytics, University College Dublin
2) School of Computer Science, University College Dublin
3) School of Engineering, Westlake University
4) Institute of Advanced Technology, Westlake Institute for Advanced Study

The problem that is addressed in this paper is regarding how deep neural models trained on sentiment analysis tasks can learn intended behavior based on what the paper terms 'spurious patterns,' or artifacts. In other words, the model may correctly identify the sentiment of a document but determine said classification by excessively weighting the importance of a word or feature with little to no importance to the sentiment. An example that the paper gives is the input sentence: "Nolan's films always shock people, thanks to his superb directing skills." Say that a trained model correctly classifies this sentence as having a positive sentiment; we should expect that it gives more importance to the work 'superb,' but it is possible that the model would learn to associate the word 'Nolan' with positive sentiment, which would indicate the model did not actually learn the intended features. This problem can cause unintended biases (including social and racial biases) to be learned by models, which when run on out-of-domain data become extremely evident.

Prior research in the area involves the introduction of counterfactual data into a corpus to prevent models from learning these spurious patterns. Going back to the Christopher Nolan example, the dataset could be augmented with the sentence "Nolan's films always bore people, thanks to his poor directing skills" (another example from the paper). By including this in the training dataset, the model would learn not to associate the word 'Nolan' with semantic meaning, and focus on the correct features like 'poor' and 'superb.' It has been shown in prior research that counterfactually augmented data (CAD) results in improved performance on out-of-domain data and the learning of the real causal relationships between input sequences and output labels. However, collecting enough counterfactual data is time consuming because most methods involve human annotators who manually generate these counterfactuals.

In contrast, this paper explores the possibility of automatically generating counterfactuals to improve model performance. The details of the implementation are very thorough, and in summary the authors propose a three-stage pipeline: first, they identify the most important causal words using a slow method of decomposition, removing words one by one and assessing how each removed word affects the overall sentiment of the remaining sentence. They then remove these causally important words and substitute them for their opposite words (in terms of sentiment). This produces a set of plausible counterfactuals, and the most humanlike ones are selected using something they call MoverScore. This contribution is unique because, as stated before, previous methods required human annotators, a method that took a long time, making CADs rare and inaccessible for many tasks. The ability to automatically generate counterfactuals can make the CADs widely accessible to many classification tasks to eliminate the learning of spurious patterns.

The results measured by the authors support this conclusion. They measured the efficacy of counterfactuals by training different types of models on the SST-2 and IMDB benchmark datasets, then training the same models on SST-2 and IMDB with counterfactual augmentation from human annotators, and then finally training on SST-2 and IMDB with counterfactual augmentation generated automatically. Then the authors compared their results using the accuracy metric. They measured significant improvements not just from the original unaugmented data to the automatic CADs, but also from the manually created CADs and the automatic CADs. Taken directly from the paper, for automatic CADs vs manual CADs we have automatic CADs outperforming by these accuracy scores on different models: SVM (+1.1%), Bi-LSTM (+0.7%), BERT-base-uncased (+2.1%), BERT-Large (+0.8%), XLNet-Large (+1.0%), and RoBERTa-Large (+0.5%), which is all the more impressive because some of these (like the BERT models) are state-the-art models.

I believe that this work is very important because it helps future machine learning models of all kinds to remain unbiased and learn true causal relationships. It is very easy to misinterpret and cherry pick data and results to fit a certain narrative, and we certainly don’t want our AI and ML models, which are supposed to analyze data objectively, to exhibit racism or other forms of discrimination by learning these things from biased data.

Linyi Yang has 185 citations, Ruihai Dong has 960 citations, Jiazheng Li has 34 citations, Padraig Cunningham has 16209 citations, Yue Zhang has 11967 citations, and Barry Smyth has 24055 citations.