Shuhui Zhu, sz649

Critique 1: “Finding Deceptive Opinion Spam by Any Stretch of the Imagination”

Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock

While the work’s analytical methods, combining context-sensitive approach (n-gram-based classifiers), genre identification with psycholinguistic deception detection is creative and reasonable, the definition of the problem, data collection process and evaluation methods show limits. This leaves doubts in the accuracy of the experiment results, thus applying the conclusions to real life is immature.

A highlight of the automated approach analytics methods is to combine n-gram and LIWC software generated psycholinguistic features with machine learning models. This could be applied to text classification task in general, deceptive opinion spam in positive hotel reviews is one scenario, further in goods reviews, restaurant reviews, and social platform posts.

Given the value and possible appliance of the proposed methods, several limits should be addressed. First limit is in data collection process, the filtering rules to generate truthful review sample does not guarantee the truthfulness of the content. If by applying these rules, deceptive contents could be eliminated, I would argue there is no need to further explore the difference between deceptive and truthful texts. In defined truthful review sample, there may contain deceptive reviews. Reflecting upon what is deceptive opinion spam, the article defines deceptive opinion review created by people who write the review without experiencing them. What about the people who could obtain the facts of the hotel, (say a customer) but incentivized monetarily to write only positive reviews? They would know the details of the hotels, capable of writing informatively, but may distort their subjective opinion. A possible but troublesome solution to data collection is to visit the hotels and ask people checking out and interested in the outcome of the project to complete truthful reviews. Monetary incentivization could be given on a completion base. Filtering of positive reviews and further selection could be made in later stage.

Another limitation is in evaluation step, three volunteers made predictions directly on a subset of the balanced sample based on their prior exposure and understanding of deception without any knowledge of the sample data. In contrast, the automated process has access to the sample, training data, but no prior understanding of deception cumulated elsewhere. I would assume learning of deception rules based on training data would be much more effective. An analogy is one student studied a set of sample questions before a test, but the other did not, the one reviewed would be more likely to do better. In the article’s experiment, automated approaches are better advantaged by the accessibility to the additional information in training data. It is unfair to draw the conclusion that human judgement is worse than automated approach in this context.

Finally, in real life, fake review detection is a progression evolved against the progression in fake review writing. The learned rules, such as “truthful opinions tend to include more sensorial and concrete language” could inform not only deceptive detection but also fake review writers. A further step could be measuring the automated approach’s performance on fake samples generated by informed writers avoiding the learned rules.