Finding Deceptive Opinion Spam by Any Stretch of the Imagination
Myle Ott, Yejin Choi, Claire Cardie, Jeffrey T. Hancock
Lillyan Pan
Ldp54

As consumer-driven review sites increase in popularity and influence, so too does the use of deceptive opinion spam to promote certain businesses. The paper describes a sophisticated method of integrating and comparing three automated approaches to detect deceptive opinion spam– genre identification, psycholinguistic deception detection, and text categorization. Further, the automated processes were contrasted with human and meta-judge performances to test the validity of the automated classifiers. An alternative to detecting deceptive opinion without using gold-standard data would be to analyze the distortion of popularity rankings as done by Yoo and Gretzel (2009) by using a statistical test to compare the psychologically relevant linguistic differences between truthful and deceptive hotel reviews. A major disadvantage with Yoo and Gretzel's method is that is must be done manually, while the method described in the paper used a much larger dataset in order to develop and evaluate automated deception classifiers. By combining and contrasting several different classifiers and metrics, the paper is able to create a more robust and credible model as previous, major studies rely on comparisons to a random guess baseline of 50% and or exclusively to poorly calibrated (to detective deceptive opinion spam) human judgments.

A possible application of the final classifier to detect deceptive opinion spam would be to detect fake users in social media platforms. Many social media sites struggle with users that clutter sites with advertisements or harmful/inappropriate material– both of which detract legitimate users from using these websites. Genre identification, psycholinguistic deception detection, and text categorization would all still be valid approaches to detecting fake users. In genre identification, test can be run to see if a relationship exists between fake and legitimate users by contracting, for each post by the user, "features based on the frequencies of each POS tag" (Ott et al., 2011). For psycholinguistic deception detection, the Linguistic Inquiry and Word Count (LIWC) software can be used to detect personality traits that are associated with fake contrasted to legitimate user content. Deception analysis can also be used in terms of posts from fake users that contain harmful content. Text categorization can be used by modeling content and context with n-gram features as well. Just as UNIGRAMS, BIGRAMS$^+$, and TRIGRAMS$^+$ were used to model travel reviews, these n-gram feature sets can be used to model social media content. Further, the three approaches mentioned above can be used to train Naïve Bayes and Support Vector Machine classifiers similarly as in the original paper. Note, human and meta-judge performances can again be leveraged to test the validity of the resulting classifiers.

Expected results include a final classifier that outperforms human classification, which an observed human bias to classify users as legitimate over fake. The superior classifier would most likely be a combination of LIWC and n-gram features as detecting fake user content would require consideration of both a universal set of illegitimate language features and contextual parameters.