**Assignment 11**

**1 Compare the results from the dictionary-based approach to the supervised machine learning approaches. Which ones outperformed? Why do you think that is?**

The machine learning approaches outperformed. The correlation of dictionary-based approach (correlation between sentiment and the difference score = .25; correlation between sentiment and the Janis-Fadner coefficient of imbalance = .12) is more *near to zero* than that of machine learning approaches (Matthews correlation coefficients >= .59), which implying that the dictionary-based result is harder to distinguish itself from a *random* predict.

However, the overwhelming outperformance of machine learning may be partly attributed to the same source of training data and test data, giving that the dictionary is developed in a different context with different corpus.

**2 How did you do when you manually fine-tuned the rules-/dictionary-based and supervised machine learning models? How accurate were you able to get them? How did the grid search approach help you in fine-tuning the supervised machine learning models?**

To fine-tune the dictionary-based model, I added the words locating at top left corner to the positive list, words locating at bottom right corner negative list. I also deleted words like "threat", "threats", "uncertain", "uncertainty", "risk", "risks", "risky" from the negative list, because they are not necessarily negative comments in a film-reviewing context. However, it didn’t improve much.

To fine-tune the supervised machine learning model, I tried on Logistic Regression Model. The *elasticnet* penalty parameter always raises error. I tried to increased *C* and *max\_iter*. The results never outperformed that of model with default parameters.

By training new models for every parameter combination, Grid search indeed improved the results, but not much. The default parameter has a LR accuracy of 88.07%, while the Grid Search LR accuracy is 88.18%. On the other hand, it is that the defaulting parameters have fortunately been too good to be improved.

**3 In organizational research, we examine multiple different forms of reliability and validity. How does this compare to the way machine learning models are evaluated? What are the implications of these differences for organizational research?**

For reliability, traditional organizational research can implement multiple items, multiple raters and multiple rounds of survey. By the consistency analysis between outcomes of these settings, reliability issues can be checked. For machine learning, we can also apply different models, algorithms and parameter settings to examine the consistency between those cases.

For validity, the traditional research approaches can take advantage of theoretical implication (e.g. by more detailed moderation or mediation test) or external data/judgement. The machine learning, however, proves to be a black box. It has difficulty in interpretation. It largely depends on the data per se. The parameter setting process is also a kind of task characterized by agnosticism. Therefore, I totally agree with Shrestha et al. (2021) that machine learning is a good tool for *pattern discover* but can’t replace the traditional research paradigm. Researchers should still rely on their own intuition and understanding of theories.

**4 Were you able to find any topic model solutions that produced interpretable results? If so, what were they? If not, why do you think that is?**

I failed to find an interpretable model solution. I tried to reduce more top frequency words from the corpus. I also tried to decrease *eta* to 0.001 which can generate class batches with fewer words in each. But all these efforts seem not working.

The reason for this may be due to the nature of film reviews. Film reviews are often short. When one comments on a film, he or she is more likely to mention different aspects of a film (e.g. characters, scenes, shots, styles) to support his or her main opinion (i.e., the film is good or bad). The reviewer maybe less likely (or has no such need) to co-mention similar elements related to a film, for example, politicians, politics, policy and government. Such kind of reviews can appear verbose and repetitive. However, it is the latter case where the topic model algorithm can learn key information from, i.e., the similarity between words. This may explain the bad performance of topic model in this research.