Book Review

Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science

Stefan Riezler and Michael Hagmann
(Heidelberg University)

Morgan & Claypool Publishers (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 55), 2022, 147 pp; hardcover, 9781636392738; paperback, 9781636392714; ebook, 9781636392721; doi:10.2200/S01137ED1V01Y202110HLT055

Reviewed by
Richard Futrell
University of California, Irvine

When we come up with a new model in NLP and machine learning more generally, we usually look at some performance metric (one number), compare it against the same performance metric for a strong baseline model (one number), and if the new model gets a better number, we mark it in bold and declare it the winner. For anyone with a background in statistics or a field where conclusions must be drawn on the basis of noisy data, this procedure is frankly shocking. Suppose model A gets a BLEU score one point higher than model B: Is that difference reliable? If you used a slightly different dataset for training and evaluation, would that one point difference still hold? Would the difference even survive running the same models on the same datasets but with different random seeds? In fields such as psychology and biology, it is standard to answer such questions using standardized statistical procedures to make sure that differences of interest are larger than some quantification of measurement noise. Making a claim based on a bare difference of two numbers is unthinkable. Yet statistical procedures remain rare in the evaluation of NLP models, whose performance metrics are arguably just as noisy.

To these objections, NLP practitioners can respond that they have faithfully followed the hallowed train-(dev-)test split paradigm. As long as proper test set discipline has been followed, the theory goes, the evaluation is secure: By testing on held-out data, we can be sure that our models are performing well in a way that is independent of random accidents of the training data, and by testing on that data only once, we guard against making claims based on differences that would not replicate if we ran the models again. But does the train-test split paradigm really guard against all problems of validity and reliability?

Into this situation comes the book under review, *Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science*, by Stefan Riezler and Michael Hagmann. The authors argue that the train-test split paradigm does not in fact insulate NLP from problems relating to the validity and reliability of its models, their features, and their performance metrics. They present numerous case studies to prove their point, and advocate and teach standard statistical methods as the solution, with rich examples

https://doi.org/10.1162/coli_r_00467

of successful application to problems of NLP system evaluation and interpretation. The statistical methods are presented with motivation from first principles, with deep citations into the fundamental statistical literature and formal derivations presented at the right level for an average NLP or ML practitioner to follow. The methods and theory behind them are frequentist, and represent the current best practices within that paradigm, based on linear mixed-effects models (LMEMs) and generalized additive models (GAMs), and using generalized likelihood ratio tests for hypothesis testing. The authors also discuss fundamental philosophical ideas including the theory of measurement and construct validity, providing an overall deeper-than-average view into this literature in a succinct, comprehensible form.

The book takes the form of a kind of taxonomy of problems that can occur in NLP analysis, with each kind of problem paired with statistical solutions to diagnose and treat it, along with case studies involving Natural Language Inference (NLI), biomedical NLP, and Machine Translation (MT) evaluation, among others. Problems are divided into *Validity*, *Robustness*, and *Significance*.

Validity (discussed in Chapter 2) refers to the extent to which a model captures real generalizable predictive information based on input features. The authors divide feature validity problems into three kinds: (1) bias features, which are spurious features that are correlated with labels in the training data; (2) illegitimate features, where features contain unexpected and non-generalizable information, such as numerical patient IDs containing information about which hospital a patient was treated in, which in turn contains information about what conditions the patient likely has; and (3) circular features, which are features in the training data that deterministically define the label. Notably, these issues are *not* guarded against by the train-test split paradigm. Although these three kinds of invalidity may seem difficult to distinguish, they can in fact be cleanly separated by different statistical analyses, which the authors motivate after a discussion of measurement theory and construct validity.

Of these, the most interesting analyses are those to diagnose illegitimate features and circular features. Illegitimate features are cast as a form of failure of transformation invariance. Based on measurement theory, the relationships between features and outcomes should be subject to certain symmetries: For example, if patient IDs are to be used as features in a model to predict medical conditions, then the effect of the patient ID on outcomes should be invariant to bijections applied to the patient ID. Any failure of this invariance is an indication that illegitimate information is present in the patient IDs: For example, if some of the digits in the ID indicate which hospital the patient was treated at. Circular features, on the other hand, are diagnosed through a procedure involving fitting GAMs to predict labels given increasing amounts of data; in the limit of large data, any circular feature will end up with a large weight in the GAM while the rest of the feature weights go to zero. The authors call this the "nullification criterion," and give extensive examples of the application of these ideas using biomedical examples.

Reliability (discussed in Chapter 3) is the extent to which a result is robust to replication—for example, if a new set of annotators is tasked with labeling the same data, how much inter-annotator agreement will there be; or if a model is run again with a different random seed or a perturbation of hyperparameters, how likely is it to give the same performance. The authors critically discuss Krippendorf's α as a measure of intra-annotator reliability, and they discuss using bootstrapping to calculate confidence intervals. The real meat of this chapter is the discussion of model-based reliability tests based on LMEMs to perform Variance Component Analysis (focusing on the random effects). These methods are applied to study reliability of data annotation of MT output in terms of error rate. Reliability of predictions with respect to hyperparameters is also discussed extensively, with LMEM analysis using hyperparameters as predictors.

Significance (discussed in Chapter 4) is the extent to which a result provides evidence against some null hypothesis, usually that there is no difference between two models. The authors' main goal in this chapter is to advocate a general likelihood ratio test performed on LMEMs as the most general and effective significance test. The idea is to compare a model class M_1, consisting of models predicting some outcome given a set of input features, against another model class M_0, which predicts the same outcome from all the same input features except for some held-out features F, representing the null hypothesis that the features F have no effect. The goodness-of-fit of these models is compared using a likelihood ratio test based on the best-fitting models within each class, and the null hypothesis M_0 can be rejected if the likelihood ratio falls in a certain range. In addition to the generalized likelihood ratio test, the authors discuss nonparameteric significance tests such as permutation tests, and include a great deal of useful statistical background on topics such as the Central Limit Theorem and the basic logic of significance testing.

One content area that NLP practitioners may be hoping for in this book, but which they will not find, is Bayesian methods as opposed to frequentist methods. Within psychology, over the last decade, there has been an increasing interest in Bayesian approaches to statistical inference, fueled in part by perceived flaws in the frequentist paradigm, for example, (1) difficulties in interpreting frequentist statistics such as p-values and confidence intervals, which can be counter-intuitive; and (2) an over-reliance on significance thresholds for drawing conclusions which leads to "p-hacking," where a researcher performs data analyses in ways that are unscrupulously targeted toward achieving significance thresholds. Without taking a stance on whether or not these arguments have merit, I think it is worth pointing out that the current book does not much engage with the arguments against frequentism that are currently on the rise, seeing these issues as out of scope.

Overall, the book provides a useful introduction to the universe of statistical testing and measurement theory in a way that is immediately applicable for analysis of NLP models. Many of the methods advocated here can be applied straightforwardly, yielding immediate improvements in terms of our confidence in the validity and reliability of our models. More fundamentally, I believe the field of NLP will benefit from engagement with the underlying ideas about validity and reliability that have originated in fields such as psychology; the book provides an ample introduction to these ideas that cuts right to their core.

Richard Futrell is an Associate Professor of Language Science and Computer Science at the University of California, Irvine. His research focuses on language processing in humans and machines. His e-mail address is rfutrell@uci.edu.