Protokoll vom 29.05.2017:
Practical Significance Testing for Experiments in Natural Language Processing

F-Score Ungenauigkeiten sind der Grund für Significance Tests, da bei verschiedenen Daten die F-Score unterschiedlich sein kann. Das Testen und Vergleichen von verschiedenen Metriken, z.B. von Accuracy, gewichtete Accuracy und F-Score sind auch wichtig.
Die Frage ist, hat man verschiedene Ergebnisse, weil die neue Technik (System) besser ist oder ist das nur zufällig? Was wir haben wollen ist :
P(observed difference due to chance | test set results)
Wenn dieser Wert niedrig ist können wir annehmen, dass die Systeme unterschiedlich sind.
“Null Hypothesis Ho”: Ist die Annahme das kein Unterschied zwischen System A und B ist.