Supervision 2

Statistical testing

1

Call the two systems A and B.

k = the number of times A outperforms B + half the number of ties.

acc_A = the number of times A is correct / 100

acc_B = the number of times B is correct / 100

$\therefore k \approx 100(acc_A)(1\text{-}acc_B) + 50(acc_A)(acc_B)$

$= 50(acc_A)(2\text{-}acc_B)$

2

If the number of ties is very high, then it might indicate that there are few simple improvements to be made. Either both systems are correct, or the data has fooled them both, suggesting that improvements to be made may have to be complex.

If the number of ties is low, it suggests that the two systems combined might perform very well (i.e. the first machine is accurate on some of the data and the second machine is accurate on different parts of the data) and so the improved machine might take features from both.

Overtraining and cross-validation

1

$\mu = 82.2$

$\sigma^2 = 13.29$

2

$\mu = 83.4$

$\sigma^2 = 13.38$

This is probably not statistically significant because the distributions are very similar. There is no reason to believe that these samples could not have come from the same population

3

The Wayne Rooney effect can impact a sentiment analysis system whose training data is from a different time period than its testing data. This is when some words (such as people or movies) can have their sentiment implications change dramatically over time as public opinion about them changes.

Similarly new slang can cause a drift in the implications of some words (e.g. over time "sick" drifting from strictly negative to positive/ambiguous)

Uncertainty and human agreement

1

The difference between the human labels (which caused the high κ) isn't entirely caused by a lack of options. While in some cases humans all agree about how to interpret the document, but an accurate label is not available and so some pick one while some pick another, in many other cases the issue is that humans differ on how to interpret the document.

Humans draw from their own personal experiences when reading certain words, and this will influence each reader differently with respect to how they label the document.

Furthermore, too many categories might result in one document being quite well described by multiple. For example, if a text is slightly positive but not glowingly so, most humans will mark it as positive when the only options are positive and negative. Introducing neutral into the mix will split the vote so to speak, with some marking it as positive and some as neutral. This will cause κ to increase.

2

This allows you to explore the discrepancy between the author's intent and the reader's interpretation.

The analysis system could train on both sets of annotations and so provide a prediction along the lines of "This review was intended to be positive but it comes across as negative to readers"