9 Machine Learning and Real-world Data (sht25)

You work at a social media company, and your task is to detect cyberbullying messages based on the text they contain. You have access to a large number of messages, N, which have been manually labelled as "OK" and "bullying".

(*a*) How can you apply a Naive Bayes classifier to the task and evaluate it? Describe the approach, including how you would estimate the parameters. [3 marks]

(*b*) You decide to use precision and recall instead of accuracy as the evaluation metric for this task. Why does this decision make sense, and how are the metrics calculated? [1 mark]

(*c*) Your colleague wants to hire two more human annotators to re-label your training data. Why might this be a good idea, how would you measure agreement in this task, and do you think this would improve your classifier in any way? [3 marks]

(*d*) Due to repeated media coverage of cyberbullying, your company introduces a new policy stating that as many cyberbullying messages as possible are to be found and deleted, while still making sure the number of non-filtered messages remains high. Some additional manual labour is made available for this change. How does this affect your evaluation strategy, and how can you adapt the classifier to comply better with the new strategy? (Tip: a development corpus could be of use.) [4 marks]

(*e*) You realise that in this particular application, language change might cause the performance of your trained classifier to drop considerably over time. You have some manual labour available to address the problem, but not enough to relabel large amounts of text.

(*i*) Why is language change relevant here, and how might you notice it? [2 marks]

(*ii*) How could you efficiently build a "cyberbullying lexicon" containing words that have been found in recent cyberbully incidents? [3 marks]

(*iii*) How can you use lexicon-based information to make your classifier more robust towards language change? [4 marks]