COMPUTER SCIENCE TRIPOS Part IA – 2017 – Paper 3

7 Machine Learning and Real-world Data (AAC)

Suppose that a very large collection of documents describing the University of Cambridge has been collected. Your task is to build a classifier which assigns sentiment to each document, using the classes: *positive*, *negative* and *neutral*.

(*a*) There is no ground truth sentiment associated with the documents, but 200 have been manually classified by three human annotators. Fleiss' Kappa is 0.65 for this set. Explain what Kappa is and outline how it is calculated. You do not need to state the full formula for Kappa. [4 marks]

(*b*) Given the limited amount of annotated data, you decide to classify the documents using a standard sentiment lexicon. Explain how you would perform an experiment to do this. [5 marks]

(*c*) (*i*) How would you evaluate the results of the system you have described in your answer to part (*b*) using the annotated data? Give details of the evaluation metrics you would use. [4 marks]

(*ii*) If the primary objective were to identify the documents with negative sentiment, how might your proposed evaluation change? [2 marks]

(*d*) It is suggested to you that the classes automatically assigned by the sentiment lexicon approach could be used to provide training data for a Naive Bayes classifier. Could such a classifier perform better than the sentiment lexicon classifier which provided the decisions it was trained on? Explain your answer. [5 marks]