IST736 Text Mining

HW4

Multinomial Naïve Bayes

**~~Use Weka’s~~**

**Use PYTHON’s Multinomial Naïve Bayes algorithm to build models to classify the customer reviews by (1) sentiment (positive or negative) (2) authenticity (true or fake, lie detection). ~~(6 points)~~**

For each of the two classification tasks, use MNB to build the models, and evaluate them using 10-fold cross validation methods. (5-fold is fine too) ~~You can choose your own options for tokenization, and make sure to report if you change any default setting.~~

**Use CountVectorizer and Python. Create labeled data. Train the NB model and test it.**

**As part of your Results section:** For each model (lie detection and sentiment classification), report the 20 most indicative words that the models have learned.

**As part of your Results (the techy part) and Conclusions (the non-techy part) include discussion of:** Based on these words, do you think the models have learned the concepts (lie or sentiment) that they are expected to learn?

**As Part of Results**: Also, compare the difficulty level of sentiment classification vs. lie detection. Discuss whether you believe computers can detect fake reviews by the words.

**Submission: Always use the required Homework Assignment FORMAT.**

~~Submit a word document in research paper format, similar to previous homework.~~

**Grading rubrics: Same as always.**

~~Same as HW1~~.