IST707 Data Analytics

Lie detection

Scenario:

Some people claimed that machine learning algorithms can figure out whether a person is lying or not. Do you believe that? To test this claim, we have collected a collection of customer reviews, some are true some are fake, and you are going to test how well you can use the algorithms we've covered for fake review detection. Note that this data set also has sentiment label for each review, and you should try to predict that as well.

For both tasks, try Naïve Bayes and SVM (note: you can use the sentiment analysis package to generate sentiment analysis instead of the existing column, but don't use sentiment analysis to **predict** sentiment), but I'd like you to explore different strategies, and report your results. For example:

1. Different preprocessing methods – e.g., with or without stop-words, lemmatization, reducing the specific tokens you've used to maximize information gain.
2. With either category (lie / sentiment) does it help to include the other category as a feature? For the lie feature
3. Use a topic model & sentiment analysis module to generate additional features and use these in combination with / instead of raw tokens.

Try to get the best results you can.

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall

In a write-up following the table, please explain the rationale for the strategies you have chosen, including the theoretical foundation for your choice. Also explain your parameter tuning approach for every attempted strategy. Where did you start? How much of a difference did parameter tuning make? Why?

For each task and feature set (sentiment classification / lie detection), use the Rank module (Gini and Information Gain Ratio) to rank the features and list top 10 features from each method. Based on these top features, can you understand what patterns the classifiers have learned from the data?

Compare performance difference in sentiment classification and lie detection, and tell us which task is harder, and try to explain why.