HW 1 Report

Name: Karan Patel

Email: pkaran@pdx.edu

Phone: (503)860-2146

1. Provide a written description of your best model and a justification of the features that you used or did not use. Why do you think they are helpful or not? You must explicitly list the features you tried here.

My best model was the SVM model with 3 features in addition to the n-gram based features. Following were the features that were used by my best performing model:

1. Unigram and Bigram based features
2. Count of repeating punctuations in each headline. Example: !!, ???
3. Set of 8 features corresponding to the number of words found in the headline that fall under an "emotion" bucket (i.e. one of the following 8 classes: 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust').
4. Set of 8 features corresponding to the emotional intensity of words found in the headline that fall under an "emotion" bucket (i.e. one of the following 8 classes: 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust').

The model had a 10-fold cross validation accuracy of 79.49% and test accuracy of about 87%.

Adding some of the additional features was helpful for the binary classification task on hand. The features were helpful since they represented more detailed aspects about the headlines that the n-grams based features couldn’t represent on their own. Addition of features representing counts of repeating punctuations for instance was helpful since it helped represent over-exaggeration / hyperbole to some extent. The addition of emotional lexicon-based features was also helpful since they helped add emotional aspects of unigrams in the headlines which, per the evaluation metrics, proved to be a helpful addition to the sarcasm classification model.

Some of the features that I tried but didn't add to the model were as follows:

1. Features representing number of words per headline – Adding the feature did not improve the classification metrics for the model at all. I believe this is the case since the word counts between sarcastic and non-sarcastic headlines were not too different in most cases.
2. I added only unigrams, only bigrams, only trigrams as well as both bigrams and trigrams based features to the model using the "GridSearchCV" in scikit-learning library but the features negatively impacted the classification accuracy of the model. I believe this to be the case since the headlines are not too long and thus have limited word counts for which the combination of unigram and bigrams worked the best out of all other possible combinations.
3. I also tried adding cue words based feature but did not notice any improvement in the classification accuracy. I believe this is the case since many of the cue words were not enough to determine if a sentence was sarcastic in nature or not. They didn't provide the needed context.

2. Do an error analysis. For your best performing model, select 3 interesting examples that did not work. Indicate why you think they were classified incorrectly. If you could add additional features, what features do you think would help improve performance?

All the headlines below are sarcastic based on their truth labels in the data but were classified by the model as not being sarcastic. Following are the 3 interesting examples that did not work:

1. "`expansive obama state of the union speech to touch on patent law, entomology, the films of robert altman`"

I believe the first headline was not classified correctly since only the last phrase "`the films of robert altman`" of the headline is sarcastic in nature. The rest of the headline does not seem to be sarcastic in nature. To prevent such misclassification, I would perhaps add a category of words that are relevant to the workings of congress which likely won't include "films".

2. "`new instant lottery game features three ways to win, 19,839,947 ways to lose`"

Personally, the second headline does not seem to be too sarcastic. It seems to be representing the probabilities correctly. Also, the number "`19,839,947`" might have to do something with the misclassification given its magnitude. I believe adding features based on the magnitude of the numbers in headline might be a good feature to prevent such misclassifications.

3. “more cats made”

This headline is only comprised of 3 grams. Perhaps the headline is too short to be able to correctly classify given very few numbers of possible features that can be inferred. To prevent such misclassification, I would perhaps get more features from the content of the article that this headline belongs to.

3. Which evaluation metric (accuracy vs. micro-averaged vs. macro-averaged F-score) do you think is most suitable for this task and dataset? Why?

Micro-average aggregates all classes to come up with the metric and thus is perhaps more suited to a multi-class classification task which is not the task at hand for this homework. I believe the macro-averaged f-score would be the most suitable for binary classification since it computes the averages independently of each class.

4. Additional question for CS 510 students: How could the performance of the classifier be improved further? What role, if any, could the associated full news articles play in boosting the performance?

Feature selectors could be leveraged to improve the performance of the classifier. If only the highest scoring features were to be used by the classifier, then the performance could improve since it would lead to less noisy data thus improving the classification accuracy. In addition to this, using sarcasm-based lexicon could also help identify powerful sarcasm based features thus helping the classifier.

The associated news articles would provide a broad range of additional features to the classifier thus potentially improving its classification performance. The full articles could include more sarcastic tone if the headline is also sarcastic and thus more features pointing in that direction which could be extracted