Project 5 - Naive Bayes Classifier

Jefferson Roylance
Justin Fairbourn

Dataset

Over 1 million reddit comments, with the following columns and labeled as sarcastic or not.

- label
- comment
- author
- subreddit
- score (# of upvotes - # of downvotes)
- ups (# of upvotes)
- downs (# of downvotes)
- date
- created_utc
- parent_comment

Analysis

- Used multinomial bayes classifier
  - This was because the data didn’t seem to be normally distributed, and tests confirmed that multinomial got better results than gaussian
- 2 sections - textual analysis of the comment and analysis of the rest of the data

Textual analysis of the comment examined the following characteristics using 4-fold cross-validation:

- Amounts of each letter
- Length of the comment (in characters)
- Presence of punctuation (boolean)
- Average word length
- Words used - this was found out by taking the top 500 words used in all comments and then finding the counts of each of those words in the comment
- Checking for predefined patterns (we only got around to checking for the presence of '...')
- Number of uppercase letters

Analysis Pt. II

- Used multinomial bayes classifier
  - This was because the data didn’t seem to be normally distributed, and tests confirmed that multinomial got better results than gaussian

Our second Naive-Bayes Analysis fed the following into the classifier:

- The total amount of "upvotes" the comment received.
- The total amount of "downvotes" the comment received.
- 15 dummy boolean variables representing the comment's subreddit.

Not satisfied with the f-scores, we also ran another analysis and instead fed the classifier *only* the subreddit data. Surprisingly, this gave slightly better performance.

In both of these analyses, cross-validation was performed 20 times.

Results

Using just the comment textual analysis, we got an f-score of 0.6209

Average f-score with subreddit and score data (using comments in the top 15 subreddits): 0.4480299005407648

Average f-score with just subreddit data (no scores): 0.5123293945443769

Use cases include automatic classification of sarcastic comments by social media as well as messaging apps.