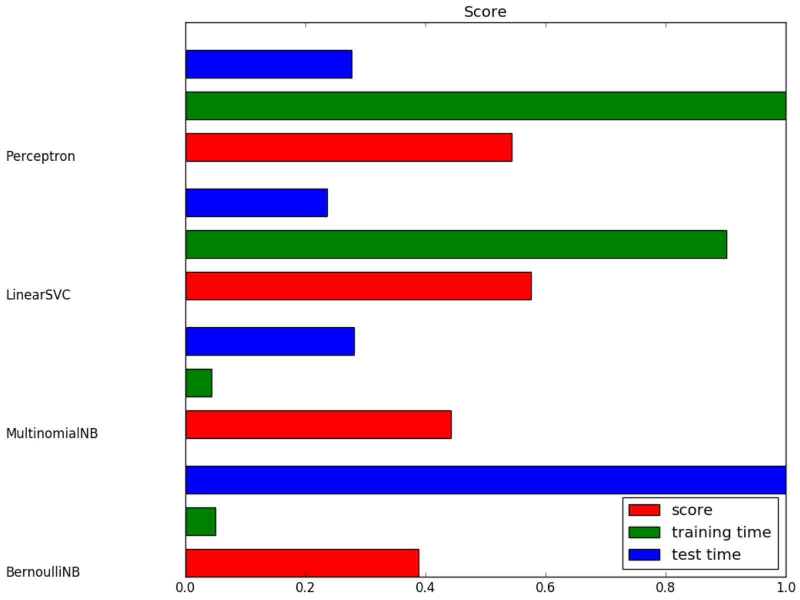
This paper attempts to explore the use of user reviews on the popular review site Yelp to classify the ratings made by those users for each review, based on term frequency for the written parts of the review. This exploration is a specific example of attempting to categorize emotional responses based on the expressed word choice, which could be of great use to human-computer interaction by guiding system responses to human users. Of particular interest might be in the use of software for diagnosing and treating mental illness and teaching empathic responses for people with developmental disorders. This particular dataset is comprised of # million yelp reviews for # businesses. Yelp reviews might provide insight in that they are very informal, however this might increase the dimensionality of the instances for analysis, as slang, non-words, typos, misspellings (intentional or otherwise) can be very prevalent.

In the interest of time, the scope of this experiment was narrowed to restaurants in English-speaking cities. As such, instances from the cities of Montreal, Canada (French) and Karlsruhe, Germany (German) were excluded. The introduction of additional languages would serve to massively increase dimensionality (depending on the prevalence and word-diversity of foreign language reviews, up to a factor of three). In addition, only instances which were categorized as “Restaurants” by Yelp were included. This is because in addition to reviews of restaurants , the dataset contained reviews for hotels, recreational services, retail outlets and others which are likely to each have a different lexicon of features associated with themselves and possibly reuse of the same features with different semantics. For example if a user says “smell”, smells are often both good and bad in a restaurant context, whereas a smell at a hotel or retail outlet may be more likely to be bad. This is the sort of comparative analysis that might be useful for further investigation, perhaps to see if there are universal terms that describe response across subjects. In addition, Yelp is most known for its reviews of restaurants and the richness of reviews on restaurants was exceed the sum of all excluded reviews in information.

Once the reviews were filtered to English-speaking restaurants, the monograms (singular words) and bigrams (consecutive two-word phrases) of the review texts were tokenized and each review text was vectorized in accordance with how often the tokens appeared in the text, relative to the overall number of tokens (Term Frequency). Then IDF (Inverse Document Frequency) was applied, that is each term was re-weighted according to how often it appeared in any document. This is to weight very frequent and very infrequent words less (as they provide less information) in the review. This process is know as TF-IDF (Term frequency-In Document Frequency). The data was then split uniformly into 2/3 training and 1/3 testing.

At this point, the training data was composed of  907484 samples with 9949748 features and even using a small subset of the instances (~1% = 100001 samples with 1960626 features) the computational overhead was already very large, taking minutes to load into memory and test. Classifier strength on this set was also very poor, likely because of the massive dimensionality causing over-fitting.



The full feature space and sample size was not resolvable for the scope of this experiment. The Chi-Squared function was attempted for feature selection to choose the best 1% of features. However, the massive feature space caused memory errors when attempting to convert it to a form usable by scikit.