> I'm a fan of TV movies in general and this was one of the good ones. The cast performances throughout were pretty solid and there were twists I didn't see coming before each commercial. To me it was kind of like Medium meets CSI.

Did anyone else think that in certain lights, the daughter looked like a young Nicole Kidman? Are they related in any way? I'd definitely watch it agin or rent it if it ever comes to video.

Dedee was great. Haven't seen in her in a lot of things and she did her job very convincingly.

If you're into to TV mystery movies, check this one out if you have a chance.

As seen above, one necessary pre-processing step prior to feature extraction was removal of HTML tags like "
". We used simple regular expressions matching to remove these HTML tags from the text. Another important step was to make the text case-insensitive as that would help us count the word occurrences across all reviews and prune unimportant words. We also removed all the punctuation marks like '!', '?', etc as they do not provide any substantial information and are used by different people with varying connotations. This was achieved using standard python libraries for text and string manipulation. We also removed stopwords[6] from the text for some of our feature extraction tasks, which is described in greater detail in later sections. One important point to note is that we did not use stemming of words as some information is lost while stemming a word to its root form.

Predictive Task:

The main aim of this project is to identify the underlying sentiment of a movie review on the basis of its textual information. In this project, we try to classify whether a person liked the movie or not based on the review they give for the movie. This is particularly useful in cases when the creator of a movie wants to measure its overall performance using reviews that critics and viewers are providing for the movie. The outcome of this project can also be used to create a recommender by providing recommendation of movies to viewers on the basis of their previous reviews. Another application of this project would be to find a group of viewers with similar movie tastes (likes or dislikes).

As a part of this project, we aim to study several feature extraction techniques used in text mining e.g. keyword spotting, lexical affinity and statistical methods, and understand their relevance to our problem. In addition to feature extraction, we also look into different classification techniques and explore how well they perform for different kinds of feature representations. We finally draw a conclusion regarding which combination of feature representations and classification techniques are most accurate for the current predictive task.

Literature:

The original work[3] on this dataset was done by researchers at Stanford University wherein they used unsupervised learning to cluster the words with close semantics and created word vectors. They ran various classification models on these word vectors to understand the polarity of the reviews. This approach is particularly useful in cases when the data has rich sentiment content and is prone to subjectivity in the semantic affinity of the words and their intended meanings.

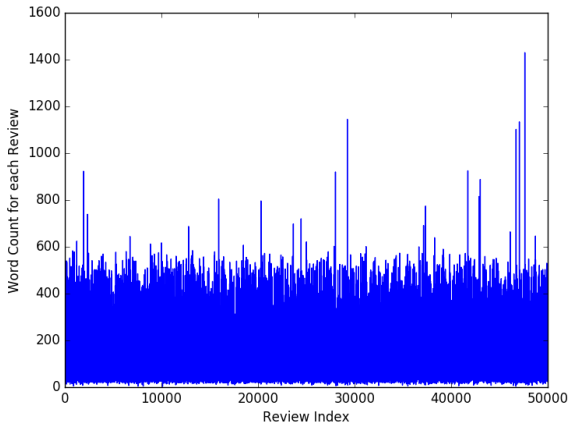
Apart from the above, a lot of work has been done by Bo Pang[7] and Peter Turnkey[8] towards polarity detection of movie reviews and product reviews. They have also worked on creating a multi-class classification of the review and predicting the reviewer rating of the movie/product.

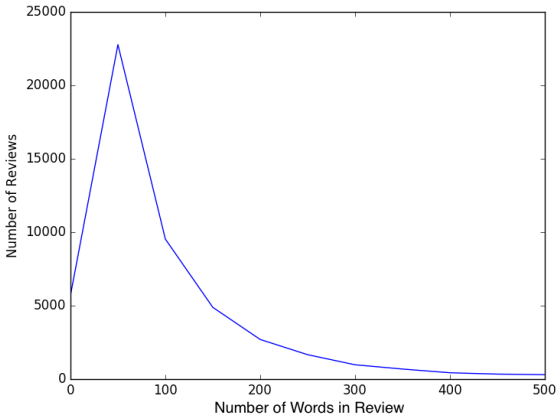
These works discussed the use of Random Forest classifier and SVMs for the classification of reviews and also on the use of various feature extraction techniques. One major point to be noted in these papers was exclusion of a neutral category in classification under the assumption that neutral texts lie close to the boundary of the binary classifiers and are disproportionately hard to classify.

There are many sentiment analysis tools and software existing today that are available for free or under commercial license. With the advent of microblogging, sentiment analysis is being widely used to analyze the general public sentiments and draw inferences out of these. One famous applications was use of Twitter to understand the political sentiment of the people in context of German Federal elections[9].

Exploratory Analysis:

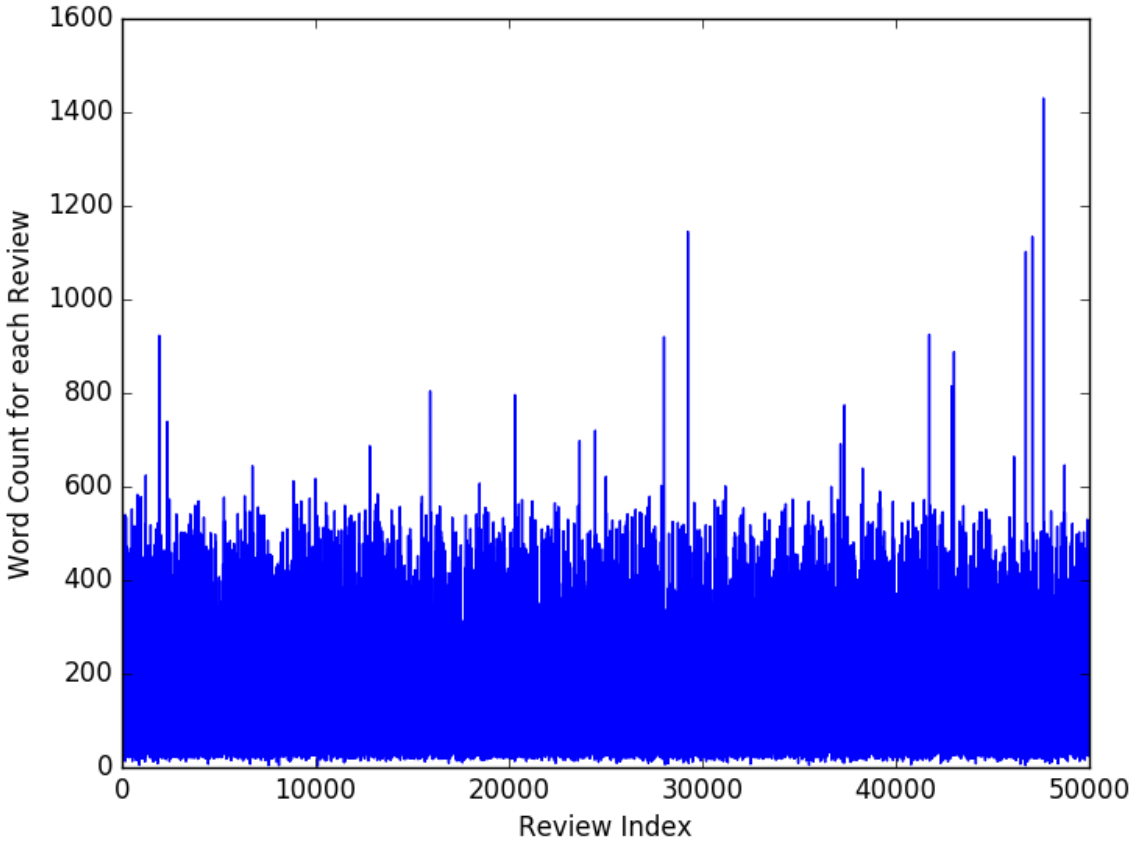
One of the starting points while working with review text is to calculate the average size of reviews to get some insight on quality of reviews. The average number of words per review is around 120. The graphs below clearly indicate the variation of the word count for each review. From this information we deduced that in general people tend to write pretty descriptive reviews for movies and as such this is a good topic for sentiment analysis. Also, people generally write reviews when they have strong opinions about a movie; they either loved it or hated it.





Apart from the word count per review another interesting metric was occurrence count of words across reviews. Some words have higher occurrence counts as compared to others depending on their relative importance. Below is the list of 20 most occurring words in negative and positive reviews along with a graph showing variability of word occurrences across all reviews. Also, the average word occurrence count was around 33 over all 50000 reviews. From all this information and the below graphs, it is clear that "Bag of Words" is not a very good model for doing sentiment analysis of reviews because similar words have high counts in both positive and negative reviews. Also, overall number of unique words is huge (1,63,353) across all the reviews and hence we use only top 50,000 and 1,00,000 of these during training. Also, this realization prompted us to move to other methods of feature extraction like n-gram modelling and TF-IDF counts of each words.

Negative Reviews		Positive Reviews	
Movie	Film	Film	Movie
Like	Even	Like	Good
Good	Bad	Great	Story
Would	Really	See	Time
Time	See	Well	Also
Don't	Get	Really	Would
Much	Story	Even	Much
People	Could	First	Films
Make	Made	Love	People
Movies	First	Best	Get



Feature Extraction:

We used 3 methods for extraction of meaningful features from the review text which could be used for training purposes. These features were then used for training several classifiers.

- Bag of Words: This is a typical way for word representation in any text mining process. We first calculated the total word counts for each word across all the reviews and then