Summary

Linguistic features for review helpfulness prediction

Submitted by:
Pallavit Aggarwal
Intelligent Systems – 22333721

In this paper, the author builds on the existing Linguistic Category Model in which the text under review is broken down into three categories – Adjectives, State Verbs and Action Verbs. A new method to extract features is proposed which is then analyzed by a binary classification model.

More inspiration is derived from multiple other works which are stated in the Related works section. As some existing work aligns the thought process of the author, he introduces the terminologies that would be used in the novel approach that will be discussed. He uses P for denoting product characteristics like name, type, description etc, C for the textual content, M for the review metadata, H for the helpfulness score computed as the ratio of helpful votes to total votes. A feature matrix F, and an n-dimensional matrix Y is described which stores 0,1 values based on a threshold on the helpfulness score.

For the feature matrix, the features are extracted by first using the NLTK-POS tagger to tag each word in the review's corpus. Now the individual Adjectives, State Verbs and Action Verbs are extracted using the tags. Action verbs are further fragmented into three different categories and to identify these separately either a list of words with each category can be defined or the other automatic approach which the author uses is that the valence of these words can be computed and used to decide which categories they fall into. SentiWordNet is used, to compute the valence, and two values tau-1 and tau-2 act as thresholds to properly judge the cumulative score of a sentence based on the average weighted values of these valences and decide the overall category that the sentence should fall into. After this an algorithm computes the average scores based on the features, by iterating over the feature matrix F for every review. The mean, standard deviation, and the z-score which is the normalized scores are now calculated. Other metrics like Review Extremity, review age, readability features and subjectivity features are also mentioned.

Two datasets are used, one of them being the comments scraped from amazon.com. They're cleaned and pre-processed by applying some minimum criteria that it should match in order to be selected as helpful and not spurious in nature. Basic threshold for helpfulness score for the above Y matrix and the two tau values are established as 0.6, 0.6, 0.6 respectively. Naïve Bayes, SVM, Random forest are used in 10-cross fold validation as models to get predictions on review helpfulness and f-measure and accuracy are used to evaluate the performance.

The analysis is done first on performance of the models on both datasets using a combination of features and the accuracy achieved is greater than 85% using RandomForest on DS1 and above 75% in DS2. When all the features are combined then the accuracy improves even more. One of the tables (Table 4) even shows that by using LF features the performance of the model is better than using any other feature and this validates the novelty achieved by the author in forming the LF features using the SentiWordNet and the algorithm. Effects of using review type, product type and changing the thresholds to improve the models are also discussed.

Finally the author concludes by stating that the techniques discussed in the paper highlight the importance of linguistic features in predicting the helpfulness of online reviews and provide a framework for identifying the types of features that are most useful for this task. By analyzing these features and using machine learning algorithms to predict review helpfulness, businesses can identify which reviews are most informative to customers and use this information to improve their products and services. The study has several implications for both researchers and practitioners. For researchers, the study suggests that linguistic features should be taken into account when predicting review helpfulness. For practitioners, the study suggests that businesses should encourage customers to write longer reviews that use certain words and phrases, as these reviews are more likely to be helpful to other customers

From a text analysis point of view, this paper highlights the importance of linguistic features, compares the performance of several machines learning algorithms, benchmarks using real-world data, and provides practical implications of this research for businesses.

References

1. Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. Expert Systems with Applications, 42(7), 3751–3759