# Prediction of Customer Ratings on a New Corpus for Opinion Mining

M. Saleh, A. Montejo-Ráez, M. T. Martín-Valdivia, L. A. Ureña-López

Universidad de Jaén, Department of Computer Science, Spain

**Abstract.** In this paper a new corpus for opinion mining is introduced. It has been generated from Amazon customer reviews on several products. Details about its generation along with a complete description of the corpus are given. Besides, a linear regression has been applied in order to study how sort comments behave as textual information for the prediction of customer rates. Our experiments show that these texts are quite informative and that the rate is an interesting measurement on the overall opinion of the customer on the product. This technique could help to summarize opinions in other web sites where rate is not explicitly given by the user.

## 1 Introduction

The number of blogs in the World Wide Web has been increased over several years. In these Weblogs people can estimate a publication, such as music, movies, video games, books, or electronic products. In addition, the author may assign rating to indicate its relative merit. Different types of reviews can be found on the net: On the one hand, "consumer reviews" are written by the owner of a product or the user of a service who has experience to comment whether or not the product or service deliver on its promises. On the other hand, some reviews can be written by an expert in that field who tested several products and can identify which offers are the best according to their features and their cost. This type of reviews refers to "Expert Reviews". Opinions in these Weblogs identify the author's viewpoint about the subject rather than simply recognize the subject itself. The opinion mining in such as Weblogs gives another magnitude to search and summarization tools. The year 2001 marked the beginning of widespread of the research problems and opportunities that sentiment analysis and opinion mining raise. Both of them denote the same field of study, which itself can be considered a sub-area of subjectivity analysis [1]. Sentiment Analysis (SA) is a discipline that deals with the quantitative and qualitative analysis of text for determining opinion properties [2]. The term sentiment analysis stands for a broad area of natural language processing, computational linguistics and text mining. It aims to extract attributes and components of the object that have been commented on a document [3]. With rapid expansion of the Web and online merchants, more people buy products on the Web. In order to enhance customer satisfaction, it becomes common for customers to submit and express

Troyano, Cruz, Díaz (Eds.): WOMSA'09, pp. 128-136, 2009

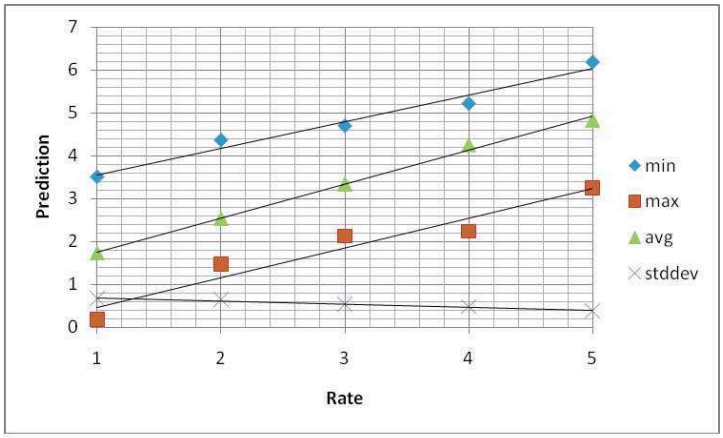


**Fig. 3.** Statistical analysis of predictions

environments (i.e. blogs) where comments are not rated. In this way, we could summarize a whole list of comments and study, using descriptive analysis, the average, standard deviation and other measurements from the distribution of predicted rates. As further work we plan to replicate our experiments using 10-fold cross validation to explore the effects of a learned model on non seen data and to provide statistical significant measurements on that setup. Also, other models, different to linear regression, have to be studied in order to identify possible better algorithms for rate prediction. In our opinion, deeper linguistic analysis has to be performed, as word vectors may not be best candidates as features before model learning. Actually, we are working on product features detection, so terms like optical lens or battery life would be considered as relevant attributes.

## Acknowledgments

This work is partially funded by project RFC/PP208/UJA-08-16-14 granted by University of Jaén, project FFIXP06-TU2301-2007/000024, granted by Andalusian Government, and project TIMOM TIN2006-15265-C06-03, granted by Spanish Government. Also another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC- AECID.

## References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1-2) (2008) 1–135