P1 – Project Statement

Project Title:

Opinion Mining for Online Product Reviews

Project Team Members:

Hardik Dalal – B00696939
Utsav Patel – B00691151
Chahna Dixit – B00695383

Problem Statement:

With the growth and advancement of Internet and E-commerce, more and more people buy products online. Along with buying the product, the buyers also give their valuable reviews on the product, its quality, and on various features of the product. These product reviews help other buyers as well as the manufacturers to get information about the product. Since large number of reviews are available for even a single product, it becomes difficult for anyone to go through all the reviews at one go; hence a summary of the reviews can come handy. This can be achieved by mining of customer's opinions (aka opinion mining). Opinion mining, also known as sentiment analysis is a field of study that analyzes the polarity of opinions and emotions from the user written reviews. In this project, we try to work upon opinion mining for such online product reviews.

List of possible approaches:

In order to carry out opinion mining, review text needs to be pre-processed, before it can be used for mining. Pre-processing of the reviews is basically removing noise from the text. Often, reviews are written in semi- or un-structured English, which can produce misleading results. The various pre-processing techniques are stop words removal, tokenization, lemmatization and stemming. S. Aravindan et al. [1] have used Stanford CoreNLP tool to perform sentence splitting, tokenization, lemmatization and part-of-speech (POS) tagging. The next step is classification of text based on its polarity. The two different methods for classification are Lexicon-based methods and Machine learning algorithms. In lexicon-based method, a lexicon based classifier is built which relies on the linguistic knowledge of the written text [2]. Machine Learning algorithms could be Support Vector Machine [1], k-NN [3], Naïve Bayes and many others. Our approach would be to implement lexicon-based method along with any of the machine learning algorithms to produce desirable results.

One other way of opinion mining is feature-based or aspect-based sentiment classification. Bross J. and Ehrig H. [4] have generated a context-aware sentiment lexicon for aspect-based review mining. Zhongwu Zhai et al. [5] modelled a semi-supervised learning method for feature classification, which outperformed the other existing methods. We would further like to study the aspect-based classification done by various authors and compare those different methods and their results.

Project Plan:

For the purpose of the project, we will be using Amazon review dataset of Cell phones and Accessories [6, 7] with number of reviews close to 78,000. The reviews are structured as follows: product/product ID, product/title, product/price, review/user ID, review/profile name, review/helpfulness, review/score, review/time, review/summary and review/text. The important fields are score, summary and text. The score field ranges from 0 to 5 and the summary field gives a short summary about the product. Taking the score and summary as main elements, we will find the polarity of the review – whether positive or negative.

Furthermore, considering the review text, we tend to preprocess the full text and then using the above mentioned approach, we will classify the reviews.

The pre-processing steps will involve removing the stop words, tokenization, lemmatization, POS tagging. After preprocessing, we will train two-third of the data and test remaining one-third of the data. We might consider using n-fold cross validation as well to achieve better results.

Since the dataset we have chosen does not promise to explore aspect-based opinion mining on the product reviews, we would make a survey on the aspect-based approaches used by various authors in their research papers.

References:

[1] S. Aravindan, A. Ekbal, "Feature Extraction and Opinion Mining in Online Product Reviews", *IEEE International Conference on Information Technology*, pp. 94-99, India 2014.

[2] Avanco L.V., Nunes M.G.V., "Lexicon-based Sentiment Analysis for Reviews of Products in Brazilian Portuguese", *Brazilian Conference on Intelligent Systems*, pp. 277-281, Sao Paulo, 18-22 Oct. 2014.

[3] Srivastava A., Singh M.P., Kumar P., "Supervised Semantic Analysis of Product Reviews Using Weighted k-NN Classifier", *11th International Conference on Information Technology: New Generations*, pp. 502-507, Las Vegas, NV, 7-9 April 2014.

[4] Bross J. ,Ehrig H., "Generating a Context-Aware Sentiment Lexicon for Aspect-Based Product Review Mining", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 1, pp. 435-439, Toronto, ON, Aug. 31 2010-Sept. 3 2010.

[5] Zhongwu Zhai, Bing Liu, Hua Xu, Peifa Jia, "Clustering product features for opinion mining", *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 347-354, NY, USA, 2011.

[6] http://snap.stanford.edu/data/web-Amazon-links.html

[7] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text", RecSys, 2013, *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165-172, NY, USA, 2013.