Data Mining and Sentiment Analysis

All the customer reviews data used were gathered from the webpage http://www.airlinequality.com/ which is owned by SKYTRAX [1], a corporate research advisors group for the air transport industry. SKYTRAX's webpage has costumer reviews about almost all the airline companies worldwide.
When collecting data we did not have access to SKYTRAX's database of costumer reviews.

Firstly we developed a Java package to go on each airline's passenger review webpages and get all the reviews and ratings raw data. Our program package extracted the personal information about the reviewer, the text opinion about the flight, and also all the ratings and information inputs that the reviewer submitted. All the raw data was classified into categories and it was stored in an Excel spreadsheet, which became known as the Master Data file.

Secondly we decided to perform sentiment analysis on each review's each sentence. Sentiment analysis is the process of analyzing people's opinions, sentiments, attitudes and emotions towards products, services, organizations, individuals, issues, events, topics, and their attributes [2]. We needed a tool that can separate text into sentences accurately, therefore we decided to use OpenNLP's [3] Sentence Detector, which can detect whether a punctuation character marks the end of a sentence or not. For the sentiment analysis we decided to use Timothy Jurka's sentiment [4], which is an R statistical package. Jurka's tool for sentiment analysis classifies text into six emotions (anger, disgust, fear, joy, sadness, and surprise) and three polarities (positive, negative and neutral). We wrote an R algorithm where Jurka's sentiment analysis functions got applied to each review's each sentence. This algorithm outputted the first sentiment analysis results.

Thirdly, for further sentiment analysis we decided to use Narayanan, Arora and Bhatia's [5] text classifier, which can perform sentiment analysis since its enhanced Naïve Bayes classifier model [5]. Narayanan, Arora and Bhatia stated that information could be better captured by looking at consecutive pairs of words like bigrams (two consecutive words) and even trigrams (three consecutive words). We decided to check for the most common bigrams in order to improve the accuracy of the sentiment analysis. For getting a list of the most frequently used bigrams we used the services of Semantria [6], which is a professional commercial text and sentiment analyzer.

Narayanan [7] created a sentiment analysis API (Application Program Interface) based on Narayanan, Arora and Bhatia's [5] work.
Fourthly we wrote an algorithm to interact with Narayanan's API and perform sentiment analysis on the opinion text reviews. The results obtained from Narayanan's sentiment analyzer became the second sentiment analysis successfully completed.

Fifthly, we selected 1000 sentences from the passenger opinion reviews and with clearly established rules we checked each sentence's polarity (positive, negative and neutral) by how a human mind would interpret it. This testing process helped to establish the main inaccuracies of the sentiment analyzer and it gave a chance to improve the accuracy of the analysis by making some changes. The change that we agreed upon is that all 'a' and 'the' words should be taken out of the opinion texts, since the algorithm is only checking for bigrams and frequently the above mentioned words stand in the middle of a 3 consecutive word negation (e.g. "not a good flight" would be transformed into "not good flight"). Removing the inconvenient words transformed many trigrams into bigrams, which made it possible for the sentiment analyzer algorithm to detect more negations, therefore improving the accuracy of the analysis. This was the final step for getting an as accurate as possible sentiment analysis for each sentence in each passenger opinion review text. For technical details check the appendix.

References:

[1]. Skytrax - Corporate Background. Skytrax research. http://www.skytraxresearch.com/main/about_skytrax.html. Last accessed on September 8th 2015

[2]. Liu B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, Ch. 1 Sentiment Analysis: A Fascinating Problem. Page 7.Web

[3]. Northedge R. (2006). Statistical parsing of English sentences. CodeProject, 2015. http://www.codeproject.com/Articles/12109/Statistical-parsing-of-English-sentences . Last Accessed on September 3rd 2015

[4]. Jurka T. (2012). Sentiment R package source code. GitHub, 2015. https://github.com/timjurka/sentiment . Last Accessed on September 1st 2015

[5]. Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naïve Bayes model. Doi:10.1007/978-3-642-41278-3_24. arXiv, http://arxiv.org/abs/1305.6143. Last Accessed on September 5th2015

[6]. Semantria Text Analytics Free Trial Sign Up. Semantria by Lexaltics. https://semantria.com/signup, Last Accessed on September 5th 2015

[7]. Narayanan, V. (2013). Sentiment Analysis API. http://sentiment.vivekn.com/ . Last Accessed on September 5th 2015