Sentiment Analysis

Nasir Arafat

MSc in Big Data

UNIVERSITY of STIRLING

Abstract

The aim of this project is to extract extracting expressed sentiments and other informative information by applying a range of technologies and comparing the effectiveness and then performing data visualisation on the information that would create an understanding of user-generated content and then develop a prototype that could bring sentiment analysis and data visualisation into one coherent application for the benefit of user swishing to analyse sentiment data. Our aim is to contribute to the research area of text classification, by way of the sentiment analysis and visualisation of user generated content.

Sentiment analysis and opinion mining is the field of study that analyses peoples opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks.

Introduction

- User-generated content on web, such as on E-commerce sites and social networking websites like twitter and Facebook has increased at very fast rate even beyond the prediction. The proliferation of new and appealing methods to share opinions, whilst bringing a constant stream of information, has created a problem for companies as well as individuals wishing to derive useful information from such content. This information informs such parties of the attitudes and experiences towards products or services that appeal to them, and creates an understanding of data that can be highly beneficial in a business context. The problem is partly due to the creation of more relevant, yet informal content, such as slang and use of different languages. Clearly this is an issue if we wish to extract information that is meaningful, and therefore *useful*.

Methodology

- First task was to get some useful data. We have collected data from freely available online datasets which contains reviews for movies, books etc. and data from twitter that could be useful for sentiment analysis and text classification. The data sets that we have collected was pre-processed using python library NLTK and then labelled so we could train our machine learning classifiers and test their accuracy. We were also able to obtain a labelled movie review dataset that came pre-loaded with python library NLTK.
- After pre-processing of data we are using Scikit-learn to implement machine learning techniques SentiWordNet, would provide us with subjectivity and objectivity scores of words grouped by parts-of-speech, useful for rule based method. After extraction of information our step will be to visualise the data and for that we will be using Python plotting library matplotlib.
- The software we are using for this project are NLTK (Natural Language Tool Kit), Scikit Learn (machine learning tool), SentiWordNet, MatPlotLib (for data visualisation)

Further work

- The further work in this project would be the development of the web based application that could dynamically update the current state using social media feeds, survey responses and press coverage to identify or predict emerging issues.