Sentiment Temporal Ranking

Problem Statement:

The Trip Advisor data set consists of comments/reviews by tourists on different Las Vegas hotels. Each comment can be categorized into Positive,Negative or Normal sentiment category, based on the textual content and ranking to the hotel provided by the user. Each comment in the data set is timestamped. Instead of just using text based analysis for studying the sentiment of the comments, the project aims to use the flow of sentiments i.e. time series analysis of sentiment rating, to better train the sentiment labeling classifier. Taking intuition from the Kinetic Model of Time Series, the rank i.e. overall average sentiment about the hotel at the current time constitutes a lot to the sentiment of comments in near future. Since the ratings for a hotel may change over time, the statistical parameters like mean and standard deviation of the hotel ranking will be deviating leading to autocorrelation, therefore temporal ranking of the hotel can't be modeled by Stationary time series models. The study aims to develop a machine learning based approach for time weighted ranking scheme to better label a comments sentiments given time stamped historical data.

Background :

The availability of large amount of information in the form of customer feedback, comments on various forums, on line discussions, blogs etc have acted as a profound source for customer or public opinion mining and sentiment analysis about a product or a business. Pang,Lee and Vaithyanathan [1], used Machine Learning based methods like Naive Bayes, maximum entropy classification, and support vector machines for classifying the sentiment of the whole document. Dave et. al. [2], worked on classifier that draws on information retrieval techniques for feature extraction and scoring to label sentiment to whole review. Though these methods can be used to classify the whole comment's sentiment but they can't be used for performing sentiment flow analysis and studying the current rank or overall sentiment.

Motivation :

Mei et al. [3] introduce a probabilistic topic model into sentiment analysis, and their model is able to track opinions and even sentiment dynamics of a topic within many documents. [4] discusses a new framework called Topic Sentiment Change Analysis (TSCA) . [4] makes the presumption that when a causal event occurs that changes the sentiment flow, this event is followed by various comments within the same time period, they introduces the notion of “topic hotness” and tries to catch the causal event via change in comment distribution. However, this may not be the case when we are looking at user feed back comments e.g. in trip advisor data set, comment frequency will not be influenced by change in some topic sentiment.

Lei et al[5] proposed a time series analysis based kinetic model temporal ranking scheme for searching in web. The model used a weighted ranking scheme where the temporal rank of a web page is the weighted sum of historical rank and current rank of the page. Further [5] models the ranking as a Kinetic Model to derive the weight matrix.

Using a similar hypothesis, we can state that the rank (or sentiment) about a hotel carries some momentum either positive or negative depending on its historical reviews. So we can form an ensemble learning process that can better label comments depending on the textual content as well as historical rank of the hotel.

The project can be broken down into following phases :

1. **Checking accuracy of conventional classification techniques** : This phase deals with sentiment labeling using

    a. Naive Bayes Algorithm
    b. Support Vector Machine
    c. Logistic Regression

2. **Time line formation** : User comments in the data set are date stamped. In this phase we will try to do time period partitioning i.e. dividing the labeled sentences corresponding to each topic into different time periods. This is necessary for the study of sentiment change analysis. The following methods can be used for time period partitioning
    a. Equal length time periods
    b. Each time period containing same number of comments.

3. **Sentiment Time Series Analysis :** In this phase we will try to use a time window based approach to train a Neural network. The neural network will be fed with label of 3 contiguous sentiments (-1 for negative, 0 for normal and 1 for positive sentiment) and the output layer neurons will try to predict the class of sentiment of fourth time period. The neural network will be trained using gradient descent and back propagation algorithm. So to train the network we will feed the output neuron with correct value of fourth time period neuron. In next iteration the neural network will be fed with sentiments of second, third and fourth time period and will try to predict the sentiment of fifth time period. In this way we will form a time series with window size 3.

4. **Sentiment Temporal Ranking :** The inherent problems with previous techniques is the dependence on the size of time window (in our case size 3) to model the time series. The last technique just try to predict the sentiment of next comment without considering its textual content. In this phase we will try to use some time decay function to get over all ranking of the hotel based on the historical sentiments about the hotel. When the text of new comment is fed to the model, we will try to use historical sentiment ranking as well as current comment's sentiment to label the current comment.

    Temporal_Sentiment = $(1 - \beta)$ Historical_Sentiment + β Reviews_Sentiment
    β = weight
    Historical_Sentiment (at time T=t) = $\sum \gamma t$ TimePeriodSentiment(t)
    γt = weight to time period t's sentiment (i.e. decay factor)
    TimePeriodSentiment(t) = sentiment at time period t (Sentiments are rated as -1,0 and 1)

In this project we will be studying two mechanism to model the temporal sentiment labeling scheme, to discover the decaying function and weight matrix

1. **Recurrent Neural Network** : In this approach we will be adding a memory layer to the neural network, that will store the historical rank of the hotel. The proposed network will have an additional weight matrix that will be used to as an ensemble criteria to assign weight to historical rank of the hotel and weight to current reviews sentiment to label the input text.

2. **Conditional Random Fields (CFR)**

References :

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. CoRR cs.CL/0205070 (2002)
2. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: WWW, pp. 519–528 (2003)
3. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: WWW, pp. 171–180 (2007)
4. Topic Sentiment Change Analysis , Yu Jiang1 , Weiyi Meng1 , and Clement Yu2 ,MLDM'11 Proceedings of the 7th international conference on Machine learning and data mining in pattern recognition,Pages 443-457,Springer-Verlag Berlin, Heidelberg ©2011
5. Link Analysis using Time Series of Web Graphs, Lei Yang, Lei Qi, Yan-Ping Zhao, Bin Gao, Tie-Yan Liu, CIKM’07 November 6-8, 2007, Lisboa, Portugal, ACM 978-1-59593-803-9/07/0011,