Predicting Polarity of Social Media Data using Wilcoxon and T- Test as Feature Optimization technique for Sentiment Analysis

Sentiment analysis is an information retrieval technique that delivers the vision of relevant users like customers regarding entities like products or services, individuals like sellers, service providers, and functional problems incorporated in events, and attributes. In the era of social web that includes social networks, forums and blogs, sentiment analysis is critical in decision making activities by an individual or an organization. With the phenomenal growth in the data quantity of social web, manually analyzing the opinion is almost impractical. Hence the machine-based sentiment analysis or opinion mining is desired. In this context the automated sentiment analysis become critical research objective that grabbed researcher's attention over a decade. Further, with respect to this, the contribution aims to design Machine Learning based learning approaches for explorative Twitter trends sentiment analysis.

Many of the contemporary Sentiment Analysis methods envisioned the intricacies because of maximum feature volume. This gap has been addressed by Feature Selection and optimization using statistical assessment schemes for choosing optimum features. In the initial phase feature selection and optimization is done using Wilcoxon Signed Rank Test which selects optimal features with linear process overhead. The work identifies a unique feature pair called term- lexicon and its feature association frequency (FAF).

The existing techniques for SA using n-grams do not consider other features like the presence of abbreviations, use of emoticons, slang and sarcasm. This gap is addressed by estimating the sentiment polarity as positive or negative through the multi objective features such as terms, emoticons, slang, and sarcasm under classifier adaptation. In order to select the optimal features, the proposal initially assesses the feature associability with sentiment lexicons and further from these optimal features are selected using T- Test. The Adaboost classifiers objective function is redefined with a novel objective function to incorporate multi objective features such as emoticons, slang and the sarcastic features.

The opinion related to trending issues such as politics, share market, sports etc. exhibit vivid dimensions of features. Feature optimization and labeling the opinion using a single classifier causes considerable false alarming. To address the false alarming rate issue ensemble classification strategies are effective. The proposed model determines the sentiment polarity under diverse dimension. AOEC defines the ensemble classification technique to address the challenge of diversity of features in voluminous training data. Experimental results show promising improvements with respect to Accuracy, Precision, Recall, F- Measure and Mathew Correlation coefficient.

S. Fouzia Sayeedunnisa