Sentiment Analysis of Streaming Data

Cloud Computing - Project Report

Aditya Sapate[1] and Chaitanya Munukutla[2]

[1] IIT Madras CS10B031
[2] IIT Madras CS10B040

1 Motivation

Social Media has creeped into everybody's lives and pockets. Constant feeds about the things that we care about, make up today's **news streams**. But, the user prespective about a specific topic is still not taken under serious consideration.

For example, a review of a recently released movie may be trending. But, it's not known whether it's trending on the positive or the negative light of user opinions. The current project aims at realtime sentiment analysis of social feeds, which serves the above purpose. THe case study uner consideration is the *Twitter Stream.*

2 Sentiment Analysis

2.1 Peter D. Turney, *Semantic Orientation Applied to Unsupervised Classification of Reviews*, 2002

The above paper presented a simple unsupervised learning algorith for classifying reviews as *recommended*(thumbs up) or *not recommended*(thumbs down)

Extracting Phrases with Adjectives or Adverbs Replying on past work by Hatzivassiloglou and Wiebe (2000) which showed that adjectives and adverbs are good indicators of subjective, evaluative sentences, the algorithm first extracts all the phrases of the review which contain an adverb or an adjective.

However, procuring an isolated adjective may lead to different opinions in different contexts. For example, "unpredictable plot" may lead to a positive opinon under the context of a movie review, but "unpredictable steering" leads to a negative opinion under the context of an automotve review. So, the lone word "unpredictable" should not be relied on, and hence the current algorithm extracts two consecutive words, of which, the second word most usually depicts the situation or context of the review.

Estimating the Semantic Orientation The estimation of semantic orientation of the extracted phrase uses the PMI-IR algorithm (Church and Hanks, 1989). The **Pointwise Mutual Information** between two words $word_1$ and $word_2$ can be defined as follows,

$$PMI(word_1, word_2) = \log_2 \left(\frac{P(word_1 \ \& \ word_2)}{P(word_1) \bullet P(word_2)} \right)$$

Here, $P(word_1 \ \& \ word_2)$ denotes the probability of the co-occurence of both $word_1$ and $word_2$. The Semantic Oreintation(SO) of a phrase is defined as follows,

$$SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$$

2.2