**IT -633 Data Mining**

Sentiment Analysis for

Reddit Comments



Project Proposal

Mentor : Prof. P.M.Jat

Team - 10

Nihit Alamuru-201301036

Ankit Taral-201301087

Rahul Kumar-201301204

Samarth Desai-201301456

**INDEX**

Abstract……………………………………………………...03

Project Objective…………………………………………...04

Project Outcome…………………………………………....04

Methodology………………………………………………...05

Tools and Techniques………………………………………06

Implementation Plan………………………………………..06

Tentative Dates and Milestones…………………………...07

References…………………………………………………. 07

**Abstract**

On a variety of online platforms, such as review sites, blogs, as well as social services such as Twitter and Reddit, internet users produce vast amounts of opinionated text about a large range of domains, such as movie reviews, travel experiences, product reviews, opinions about news and others.Automatic opinion mining - the ability to process large amounts of opinionated textual information from online sources without human interference - is necessary. Automatic opinion mining is divided into two categories; qualitative opinion mining, which attempts to extract pieces of literal information from the data, such as sentences describing an experience relevant to the target of the opinion and quantitative opinion mining, which attempts to determine quantifiable dimensions of opinion, such as sentiment. Sentiment analysis is utilized in order to determine the polarity of opinions (positive/neutral/negative), or the emotional charge of opinions across a range of possible emotions (love, fear, anger, understanding etc).

Reddit released dataset containing all ~1.7 billion of their publicly available comments. The full dataset is and unwieldy 1+ TB uncompressed, so we’ve decided to to work on a subset from a particular subreddit during May 2015 for this project.The dataset is categorised with 22 fields.

**Objective**

Sentiment analysis can be approached from various directions , such as application of lexicons with manually or semi-automatically annotated word polarities ,Natural Language processing methods and machine learning based approaches.The goal of our project is to apply machine learning for sentiment analysis or opinion mining on user generated comments on our subset(subreddit). We are looking to classify comments as being positive or negative. An opinion is defined as a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion holder.

**Outcome**

The outcome of the project is to classify the comment as being positive or negative. Given a comment of a subreddit our model will predict the class of the comment doing sentiment analysis. Using the comment of the user we would be able to predict his/her sentiment regarding the product. This would help companies to improve their product. We would be able to predict the sentiment of a redditor on some important matter.

For example, we can classify them as liberals and conservatives using their comment on news post.

**Methodology**

Reddit offers a simple Application Programming Interface (API). To facilitate data collection from Reddit, we will implement the Reddit API and a scraper application to get data of particular subreddit we are interested in. The data is classified in 22 fields.

We explore methods for sentiment prediction of comments to posts. We utilise a machine learning approach for the training of our sentiment prediction systems. We train our prediction systems on (objective) features from posts and the prevalent subjective opinion label of the comments to that post.

We use basic approach proposed by Go et al [1] for construction of our base classifier. In order to improve the performance of our base classifier we will additionally try Naive Bayes classifier.

We train our sentiment prediction systems on various postings on different categories with comments. We will calculate the distribution of the positive polarity strengths over categories and user groups. In order to establish baseline for our work, we will implement a sentiment prediction proposed by Balasubramanyam [2]. We will only implement Support Vector Machine.

We will use a dictionary of bad words to identify bad words in a comment. This will helpful in predicting the positivity or negativity of the comment subjectively. This will help in predicting if the comment is an insult. For eg. a comment is an insult if the is directed towards a second person. For this we have to search bad words around variations of “you” (such as “u”, “your” etc).

**Tools and Technologies**

* Reddit API
* Google Visualisation API
* Google charts API
* Python Machine learning and NLP Libraries
* GitHub
* AlchemyLanguage by IBM

**Implementation Plan**

The implementation plan consists of the following steps:

* Gather Dataset
* Preprocess Dataset
* Collect Training Data
* Design and Implement the system using tools mentioned
* Train the system
* Evaluate the system

**Milestones and Expected dates**

|  |  |
| --- | --- |
| Milestone | Expected date of completion |
| Phase 1: Gather dataset, preprocess dataset, collect training data. | 3rd Oct. |
| Phase 2: Design and Implement the system using tools mentioned | 29th Oct. |
| Phase 3: Train system followed by evaluation of system | 12th Nov. |

**References**

**[1]** A. Go, L. Huang and R. Bhayani, "Twitter Sentiment Classification using Distant Supervision," The Stanford Natural Language Processing Group, 2008/2009.

**[2]** R. Balasubramanyan, W. W. Cohen, D. Pierce and D. P. Redlawsk, "What pushes their buttons? Predicting comment polarity from the content of political blog posts," in Workshop on Language in Social Media, Portland, Oregon, USA, 2011**.**