Sentiment Analysis of news comments

1. Problem or hypothesis

With the proliferation of Web 2 applications such as microbloging, forums and social networks, there came reviews, comments, recommendation, ratings and feedbacks generated by users. The user generated content can be about virtually anything including politicians, products, people, events, etc. With the explosion of user generated content came the need by companies, politicians, service providers, social psychologists, researchers to mine and analyze the content for different uses. The bulk of this user generated content required the use of automated techniques for mining and analyzing since manually mining and analyzing is difficult for such a huge content.

Today, traditional news companies have gone online or have an online version of their news allowing news consumers to express their opinions about the news article, something that is almost impossible in the traditional printed-news. And this, like in blogs, product /movie reviews, has brought about a bulk of user generated content. and with it, the need by news agencies to know how their news consumers have been affected. they are interested if news consumers reacted in a negative or a positive way. The hypothesis here is that the polarity of news comments can be automatically and reasonably identified. The objective of this thesis is to find techniques to automatically classify news comments as positive or negative with respect to the news article.

2. Importance of research

The general area of sentiment analysis has importance in helping companies and service provides to tailor their products to user needs. **Wright (2009) claims that "for many businesses,** online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace". But it is also of paramount importance for pure academicians, social psychologists since it opens a window to tap into the psychological thinking and reactions of online communities to study the general mind-state of communities at a particular time with regard to some issue. It can also be of use for political analysts to predict election results during the campaign stage of political elections.

Since news articles are made available to communities with the intent of impacting them in different ways, it is of great importance for news agencies to know how their news consumers have been affected or impacted by reading the news articles. Knowing news consumer reactions to news can also be used in decision making by politicians and policy makers. News agencies can also use consumer reactions to better their news coverage, presentation, and content.

3. Prior research

There have been many researches on the subject of sentiment analysis. There already exist sentiment analysis studies for movie/product reviews, blogs, twitter messages. The main research on sentiment analysis so far has mainly focused on two things: identifying whether a given textual entity is subjective or objective, and identifying the polarity of that subjective textual entity after removing the objective content. Most sentiment analysis research is done with help of supervised machine learning techniques, but of course there are also some researches that use unsupervised machine learning techniques and/or statistical approaches. For machine learning approaches, a bag of words representation is used. Recently, there have been feature based approaches to improve results. A baseline for sentiment analysis studies is the a classification result by support vector machine(SVM), a supervised machine learning technique.

Some related research to what this thesis is going to do is a study by Ahmad etal (2007) which studied a computable metric to find polarity financial news articles. There are also some studies about sentiment on a given topic (Nasukawa, 2003) . Also related is a term based sentiment analysis (Bhayani, 2009).

4. Possible research approach or methodology

The possible research methodology will be, first to manually annotate news comments provided by users as negative or positive. This will serve as a gold standard. Then I will use two approaches: one is by using key-word based approach. In this keyword based approach, I will first create two groups of keywords – one class of positive keywords, and another class of negative keywords. Then, on the basis of these keywords, I will classify news comments as positive or negative comments. Neutral comments will be ignored. The second approach will be to use machine learning techniques. The machine learning approach will develop a classifier that will classify news comments as positive or negative. The classifier will be trained on the basis of some manually classified positive and negative comments. Since I will be using news articles whose comments come from twitter, the use of emoticons will be useful in training the classifier.

4. Data

Data for this thesis is obtained from a Groningen IT Company called Heeii. The data contains news tittles with a small description of the news and the comments about that news.The news comments are extracted from twiitter for the news posted on a news agency website called the denever post

5. Planning

5.1 Time Table

The subtasks of the thesis and estimated finish time are provided in the table below

Subtasks	**Finish Time**
Literature Review	Nov 10
Manual data annotation	Nov 24
Feature selection	Nov 30
Implementing and Testing	Dec 30
Tsesis Writing	Jan 30

6. Potential outcomes of research

The main outcome of the research will be a classifier that can be used to automatically classify news comments into positive and negative. Of course, it will also provide insights into better machine learning techniques for such a task, features, and important trends in news comments on the way.