Text Analysis Platform

Author: Ashish Mukherjee **Date: 21 October 2013**

Document Scope

This document encompasses use cases and high-level technology information about text analysis tools, particularly for email analysis.

Use Cases

#1 Customer sentiment

Customer service/support centres handle large volumes of mail. The following may be supported --

1) Time series analysis of levels of customer satisfcation/dissatisfaction
2) Metrics reporting of satisfcation/dissatisfaction handling by different CSRs
3) Detecting high-levels of customer dissatisfaction for timely escalation and intervention

#2 Detecting employee morale

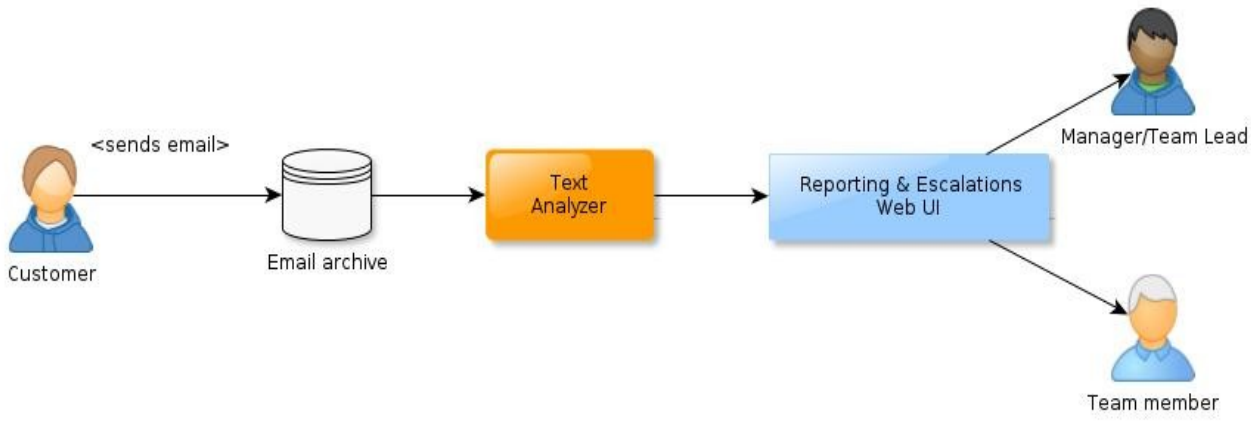
Derive insights into employee mood/sentiment about policy changes, products/services of the company or events based on a search. eg. Search the mail repository for 'leave policy' to gauge the mood related to recent changes in leave policy. The mails which come up in search result would have associated sentiment indicated against them.

#3 Issue Detection - Detect most frequently faced customer issues and pain-points by watching out for keywords in email. eg. Shipping, handling

Technolog High-level View

The NLP layer uses Apache OpenNLP to perform text analysis. Different lexicons and rule-based algorithm are used in the course of analysis. These are derived from lexicons arising out of widely published research.

Email Analysis Scenario
The analysis service runs on the same server infrastructure as the IDEA email archival platform for security and performance. A Java server daemon reads new emails from the IDEA queue and returns percentage of positive and negative sentiment and the total sentiment expressed in the mail, which are then stored back into the mail repository.



Analysis of other Text Sources
If the solution is used for any other text data outside email content, then any input which can be converted to text can be provided and a text output would be generated which can be imported into an application which can read text (such as Excel). The application would be hosted on a multi-tenanted Cloud server on the Web.

Comparison with other Sentiment Analyzers

We considered Textalytics and Text-Processing as our nearest competitiors.

1. Number of services are tailored towards social network posts like Tweets or Facebook comments. These are of a different nature (shorter in length, different language) and have training inputs other than annotated text (eg. #tag, 'Like' etc.). Enterprise email or user reviews content is often different from this kind of input.

2. Another category of services are trained on product review data such as books, movie reviews which are rather different from email or enterprise content. These are based on very high training requirements to make them perform to acceptable levels.

3. The existing applications are exposed as third-party Web Services and this would not be able to meet our performance needs at high volume nor address our privacy concerns. Our solution can be deployed in different environments based on needs.

4. All the applications evaluated provide a binary classification of positive, negative or neutral (in absence of any class). We saw a need to be able to arrive at an outcome of mixed sentiments commonly conveyed within email or other types of content.

5. Customization to the lexicon is facilitated by the platform, as opposed to the standard trained models

6. Not just word-based sentiment, but also sentiments of phrases are supported.

We follow a combination of rule-based and lexicon/meta-data driven methodology which ensures better results and allows room for customization to your needs as best as possible.

Operational Deployment

Customer will be provided with periodic updates to the lexicon & application on subscription basis. Further, based on requirements, we can provide front-end tools for to customize the lexicon or set configuration thresholds for sentiment-based categorization.

Testing

Test Data-set: Enron email databank
Sample size: ~ 500
Sample Selection methodology: Random within Ken Lay's inbox

Testing Objective

Estimate accuracy of analysis

Test Observations

Total mails analyzed =~ 489 (excluding some empty ones)
Wrong analysis = 40
Accuracy rate = 91.8%

#mails which are approximately neutral (Score <= 1.50) - 337
#mails with score > 1.50 and pos = neg - 1
#mails with score > 1.50 and pos < neg - 40
#mails with score > 1.50 and pos > neg - 121

We chose the score of 1.50 as threshold to indicate whether mail has negligible sentiment or not. This value was based on human perusal/understanding of the data-set.

This has more than average negative mails for internal corporate context, because Enron was going through troubled times. Many of the mails with positive tilt were actually encouraging mails or mails containing praise or some good news.

Outcomes

From our understanding of the technology and its limitations and testing carried out, we arrive at the following conclusions -

1) Usability

The understanding of what emails are important might vary according to data-set and business context and is a control which should be with the user to decide what kind of mails deserve his/her attention (% threshold of positive & negative sentiments respectively and threshold of absolute sentiment value).

2) Sarcasm Detection

Sarcasm cannot be detected due to sentence constructions. This is a NLP research field in sentiment analysis and still not mature. We will pursue work on this in future releases.

3) Improvement of Accuracy

NLP, like all machine learning areas is a probabilistic science and is never perfect. We use controlled vocabularies based on reputed online lexicon and corpora and continually improve the quality of these. This involves intervention of engineers and an English language resource.

If the user would like to flag an email which was incorrectly analyzed in his view, he/she may do so and our team will examine it and take appropriate action. Another hurdle in the way of accuracy could be context and to address this, we could provide a mechanism to the user to directly contribute to the lexicon to support organization-specific terminology or context-specific vocabulary.

Background Research

A study of the research done in this field was carried out to form a broader understanding of the techniques successfully employed in Sentiment Analysis. This included publications, PowerPoint presentations and developer blogs. Here are a few relevant citations from the Web -

http://www.academia.edu/1336655/Reviews_Classification_Using_SentiWordNet_Lexicon
http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76.2378&rep=rep1&type=pdf
http://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf