User-directed Sentiment Analysis: Visualizing the Affective Content of Documents

Michelle L. Gregory
PNNL
902 Battelle Blvd.
Richland Wa. 99354
michelle.gregory@pnl.gov

Nancy Chinchor
Consultant
chinchor@earthlink.net

Paul Whitney
PNNL
902 Battelle Blvd.
Richland Wa. 99354
paul.whitney@pnl.gov

Richard Carter
PNNL
902 Battelle Blvd.
Richland Wa. 99354
richard.carter@pnl.gov

Elizabeth Hetzler
PNNL
902 Battelle Blvd.
Richland Wa. 99354
beth.hetzler@pnl.gov

Alan Turner
PNNL
902 Battelle Blvd.
Richland Wa. 99354
alan.turner@pnl.gov

Abstract

Recent advances in text analysis have led to finer-grained semantic analysis, including *automatic sentiment analysis*—the task of measuring documents, or chunks of text, based on emotive categories, such as *positive* or *negative*. However, considerably less progress has been made on efficient ways of exploring these measurements. This paper discusses approaches for visualizing the affective content of documents and describes an interactive capability for exploring emotion in a large document collection.

1 Introduction

Recent advances in text analysis have led to finer-grained semantic classification, which enables the automatic exploration of subtle areas of meaning. One area that has received a lot of attention is *automatic sentiment analysis*—the task of classifying documents, or chunks of text, into emotive categories, such as *positive* or *negative*. Sentiment analysis is generally used for tracking people's attitudes about particular individuals or items. For example, corporations use sentiment analysis to determine employee attitude and customer satisfaction with their products. Given the plethora of data in digital form, the ability to accurately and efficiently measure the emotional content of documents is paramount.

The focus of much of the automatic sentiment analysis research is on identifying the *affect bearing* words (words with emotional content) and on measurement approaches for sentiment (Turney & Littman, 2003; Pang & Lee, 2004; Wilson et al., 2005). While identifying related content is an essential component for automatic sentiment analysis, it only provides half the story. A useful area of research that has received much less attention is how these measurements might be presented to the users for exploration and added value.

This paper discusses approaches for visualizing affect and describes an interactive capability for exploring emotion in a large document collection. In Section 2 we review current approaches to identifying the affective content of documents, as well as possible ways of visualizing it. In Section 3 we describe our approach: The combination of a lexical scoring method to determine the affective content of documents and a visual analytics tool for visualizing it. We provide a detailed case study in Section 4, followed by a discussion of possible evaluations.

2 Background

At the AAAI Symposium on Attitude and Affect held at Stanford in 2004 (Qu et al., 2005), it was clear that the lexical approach to capturing affect was adequate for broad brush results, but there were no production quality visualizations for presenting those results analytically. Thus, we began exploring methods and tools for the visualization of lexically-based approaches for measuring affect which could facilitate the exploration of affect within a text collection.

2.1 Affect Extraction

Following the general methodology of informational retrieval, there are two pre-dominant methods for identifying sentiment in text: Text classification models and lexical approaches. Classification models require that a set of documents are hand labeled for affect, and a system is

Proceedings of the Workshop on Sentiment and Subjectivity in Text, pages 23–30,
Sydney, July 2006. ©2006 Association for Computational Linguistics

trained on the feature vectors associated with labels. New text is automatically classified by comparing the feature vectors with the training set. (Pang & Lee, 2004; Aue & Gamon, 2005). This methodology generally requires a large amount of training data and is domain dependent.

In the lexical approach, documents (Turney & Littman, 2003), phrases (see Wilson et al., 2005), or sentences (Weibe & Riloff, 2005) are categorized as *positive* or *negative*, for example, based on the number of words in them that match a lexicon of sentiment bearing terms. Major drawbacks of this approach include the contextual variability of sentiment (what is *positive* in one domain may not be in another) and incomplete coverage of the lexicon. This latter drawback is often circumvented by employing *bootstrapping* (Turney & Littman, 2003; Weibe & Riloff, 2005) which allows one to create a larger lexicon from a small number of seed words, and potentially one specific to a particular domain.

2.2 Affect Visualization

The uses of automatic sentiment classification are clear (public opinion, customer reviews, product analysis, etc.). However, there has not been a great deal of research into ways of visualizing affective content in ways that might aid data exploration and the analytic process.

There are a number of visualizations designed to reveal the emotional content of text, in particular, text that is thought to be highly emotively charged such as conversational transcripts and chat room transcripts (see DiMicco et al., 2002; Tat & Carpendale, 2002; Lieberman et al., 2004; Wang et al., 2004, for example). Aside from using color and emoticons to explore individual documents (Liu et al., 2003) or email inboxes (Mandic & Kerne, 2004), there are very few visualizations suitable for exploring the affect of large collections of text. One exception is the work of Liu et al. (2005) in which they provide a visualization tool to compare reviews of products,using a bar graph metaphor. Their system automatically extracts product features (with associated affect) through parsing and pos tagging, having to handle exceptional cases individually. Their Opinion Observer is a powerful tool designed for a single purpose: comparing customer reviews.

In this paper, we introduce a visual analytic tool designed to explore the emotional content of large collections of open domain documents. The tools described here work with document collections of all sizes, structures (html, xml, .doc, email, etc), sources (private collections, web, etc.), and types of document collections. The visualization tool is a mature tool that supports the analytical process by enabling users to explore the thematic content of the collection, use natural language to query the collection, make groups, view documents by time, etc. The ability to explore the emotional content of an entire collection of documents not only enables users to compare the range of affect in documents within the collection, but also allows them to relate affect to other dimensions in the collection, such as major topics and themes, time, and source.

3 The Approach

Our methodology combines a traditional lexical approach to scoring documents for affect with a mature visualization tool. We first automatically identify affect by comparing each document against a lexicon of affect-bearing words and obtain an affect score for each document. We provide a number of visual metaphors to represent the affect in the collection and a number of tools that can be used to interactively explore the affective content of the data.

3.1 Lexicon and Measurement

We use a lexicon of affect-bearing words to identify the distribution of affect in the documents. Our lexicon authoring system allows affect-bearing terms, and their associated strengths, to be bulk loaded, declared manually, or algorithmically suggested. In this paper, we use a lexicon derived from the General Inquirer (GI) and supplemented with lexical items derived from a semi-supervised bootstrapping task. The GI tool is a computer-assisted approach for content analyses of textual data (Stone, 1977). It includes an extensive lexicon of over 11,000 hand-coded word stems and 182 categories.

We used this lexicon, specifically the *positive* and *negative* axes, to create a larger lexicon by bootstrapping. Lexical bootstrapping is a method used to help expand dictionaries of semantic categories (Riloff & Jones, 1999) in the context of a document set of interest. The approach we have adopted begins with a lexicon of affect bearing words (POS and NEG) and a corpus. Each document in the corpus receives an affect score by counting the number of words from the seed lexicon that occur in the document; a separate score is given for each affect axis. Words in the corpus are scored for affect potential by comparing their distribution (using an L1 Distribution metric) of occurrence over the set if documents to the distribution of affect bearing words. Words that compare favorably with affect are hypothesized as affect bearing words. Results are then manually culled to determine if in fact they should be included in the lexicon.

Here we report on results using a lexicon built from 8 affect categories, comprising 4 concept pairs:

- Positive (*n*=2236)-Negative (*n*=2708)
- Virtue (*n*=638)-Vice (*n*=649)
- Pleasure (*n*=151)-Pain (*n*=220)
- Power Cooperative (*n*=103)-Power Conflict (*n*=194)

Each document in the collection is compared against all 8 affect categories and receives a score for each. Scores are based on the summation of each affect axis in the document, normalized by the number of words in the documents. This provides an overall proportion of *positive* words, for example, per document. Scores can also be calculated as the summation of each axis, normalized by the total number of affect words for all axes. This allows one to quickly estimate the balance of affect in the documents. For example, using this measurement, one could see that a particular document contains as many *positive* as *negative* terms, or if it is heavily skewed towards one or the other.

While the results reported here are based on a predefined lexicon, our system does include a *Lexicon Editor* in which a user can manually enter their own lexicon or add strengths to lexical items. Included in the editor is a *Lexicon Bootstrapping Utility* which the user can use to help create a specialized lexicon of their own. This utility runs as described above. Note that while we enable the capability of strength, we have not experimented with that variable here. All words for all axes have a default strength of .5.

3.2 Visualization

To visualize the affective content of a collection of documents, we combined a variety of visual metaphors with a tool designed for visual analytics of documents, IN-SPIRE.

3.2.1 The IN-SPIRE System

IN-SPIRE (Hetzler and Turner, 2004) is a visual analytics tool designed to facilitate rapid understanding of large textual corpora. IN-SPIRE generates a compiled document set from *mathematical signatures* for each document in a set. Document signatures are clustered according to common themes to enable information exploration and visualizations. Information is presented to the user using several *visual metaphors* to expose different facets of the textual data. The central visual metaphor is a **Galaxy view** of the corpus that allows users to intuitively interact with thousands of documents, examining them by theme (see Figure 4, below). IN-SPIRE leverages the use of context vectors such as LSA (Deerwester et al., 1990) for document clustering and projection. Additional analytic tools allow exploration of temporal trends, thematic distribution by source or other metadata, and query relationships and overlaps. IN-SPIRE was recently enhanced to support visual analysis of sentiment.

3.2.2 Visual Metaphors

In selecting metaphors to represent the affect scores of documents, we started by identifying the kinds of questions that users would want to explore. Consider, as a guiding example, a set of customer reviews for several commercial products (Hu & Liu, 2004). A user reviewing this data might be interested in a number of questions, such as:

- What is the range of affect overall?
- Which products are viewed most positively? Most negatively?
- What is the range of affect for a particular product?
- How does the affect in the reviews deviate from the norm? Which are more negative or positive than would be expected from the averages?
- How does the feedback of one product compare to that of another?
- Can we isolate the affect as it pertains to different features of the products?

In selecting a base metaphor for affect, we wanted to be able to address these kinds of questions. We wanted a metaphor that would support viewing affect axes individually as well as in pairs. In addition to representing the most common axes, negative and positive, we wanted to provide more flexibility by incorporating the ability to portray multiple pairs because we suspect that additional axes will help the user explore nuances of emotion in the data. For our current metaphor, we drew inspiration from the Rose plot used by Florence Nightingale (Wainer, 1997). This metaphor is appealing in that it is easily interpreted, that larger scores draw more attention, and that measures are shown in consistent relative location, making it easier to compare measures across document groups. We use a modified version of this metaphor in which each axis is represented individually but is also paired with its opposite to aid in direct comparisons. To this end, we vary the spacing between the rose petals to reinforce the pairing. We also use color; each pair has a common hue, with the more positive of the pair shown in a lighter shade and the more negative one in a darker shade (see Figure 1).

To address how much the range of affect varies across a set of documents, we adapted the concept of a box plot to the rose petal. For each axis, we show the median and quartile values as shown in the figure below. The dark line indicates the median value and the color band portrays the quartiles. In the plot in Figure 1, for example, the scores vary quite a bit.

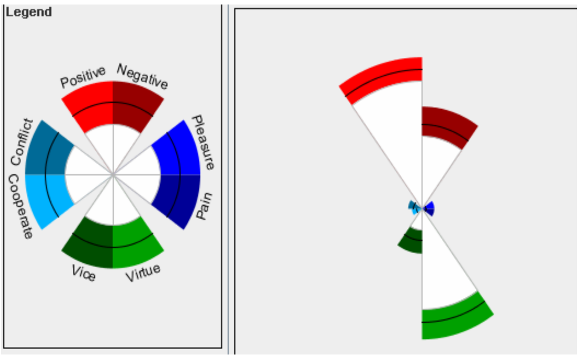


Figure 1. Rose plot adapted to show median and quartile variation.

Another variation we made on the base metaphor was to address a more subtle set of questions. It may happen that the affect scores within a dataset are largely driven by document membership in particular groups. For example, in our customer data, it may be that all documents about Product A are relatively positive while those about Product B are relatively negative. A user wanting to understand customer complaints may have a subtle need. It is not sufficient to just look at the most negative documents in the dataset, because none of the Product A documents may pass this threshold. What may also help is to look at all documents that are more negative than one would expect, given the product they discuss. To carry out this calculation, we use a statistical technique to calculate the Main (or expected) affect value for each group and the Residual (or deviation) affect value for each document with respect to its group (Scheffe, 1999).

To convey the Residual concept, we needed a representation of deviation from expected value. We also wanted this portrayal to be similar to the base metaphor. We use a unit circle to portray the expected value and show deviation by drawing the appropriate rose petals either outside (larger than expected) or inside (smaller than expected) the unit circle, with the color amount showing the amount of deviation from expected. In the figures below, the dotted circle represents expected value. The glyph on the left shows a cluster with scores slightly higher than expected for Positive and for Cooperation affect. The glyph on the right shows a cluster with scores slightly higher than expected for the Negative and Vice affect axes (Figure 2).

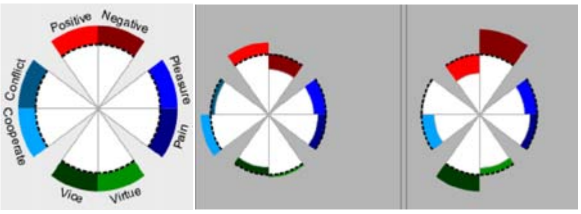


Figure 2. Rose plot adapted to show deviation from expected values.

3.2.3 Visual Interaction

IN-SPIRE includes a variety of analytic tools that allow exploration of temporal trends, thematic distribution by source or other metadata, and query relationships and overlaps. We have incorporated several interaction capabilities for further exploration of the affect. Our analysis system allows users to group documents in numerous ways, such as by query results, by metadata (such as the product), by time frame, and by similarity in themes. A user can select one or more of these groups and see a summary of affect and its variation in those groups. In addition, the group members are clustered by their affect scores and glyphs of the residual, or variation from expected value, are shown for each of these sub-group clusters.

Below each rose we display a small histogram showing the number of documents represented by that glyph (see Figure 3). These allow comparison of affect to cluster or group size. For example, we find that extreme affect scores are typically found in the smaller clusters, while larger ones often show more mid-range scores. As the user selects document groups or clusters, we show the proportion of documents selected.

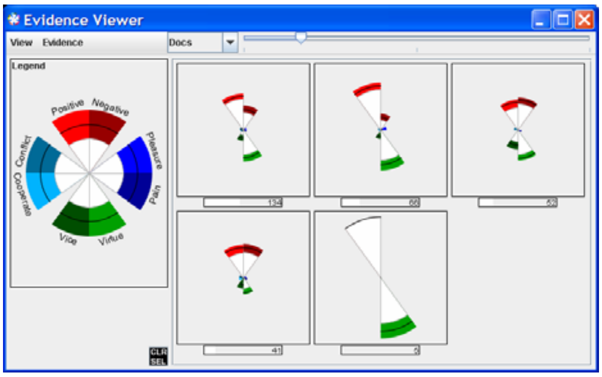


Figure 3. Clusters by affect score, with one rose plot per cluster.

The interaction may also be driven from the affect size. If a given clustering of affect characteristics is selected, the user can see the themes they represent, how they correlate to metadata, or the time distribution. We illustrate how the affect visualization and interaction fit into a larger analysis with a brief case study.

4 Case study

The IN-SPIRE visualization tool is a non-data specific tool, designed to explore large amounts of textual data for a variety of genres and document types (doc, xml, etc). Many users of the system have their own data sets they wish to explore (company internal documents), or data can be harvested directly from the web, either in a single web harvest, or dynamically. The case study and dataset presented here is intended as an example only, it does not represent the full range of exploration capabilities of the affective content of datasets.

We explore a set of customer reviews, comprising a collection of Amazon reviews for five products (Hu & Liu, 2004). While a customer may not want to explore reviews for 5 different product types at once, the dataset is realistic in that a web harvest of one review site will contain reviews of multiple products. This allows us to demonstrate how the tool enables users to focus on the data and comparisons that they are interested in exploring. The 5 products in this dataset are:

- Canon G3; digital camera
- Nikon coolpix 4300; digital camera
- Nokia 6610; cell phone
- Creative Labs Nomad Jukebox Zen Xtra 40GB; mp3 player
- Apex AD2600 Progressive-scan DVD player

We begin by clustering the reviews, based on overall thematic content. The labels are automatically generated and indicate some of the stronger theme combinations in this dataset. These clusters are driven largely by product vocabulary. The two cameras cluster in the lower portion; the Zen shows up in the upper right clusters, with the phone in the middle and the Apex DVD player in the upper left and upper middle. In this image, the pink dots are the Apex DVD reviews.

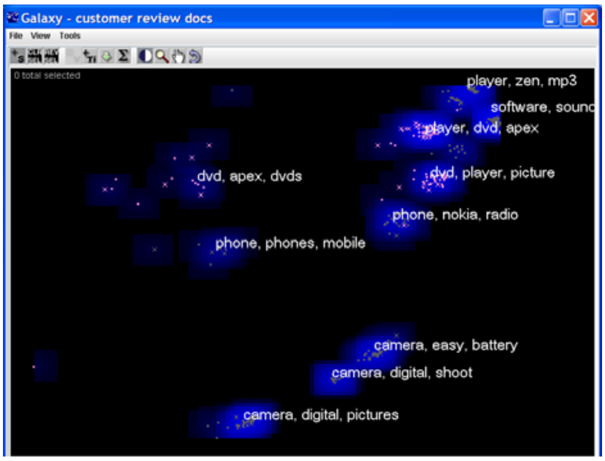


Figure 4. Thematic clustering of product review

The affect measurements on these documents generate five clusters in our system, each of which is summarized with a rose plot showing affect variation. This gives us information on the range and distribution of affect overall in this data. We can select one of these plots, either to review the documents or to interact further. Selection is indicated with a green border, as shown in the upper middle plot of Figure 5.

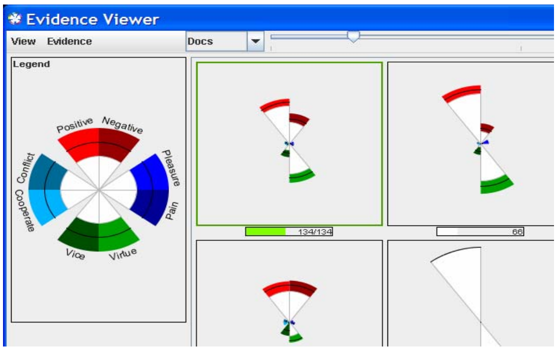


Figure 5. Clusters by affect, with one cluster glyph selected.

The selected documents are relatively positive; they have higher scores in the Positive and Virtue axes and lower scores in the Negative axis. We may want to see how the documents in this affect cluster distribute over the five products. This question is answered by the correlation tool, shown in Figure 6; the positive affect cluster contains more reviews on the Zen MP3 player than any of the other products.

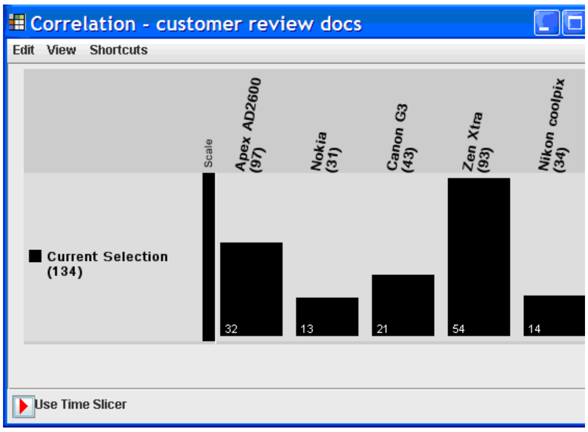


Figure 6. Products represented in one of the positive affect clusters.

Alternatively we could get a summary of affect per product. Figure 7 shows the affect for the Apex DVD player and the Nokia cell phone. While both are positive, the Apex has stronger negative ratings than the Nokia.

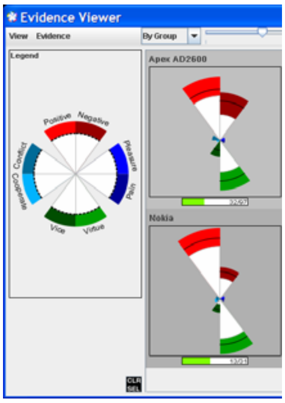


Figure 7. Comparison of Affect Scores of Nokia to Apex

More detail is apparent by looking at the clusters within one or more groups and examining the deviations. Figure 8 shows the sub-clusters within the Apex group. We include the summary for the group as a whole (directly beneath the Apex label), and then show the four sub-clusters by illustrating how they deviate from expected value. We see that two of these tend to be more positive than expected and two are more negative than expected.

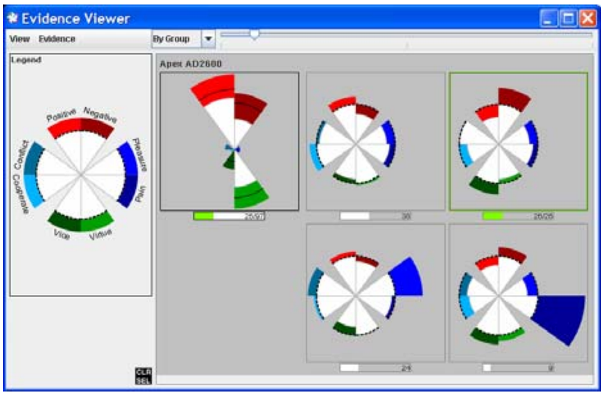


Figure 8. Summary of Apex products with sub-clusters showing deviations.

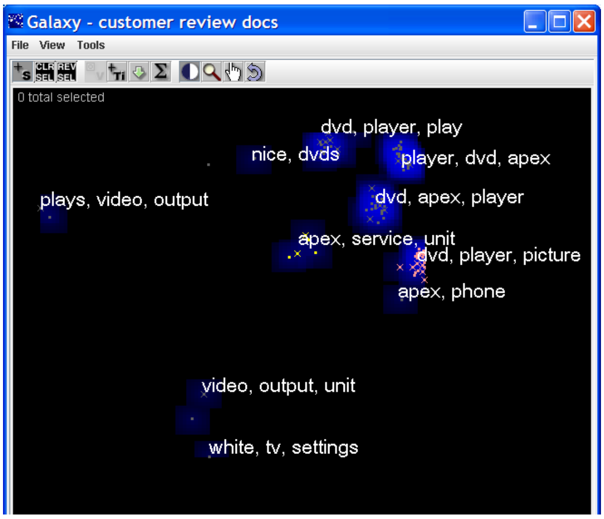


Figure 9. Thematic distribution of reviews for one product (Apex).

Looking at the thematic distribution among the Apex documents shows topics that dominate its reviews (Figure 9).

We can examine the affect across these various clusters. Figure 10 shows the comparison of the “service” cluster to the “dvd player picture” cluster. This graphic demonstrates that documents with “service” as a main theme tend to be much more negative, while documents with “picture” as a main theme are much more positive.

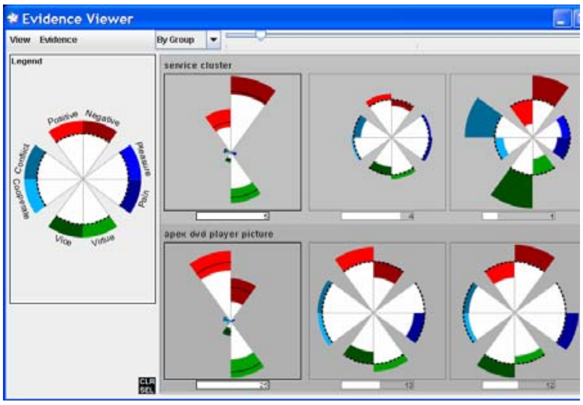
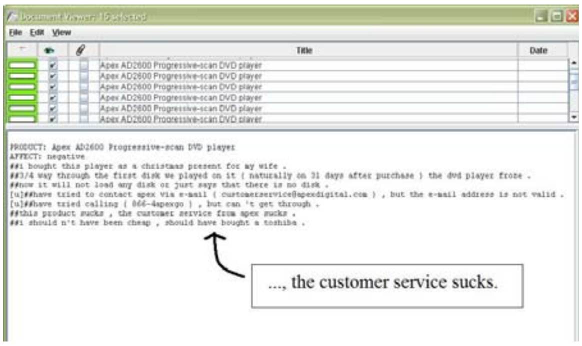


Figure 10. Affect summary and variation for "service" cluster and "picture" cluster.

The visualization tool includes a document viewer so that any selection of documents can be reviewed. For example, a user may be interested in why the "service" documents tend to be negative, in which case they can review the original reviews. The doc viewer, shown in Figure 11, can be used at any stage in the process with any number of documents selected. Individual documents can be viewed by clicking on a document title in the upper portion of the doc viewer.



			Title	Date
	✓	□	Apex AD2600 Progressive-scan DVD player	
	✓	□	Apex AD2600 Progressive-scan DVD player	
	✓	□	Apex AD2600 Progressive-scan DVD player	
	✓	□	Apex AD2600 Progressive-scan DVD player	
	✓	□	Apex AD2600 Progressive-scan DVD player	
	✓	□	Apex AD2600 Progressive-scan DVD player	

```
PRODUCT: Apex AD2600 Progressive-scan DVD player
AFFECT: negative
##i bought this player as a christmas present for my wife .
##3/4 way through the first disk we played on it ( naturally on 31 days after purchase ) the dvd player froze .
##now it will not load any disk or just says that there is no disk .
[u]##have tried to contact apex via e-mail ( customerservice@apexdigital.com ) , but the e-mail address is not valid .
[u]##have tried calling ( 866-4apexgo ) , but can 't get through .
##this product sucks , the customer service from apex sucks .
##i should n't have been cheap , should have bought a toshiba .
```

Figure 11: The Doc Viewer.

In this case study, we have illustrated the usefulness of visualizing the emotional content of a document collection. Using the tools presented here, we can summarize the dataset by saying that in general, the customer reviews are positive (Figure 5), but reviews for some products are more positive than others (Figures 6 and 7). In addition to the general content of the reviews, we can narrow our focus to the features contained in the reviews. We saw that while reviews for Apex are generally positive (Figure 8), reviews about Apex "service" tend to be much more negative than reviews about Apex "picture" (Figure 10).

5 Evaluation

IN-SPIRE is a document visualization tool that is designed to explore the thematic content of a large collection of documents. In this paper, we have described the added functionality of exploring affect as one of the possible dimensions. As an exploratory system, it is difficult to define appropriate evaluation metric. Because the goal of our system is not to discretely bin the documents into affect categories, traditional metrics such as precision are not applicable. However, to get a sense of the coverage of our lexicon, we did compare our measurements to the hand annotations provided for the customer review dataset.

The dataset had hand scores (-3-3) for each feature contained in each review. We summed these scores to discretely bin them into positive (>0) or negative (<0). We did this both at the feature level and the review level (by looking at the cumulative score for all the features in the review). We compared these categorizations to the scores output by our measurement tool. If a document had a higher proportion of positive words than negative, we classified it as positive, and negative if it had a higher proportion of negative words. Using a chi-square, we found that the categorizations from our system were related with the hand annotations for both the whole reviews (chi-square=33.02, df=4, p<0.0001) and the individual features (chi-square=150.6, df=4, p<0.0001), with actual agreement around 71% for both datasets. While this number is not in itself impressive, recall that our lexicon was built independently of the data for which is was applied. W also expect some agreement to be lost by conflating all scores into discrete bins, we expect that if we compared the numeric values of the hand annotations and our scores, we would have stronger correlations.

These scores only provide an indication that the lexicon we used correlates with the hand annotations for the same data. As an exploratory system, however, a better evaluation metric would be a user study in which we get feedback on the usefulness of this capability in accomplishing a variety of analytical tasks. IN-SPIRE is currently deployed in a number of settings, both commercial and government. The added capabilities for interactively exploring affect have recently been deployed. We plan to conduct a variety of user evaluations *in-situ* that focus on its utility in a number of different tasks. Results of these studies will help steer the further development of this methodology.

6 Conclusion

We have developed a measurement and visualization approach to affect that we expect to be useful in the context of the IN-SPIRE text analysis toolkit. Our innovations include the flexibility of the lexicons used, the measurement options, the bootstrapping method and utility for lexicon development, and the visualization of affect using rose plots and interactive exploration in the context of an established text analysis toolkit. While the case study presented here was conducted in English, all tools described are language independent and we have begun exploring and creating lexicons of affect bearing words in multiple languages.

References

A. Aue. & M. Gamon. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study. Submitted RANLP.

S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science,* 41(6):391–407.

J. M. DiMicco, V. Lakshmipathy, A. T. Fiore. 2002. Conductive Chat: Instant Messaging With a Skin Conductivity Channel. In *Proceedings of Conference on Computer Supported Cooperative Work.*

D. G. Feitelson. 2003. Comparing Partitions with Spie Charts. *Technical Report 2003-87,* School of Computer Science and Engineering, The Hebrew University of Jerusalem.

E. Hetzler and A. Turner. 2004. Analysis Experiences Using Information Visualization. *IEEE Computer Graphics and Applications*, 24(5):22-26, 2004.

M. Hu and B. Liu. 2004. Mining Opinion Features in Customer Reviews. In *Proceedings of Nineteenth National Conference on Artificial Intelligence* (AAAI-2004).

H. Lieberman, H. Liu, P. Singh and B. Barry. 2004. Beating Common Sense into Interactive Applications. *AI Magazine* 25(4): Winter 2004, 63-76.

B. Liu, M. Hu and J. Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th international World Wide Web conference (WWW-2005),* May 10-14, 2005: Chiba, Japan.

H. Liu, T. Selker, H. Lieberman. 2003. Visualizing the Affective Structure of a Text Document. *Computer Human Interaction,* April 5-10, 2003: Fort Lauderdale.

M. Mandic and A. Kerne. 2004. faMailiar—Intimacy-based Email Visualization. In *Proceedings of IEEE Information Visualization 2004,* Austin Texas, 31-32.

B. Pang and L. Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd ACL,* pp. 271-278, 2004.

Y. Qu,, J. Shanahan, and J. Weibe. 2004. Exploring Attitude and Affect in Text: Theories and Applications. Technical Report SS-04-07.

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (AAAI-99) pp. 474-479.

H. Scheffé. 1999. *The Analysis of Variance*, Wiley-Interscience.

P. Stone. 1977. Thematic Text Analysis: New Agendas for Analyzing Text Content. In *Text Analysis for the Social Sciences,* ed. Carl Roberts, Lawrence Erlbaum Associates.

A. Tat and S. Carpendale. 2002. Visualizing Human Dialog. In *Proceedings of IEEE Conference on Information Visualization,* IV'02, p.16-24, London, UK.

P. Turney and M. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems* (TOIS) 21:315-346.

H. Wainer. 1997. A Rose by Another Name.” *Visual Revelations*, Copernicus Books, New York.

H. Wang, H. Prendinger, and T. Igarashi. 2004. Communicating Emotions in Online Chat Using Physiological Sensors and Animated Text.” In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems* (CHI'04), Vienna, Austria, April 24-29.

J. Wiebe and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts.” In *Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.” In *Proceeding of HLT-EMNLP-2005*.