Exploring Sentiment Summarization

Philip Beineke and **Trevor Hastie**
Dept. of Statistics
Stanford University
Stanford, CA 94305

Christopher Manning
Dept. of Computer Science
Stanford University
Stanford CA 94305-9040

Shivakumar Vaithyanathan
IBM Almaden Research Center
650 Harry Rd.
San Jose, CA 95120-6099

Abstract

We introduce the idea of a sentiment summary, a single passage from a document that captures a key aspect of the author's opinion about his or her subject. Using supervised data from the Rotten Tomatoes website, we examine features that appear to be helpful in locating a good summary sentence. These features are used to fit Naive Bayes and regularized logistic regression models for summary extraction.

Introduction

The website Rotten Tomatoes, located at `www.rottentomatoes.com`, is primarily an online repository of movie reviews. For each movie review document, the site provides a link to the full review, along with a brief description of its sentiment. The description consists of a rating ("fresh" or "rotten") and a short quotation from the review. Other research, e.g. Pang et al. (2002), has predicted a movie review's rating from its text. In this paper, we focus on the quotation, which is a main attraction to site users.

A Rotten Tomatoes quotation is typically about one sentence in length and expresses concisely the reviewer's opinion of the movie. To illustrate, Curtis Edmonds's review of the documentary *Spellbound* is encapsulated, "Hitchcock couldn't have asked for a more suspenseful situation." A.O. Scott's review of *Once upon a Time in Mexico* is encapsulated, "A noisy, unholy mess, with moments of wit and surprise that ultimately make its brutal tedium all the more disappointing." A reader can infer from these statements whether or not the overall sentiment is favorable, and get an impression about why. Consequently, we refer to them as *sentiment summaries.*

Apart from movie reviews, it is easy to envision other situations where obtaining such quotations would be useful. A manufacturer may wish to see sentiment summaries for its product reviews; a policy-maker may wish to see them for newspaper editorials; a university lecturer may wish to see them for his or her student feedback. In general, whenever an evaluation is desired, a sentiment summary will enable an interpretation of the reasoning behind it. In this paper, we seek a more precise understanding of the features that can be used productively in the automatic extraction of sentiment summaries.

The task of sentiment summarization is clearly connected with that of text summarization, a field with a rich literature (Mallett 2003). In comparison, sentiment summarization is more narrowly focussed, seeking a broad statement of opinion. As such, it be viewed as a special case of extracting opinion-oriented information. Some issues involved in this task are addressed in Cardie, et al. (2003).

This paper is an exploration of the readily available features that can be used productively for automatically locating sentiment summaries. We examine the summary quotations and identify several features that can be used to predict them. This enables us to fit statistical models to evaluate the usefulness of these features.

Rotten Tomatoes Data

Over 2,500 critics are listed by Rotten Tomatoes. Some write independently, others for a wide range of publications. As a consequence, reviews vary considerably in format, length, and writing style. There are even non-English language reviews included. In order to obtain a reasonably homogeneous collection of full-length reviews, we restrict our attention to the source publications that Rotten Tomatoes terms "The Cream of the Crop." For various reasons,[1] several of these publications were excluded. Thus our final list of sources is restricted to 14 of them.[2] Combined, there are reviews from over 200 critics.

Several thousand (3897) full-text reviews were downloaded and extracted from the web pages on which they

[1] There are many possible reasons that a publication was removed from consideration. These include: few of its reviews have valid links or accompanying quotations; its reviews exist only in audio or video format; its reviews are only a single paragraph in length; its reviews critique multiple movies within a single document; it requires user log-in; or, its reviews reside on web pages whose format makes it cumbersome to separate the review text from other material.

[2] Included publications are The Arizona Republic, The Atlanta Journal-Constitution, The Boston Globe, The Chicago Reader, The Chicago Sun-Times, CNN, The Detroit Free Press, the Detroit News, Entertainment Weekly, the Hollywood Reporter, the Minneapolis Star-Tribune, the New York Post, the San Francisco Chronicle, the Philadelphia Inquirer, and USA Today.

Location	Pct.
Opening	9.2%
Ending	6.0%
Interior	3.7%
Complete	4.4%

Figure 1: Frequency of Quotations by Location within Paragraph

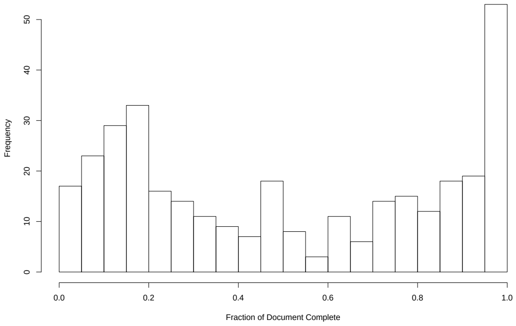

Figure 2: Quotation location within document

reside. Most HTML formatting was removed, although a few features (e.g. paragraph breaks, italics) were retained for modeling purposes, where they are treated identically to word types. The text was then tokenized. All words were shifted to lower-case, and passed through a Porter Stemmer (Porter 1980). In addition, some precautions were taken to ensure that different writing conventions produce the same output. For instance, different methods of writing ". . ." were pooled together.

From there, the Unix command *diff* was used to identify matching substrings between review quotations and their corresponding full text. When the quotation and full text review matched with at most three alterations (inserted strings, deleted strings, or type mis-matches), the capsule was deemed to have been found. If a quotation was shorter than five tokens in length, it was required to match the full text precisely. To avoid artifacts of editing, the Rotten Tomatoes quotations themselves are not used for prediction. Rather, we use the original sentences from which they are drawn.

Descriptive Results

Based on exploratory data analysis, several features appear to be predictive of whether a particular span of text will be chosen as a quotation. These include the following.

1. **Location within Paragraph**

   Figure 1 shows the likelihood of being chosen as a summary quotation for sentences in different locations within a paragraph. In general, the most likely sentence to be chosen is a paragraph's first. The row labelled "Complete" refers to the situation where a paragraph consists of a single sentence.

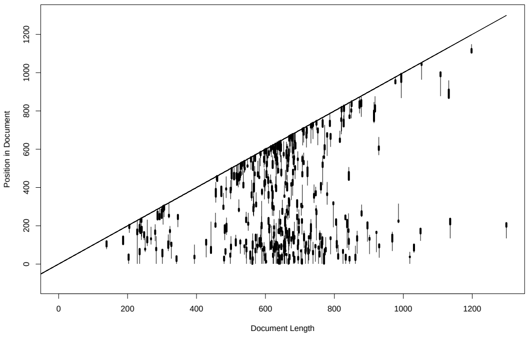

Figure 3: Quotation location versus document length

2. **Location within Document**

   Figure 2 is a histogram of the location of the midpoints of summary quotations within documents. The x-axis indicates the fraction of the document that has been completed. There are two peaks in this plot: one is early in the document, while the other is in the final five percent of the text.

   Figure 3 graphically depicts the position of quotations within documents, using a sample of size 300. In the plot, each vertical line identifies the location of a paragraph that contains a summary quotation. The thicker portion of each line identifies the location of the quotation itself.

   Here we observe a strong interaction between a sentence's location within its paragraph and location within its document. As seen in Figure 1, it is not common for a quotation to be the final sentence in a paragraph; however, the last sentence of the final paragraph is chosen for 16.1% of all quotations.

3. **Word Choice**

   In the entire corpus, 6.7% of tokens occur within summary sentences. Given that dictionary element $w_k \in \mathcal{W}$ occurs n_k times in the corpus, we can compare the number of times it occurs in summary quotations with a binomial random variable whose parameters are $n = n_k$ and $p = 0.067$. Of the 3,537 types that appear between 50 and 500 times in the corpus, 348 of their counts (9.8%) are above the 99-th percentile of the corresponding binomial distribution. By chance, we would only have expected 35 counts to pass this threshold. This suggests that many word types are useful in distinguishing between quotations and other text.

   Figure 4 lists some examples of words that appear in summary quotations more frequently than expected. Often, these words express emotion directly. Words that are interchangeable with "movie" are also more common, as are several other words with a variety meanings and uses. In addition to word types, formatting is also associated with summary quotations. Italicized words and phrases (such as titles) make 8.9 % (893 of 10152) of their appearances

Stemmed Word	In Quot.	Elsewhere	Pct. in Quot.
well-made	10	8	55.6%
craft	34	91	27.2%
mildly	17	46	27.0%
to dazzle	23	77	23.0%
to entertain	144	577	20.0%
nevertheless	19	110	16.6%
movie	915	6881	11.7%
film	797	6990	10.2%
to be	719	7963	8.3 %

Figure 4: Selected words with higher-than-average frequency in quotations

in quotations, while parentheses make only 2.9% (569 of 18375).

Although it is clear that an association exists between many word types and summary quotations, few of these associations appear to be strong. For instance, a sentence containing the word "movie" is less than twice as likely to be selected as a summary as one that does not.

Statistical Models

We approach sentiment summarization as a classification problem at the sentence level. A review document d_i is viewed as a collection of sentences, $(s_{i1}, \ldots, s_{im(d_i)})$. In training data, each sentence s_{ij} is associated with a label y_{ij}.

$$y_{ij} = \begin{cases} 1 & \text{if } s_{ij} \text{ is summary of } d_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In the event that a quotation spans multiple sentences, each is separately labeled as a summary. This is a fairly uncommon occurrence, as the average quotation is 1.15 sentences in length.

For each sentence s_{ij}, we also have two vectors of of features: type features $\mathbf{n}_{ij}$ and location features $\mathbf{l}_{ij}$. Given a dictionary $\mathcal{W}$, the feature n_{ijk} is the number of times that type $w_k \in \mathcal{W}$ occurs in sentence s_{ij}. Because most types occur only infrequently in quotations, we restrict attention to the 1000 most frequent.

The vector $\mathbf{l}_{ij}$ consists of binary variables that indicate where in a document a sentence occurs. For example,

$$l_{ij1} = \begin{cases} 1 & \text{if } s_{ij} \text{ is in the first paragraph} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Other location variables are used to indicate whether a sentence occurs in the final paragraph, whether it is the first sentence in a paragraph, and whether it is the last sentence in its paragraph. Interaction terms between location variables are also included.

Using these features as predictors, we fit statistical models to estimate

$$\widehat{\Pr}(y_{ij} = 1|\mathbf{l}_{ij}, \mathbf{n}_{ij}) \tag{3}$$

Our chosen summary sentence for document d_i is the one that maximizes the above quantity. We fit these models by two different methods: Naive Bayes and regularized logistic regression.

Naive Bayes

The multinomial Naive Bayes model on a dictionary W is a familiar option for text classification, e.g. Gale et al. (1992), McCallum & Nigam (1998) When there are additional features, the Naive Bayes model has also a natural extension: We simply assume that each additional feature is independent of all the others, conditional upon y. In this case, we invert Bayes' Law by observing:

$$\frac{\Pr(y=1|\mathbf{l},\mathbf{n})}{\Pr(y=0|\mathbf{l},\mathbf{n})} = \frac{\Pr(y=1)}{\Pr(y=0)} \frac{\Pr(\mathbf{n}|y=1)}{\Pr(\mathbf{n}|y=0)} \frac{\Pr(\mathbf{l}|y=1)}{\Pr(\mathbf{l}|y=0)} \tag{4}$$

Regularized Logistic Regression

Given feature vectors $\mathbf{l}_{ij}$ and $\mathbf{n}_{ij}$, a linear logistic regression model takes the form:

$$\log \frac{\widehat{\Pr}(y_{ij}=1|\mathbf{l}_{ij},\mathbf{n}_{ij})}{\widehat{\Pr}(y_{ij}=0|\mathbf{l}_{ij},\mathbf{n}_{ij})} = \alpha_0 + \alpha^T \mathbf{l}_{ij} + \beta^T \mathbf{n}_{ij} \tag{5}$$

Most often, this model's parameters are fit by maximizing the conditional likelihood of the response given the feature values. However, this is not desirable when the number of features is too large. In order to prevent over-fitting, a regularization parameter λ is introduced. Then we have a modified maximization problem.

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha,\beta} \log lik(\alpha,\beta) - \lambda\|\beta\|^2 \tag{6}$$

Here we penalize the coefficients that are associated with type features but not the ones associated with location features. This is because type features are only rarely active, whereas location features are frequently active, so their coefficients can be estimated more accurately.

Regularized logistic regression has been used in other text classification problems, as in Zhang and Yang (2003). For further information on regularized model fitting, see for instance, Hastie, Tibshirani, and Friedman (2001).

Experimental Results

Models are fit using 25 randomly chosen sets of 2000 training documents each. Figure 5 shows their success rate at identifying the correct sentence in test documents. The model that uses only location features always selects the final sentence of the document as its summary. By incorporating word type features, model performance improves substantially.

One possible complication in viewing these results is that some review documents contain multiple statements of their overall opinion. For instance, the following sentence is predicted as a sentiment summary: "*Mulholland Drive* is rapt and beautiful and absorbing, but apart from a few scenes ...

Method	Features	Pct. Correct	Std. Error
Random	none	6.3%	–
Logist. Reg.	loc.	16.1%	0.3%
Naive Bayes	loc.; type	23.1%	0.5%
Logist. Reg.	loc.; type	25.8%	0.6%

Figure 5: Prediction match rate

it lacks the revelatory charge that *Blue Velvet* had 15 years ago." Although this does not match the Rotten Tomatoes quotation, it is otherwise a suitable choice.

If this complication has a significant influence on classification results, then we would expect the correct summary quotation to be ranked highly, even when not the first choice. However, this is not the case. In only 11.0% of test cases is the true summary ranked second among sentences in the document, and in only 5.7% is it rated third.

Summaries and Overall Sentiment

A typical document contains 15 sentences that do not overlap with the summary, and only one that does. As a result, it is much easier to estimate the frequency of a particular word type outside of summary quotations than within. This leads us to consider whether it is possible to draw upon other information to produce more accurate estimates. One such piece of information is a word's relative frequency in favorable and unfavorable reviews. If a word type occurs more often in favorable reviews, then it may convey positive sentiment. In that case, it would rarely be appropriate for the sentiment summary of an unfavorable review.

Figure 6 provides graphical evidence that such a relationship exists. Each data point represents a single word type, with attention restricted to types that appear at least 30 times in summary quotations. Each axis show the fraction of a type's occurrences in positive reviews. The y-value for a particular type is this fraction for occurrences in summary quotations; the x-value is the same fraction for other review text. The accompanying regression line has slope significantly greater than one (p-value 0.019). This enables us to conclude that when a word type appears more often in reviews of a particular sentiment, that this imbalance becomes even more pronounced in summary statements. This information could enable one to predict a type's frequency in summaries from its frequency in other text.

A simpler approach is to fit separate models for positive and negative reviews. This has a drawback, however, because it further reduces the amount of training data for each class. Perhaps for this reason, a Naive Bayes classifier built in this way achieved a success rate of only 18.3%. Thus, in order to gain improvements in performance from this insight, a more sophisticated approach is required.

Conclusion

We have introduced the problem of sentiment summarization and shown that a sentence's location and constituent words are valuable predictors of whether it will be chosen as a sentiment summary. However, when one applies standard classification techniques, these features are typically insufficient to identify the summary correctly. In seeking to improve performance, one natural step would be to incorporate more sophisticated summarization techniques. We note, however, that such techniques would be unlikely to utilize many of the word type features that we have observed to be informative. Possibly, the features identified here can be used in conjunction with text summarization to produce improved predictions.

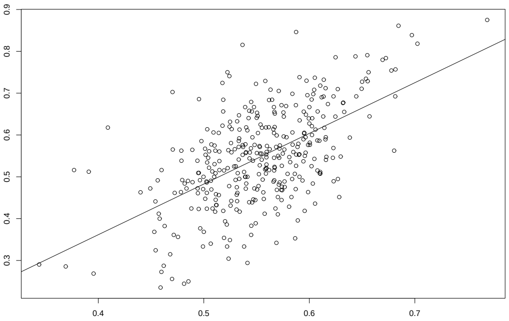

Figure 6: Percentage of Favorable Reviews for Features in Summary versus Other Text

References

Cardie, C.; Wiebe, J.; Wilson, T.; and Litman, D. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*. AAAI Press.

Gale, W. A.; Church, K. W.; and Yarowski, D. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26:415–439.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Mallett, D. 2003. Text summarization: An annotated bibliography.

McCallum, A., and Nigam, K. 1998. A comparison of learning models for naive bayes text classification. In *Working Notes of the 1998 AAAI/ICML Workshop on Learning for Text Categorization*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Porter. 1980. An algorithm for suffix stripping. *Program* 14 (3):130–137.

Zhang, J., and Yang, Y. 2003. "robustness of regularized linear classification methods in text categorization". In *Proceedings of the 26th Annual International ACM SIGIR Conference (SIGIR 2003)*.