The Impact of Time on the Accuracy of Sentiment Classifiers Created from a Web Log Corpus

Kathleen T. Durant and Michael D. Smith

School of Engineering and Applied Sciences
Harvard University
Cambridge, MA USA
{ Kathleen, Smith}@seas.harvard.edu

Abstract

We investigate the impact of time on the predictability of sentiment classification research for models created from web logs. We show that sentiment classifiers are time dependent and through a series of methodical experiments quantify the size of the dependence. In particular, we measure the accuracies of 25 different time-specific sentiment classifiers on 24 different testing timeframes. We use the Naive Bayes induction technique and the holdout validation technique using equal-sized but separate training and testing data sets. We conducted over 600 experiments and organize our results by the size of the interval (in months) between the training and testing timeframes. Our findings show a significant decrease in accuracy as this interval grows. Using a paired t-test we show classifiers trained on future data and tested on past data significantly outperform classifiers trained on past data and tested on future data. These findings are for a topic-specific corpus created from political web log posts originating from 160 different web logs. We then define concepts that classify months as exemplar, infrequent thread, frequent thread or outlier; this classification reveals knowledge on the topic's evolution and the utility of the month's data for the timeframe.

Introduction

Supervised machine learning uses statistics to build mathematical models to perform a particular performance criterion based on example data, known as training data. Once the machine learning classifier is created, it can be applied to other examples known as testing data. The basic assumption is that the training data does not vary from the testing data, since we are using the training data as exemplars of the testing data. There are many ways the training and testing data can vary from one another; the training examples could be from a different domain source, on a different topic, or from a different time period. We explore the effect of different creation dates between the training and the testing examples; we refer to this factor as the *testing-training difference*.

When the training data is from a particular time period, it may not accurately represent data from a different time period. All data have the aspect of time, but we believe people's opinions change more frequently than facts or belief systems. We run our experiments on a corpus of opinions found in web logs. We find that research performing web mining analysis should pay particular attention to the timeliness of the data they are reaping. In particular, the question we investigate is how chronologically close the training set needs to be to the testing set to achieve adequate sentiment classification results for a corpus of opinions collected from web log posts.

We apply sentiment classification to a collection of political web log posts varying the testing-training difference. Sentiment classification is the ability to judge an example as positive or negative. Previous work has shown that traditional text classification methods work well when applied to sentiment classification. However, the corpus of the previous study was not topic specific. The study did not investigate the effect of time and its results may not be applicable to the web log domain (Pang et al., 2002).

We run a rigorous study measuring time dependency. We show sentiment classification models are time-dependent by measuring the accuracy results of twenty-five different time-specific sentiment classifiers on twenty-four different testing timeframes. Our results show a significant degradation in accuracy results as the testing timeframe is distanced from the training timeframe. The creation date of the training and testing data does affect the result of a sentiment classifier.

We then consider each month's results individually in order to derive its utility for the complete timeframe. We define a classification method using the statistical results of each month to determine if its data is applicable only to itself (*outlier*), to a few months (*infrequent-thread*), to many months *(frequent- thread*) or on average to the whole timeframe (*exemplar*). We then consider the effects of the testing-training difference on the four classifications.

Studies on Machine Learning Dependencies

There have been few studies measuring accuracy dependencies in machine learning classification. Engström (2004) showed that the bag-of-features approach used by machine models is topic-dependent. Machine models trained on a particular product's reviews are unlikely to perform well on another product. Read (2005) states models are domain-dependent and time-dependent. The source of the data and the creation date of the passage impact the accuracy of sentiment classification results. He measured the accuracies of two different movie databases, one that spans a two-year time period and another collection from a previous time period whose length was not measured. He grouped his samples such that models trained and tested on the same timeframe formed one group and models trained and tested on different timeframes formed a second group. He used two different machine learning induction techniques, Naïve Bayes and Support Vector Machines, to measure the accuracy of the models on different testing datasets. He determined datasets trained and tested on the same timeframe yielded a higher result in sentiment classification than datasets that were trained from one timeframe and tested on another timeframe. In particular his best result for the same timeframe dataset was 81.5%, for the mixed timeframe result 77.5%.

Read's study paired experiments with differing amounts of testing and training data and different validation techniques. For the same timeframe experiments Read's research (2005) used three-fold cross validation as depicted in Figure 1; for the mixed timeframe experiments he used the holdout technique as depicted in Figure 2. The error in the accuracy estimate differs with different validation techniques and cross validation techniques are known to improve the generalization accuracy of the learned classifiers (Kohavi, 95). Also the size of both the training and the testing data vary substantially between the two experiment groups; training size varies from 1400 to 932 and the testing size varies from 1400 to 466. Using different amounts of testing data within the experiments introduces a bias in the comparison (Kohavi, 1995).

Experiments

Our experiments investigate the temporal relationship between the testing and training data. We use the Naïve Bayes induction technique as our classification technique and the holdout validation technique using equal-sized but separate training and testing data sets. We ran two collections of experiments, one for the maximum-sized datasets for each month and the other for the minimum-sized datasets. Each collection contains a series of 600 experiments. Figure 3 illustrates our methodology in contrast to Read's. We assign each month an ordinal number, the first month, March 2003, being labeled 1, continuing to our last month in our dataset March 2005, being labeled as 25. We group our experiments by the number of months separating our testing and training datasets. For example, results for the

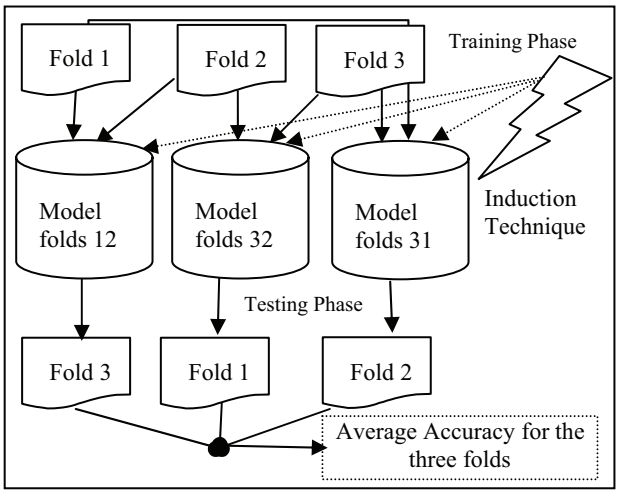

Figure 2 Three-fold cross validation technique used by Read (2005) for the same test/training timeframe samples. Each fold contains 466 data elements. The final accuracy was the mean of the accuracy for the three folds, lowering the generalization error rate.

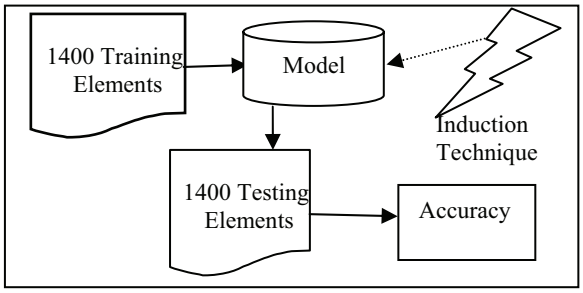

Figure 1 The validation technique used by Read (2005) in his different test/training timeframe experiments. The training and the testing data are each 1400 elements.

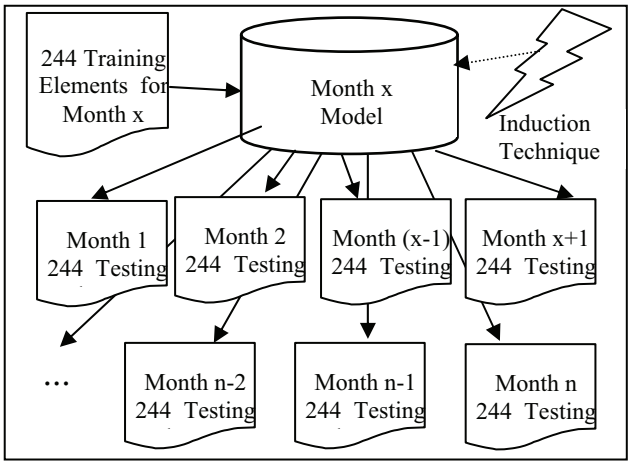

Figure 3 The size of our testing and training sets and our holdout validation technique used by our minimum experiments. Each month's data is tested on 24 different testing sets. We run these experiments for 25 different classifiers, each containing one month's worth of web log posts.

ordinal number -1 are the accuracy results of classifiers that were trained on month *n* but tested on month *n - 1*. The testing and the training datasets are always from different timeframes.

Our first collection of experiments using the maximum-sized datasets takes advantage of all data that was available from the web logs. This led to various amounts of testing and training data and may lead to unfair comparisons of models. These results are interesting because they indicate the performance of the models when the data size is determined by the interest in the given topic for the particular month.

We also ran experiments limiting the number of elements of each dataset to the minimum-sized dataset, 244 elements. This gave each experiment equal-sized testing and training sets. The contributing posts were randomly selected from the total collection of posts for the month.

The origin of our opinion corpus is Gandelman's political blog, themoderatevoice.com. It categorizes over 250 web logs as a *left-voice*, *right-voice* or *moderate-voice*. We predict Gandelman's political categorization of the originating web log, allowing postings to inherit the categorization of the web log. We collect the categorization and the web log posts from the referenced blogs, limiting our corpus to a specific topic. We classify the post's sentiment toward our topic as *right-voice* or *left-voice*. We use a subset of the terms found within the corpus as features. Previous research on this corpus (Durant and Smith, 2006) showed that the sentiment of a political web log post can be predicted on average 78.45% of the time using a Boolean presence feature set representation and boosted to 89.77% when an off-the-shelf feature selection technique is applied (Durant and Smith, 2007). We use the same feature set representation, time segmentation scheme (a month) and topic as the previous study. However our study uses the hold-out validation technique; whereas the previous study used stratified 10-fold cross validation.

Results

As demonstrated in Figure 4, when the training time period is chronologically closer to the testing time period (in the area of the graph closer to the origin), we achieve a higher accuracy than when they are farther apart. There is a slow degradation in the accuracy in the classifiers as the difference between the testing and the training data increases. There is a noticeable difference in the plot on the different sides of the origin. If it is a negative testing-training difference then the training month is from the future. Classifiers trained on future data seem to do a better job predicting past data than classifiers trained on past data do at predicting future data.

Figure 5 demonstrates the similarly shaped results for the collection of minimum-sized datasets; the difference is an average degradation in the accuracy and an increase in the variance, a vertical stretching of the plotted graph in Figure 4. A statistical comparison of the results of the two collections can be found in Table 1.

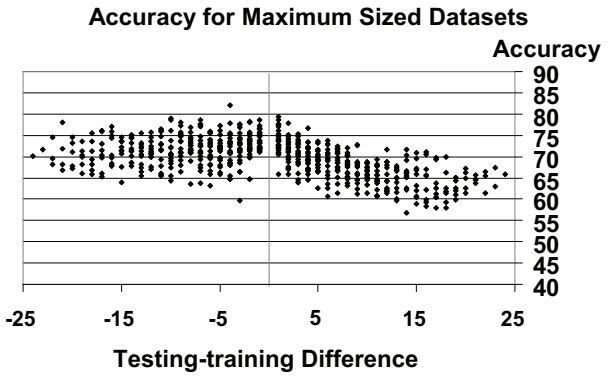

Figure 4 Accuracy results for the maximum-sized datasets. The graph shows the degradation of sentiment classifier accuracies as the training-time difference varies. Points graphed left of the origin are the results of classifiers that are tested on previous time periods. Points graphed right of the origin are results of classifiers tested on future time periods.

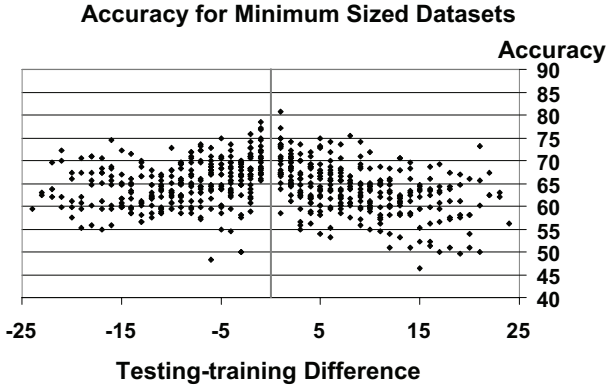

Figure 5 Accuracy results for the minimum-sized datsets. The overall shape above resembles a vertically stretched version of the graph displayed in Figure 4.

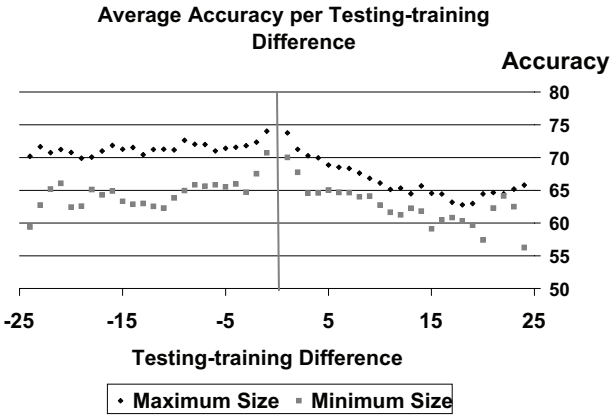

Figure 6 Average accuracy per testing-training difference for both maximum-sized and minimum-sized datasets. The shapes of the results are quite similar.

Table 1 A comparison of the average mean and average variance for the maximum and minimum datasets. The variance increases and the average accuracy decreases for the minimum dataset.

Collection	Accuracy Variance	Accuracy Average
Maximum-sized	19.157	69.708
Minimum-sized	27.872	64.550

In Figure 6 we compare the average accuracy for each testing-training difference between the minimum-sized and the maximum-sized datasets. The accuracy decreases and the variance increases for our minimum-sized collection in comparison to our maximum-sized collection. However, the average degradation between accuracy results as the testing-training difference increases remains. The slow degradation of our results as we move away from the origin displays the time dependency of our machine classifiers. In both the minimum and maximum collections, training months four months away from the testing month do a poor job representing the testing data.

Figures 4 and 5 indicate models trained with future data in general produce higher accuracy results. To quantify this observation we paired our average results by the absolute difference between the testing and training months and ran a paired t-test to determine if future models predict the past at a higher accuracy than past models do at predicting the future. The models contributing to each paired distance are the same; the collection of testing months and the direction of the movement varies. For one group the testing months are x months in the future, for the other group the testing months are x months in the past. The accuracy improvement of the future models (tested on the past) compared to the past models (tested on the future) is [2.065 to 5.441] for the maximum datasets and [.95 to 2.81] for the minimum-sized datasets at a 99.9% confidence interval. We also verified our results using the nonparametric Wilcoxon signed-ranked test. By comparing the medians of the samples it verified the superiority of the future models to the past models at a significant level of 0.1%. We believe this difference can be attributed to authors of future data having access to all historical information; whereas authors from the past do not have access to topics and events in the future.

Thread, Exemplar, and Outlier Months

In the previous section we looked at the overall effect of the training timeframe and testing timeframe on the accuracy of sentiment classifiers. We discovered on average, when the difference between the testing timeframe and the training timeframe is small; the classifier on average performs better. Now we consider the predictive power of each month individually. Our goal is two faceted. First we wish to further investigate the testing-training difference on accuracy results. We consider different statistically related subsets of months and examine the effect of the testing-training difference on each subset. Second we wish to define a method that allows us to automatically glean information about our topic and the evolution of the topic throughout the timeframe.

We do this by defining a method that classifies each month's predictive power on other months' data across the total timeframe. Our method identifies four classes of months: *exemplar, frequent- thread, infrequent-thread,* and *outlier months*. We use four different metrics to identify a month's class: *training accuracy, training variance*, *testing accuracy* and *testing variance*.

The *training* metrics select a month x as the training dataset and then test the resulting machine classifier on all other months in the timeframe. This sets the data for the chosen month as the only representation of the topic for the whole timeframe, giving a very limited description of the topic. The *training accuracy* for a particular month x is defined as the average accuracy of the machine classifier tested on all other twenty-four months. The *training variance* for month x is defined as the variance of the accuracy results. The training variance shows how well month x represents the total timeframe since the feature set for month x determines the identifying features for the complete timeframe.

The *testing* metrics select a month x as the single testing dataset and then test the machine classifiers trained on every other month in the timeframe. In this collection of classifiers there is no single representation of the topic across the timeframe; instead the representation of the topic is a collective representation of all the feature sets from all the contributing classifiers. This representation allows the topic to evolve across the timeframe; each month determines the representation of the topic for that given month. The t*esting accuracy* for a particular month x is defined as the average accuracy of all the machine classifiers tested on month x. The *testing variance* for a particular month x is defined as the accuracy variance. If a particular month has a low testing variance and a higher than average testing accuracy, the model may be beneficial for the long term.

Table 2 displays our four classes of interest and the values of the metrics that define them. The classes indicate how the metrics for a particular month compare in relation to the median of the metrics for all months. This roughly divides the months into classes based on their utility for the complete timeframe and those that would be good predictors for the topic. We measure our results to the median rather than the mean, since our data is not normally distributed. To create a soft boundary around our measurements, we use a 95% confidence interval with each metric.

The testing and training metrics are two related similarity metrics and viewed as coupled in our classification. Our classes highlight months that show either degradation or an improvement in both types of accuracies and variances. In other words, time segments that can successfully represent the total timeframe and also be successfully represented by the total timeframe.

Exemplar months contain data that is representative of the total timeframe. The overall arching theme of the time-

Table 2 Four classes of interest for months in the timeframe. Above (below) means that the month's metric was greater (less) than the median value of all months in the timeframe, within the 95% confidence interval for the median measurement.

Class	Accuracy		Variance	
	Testing	Training	Testing	Training
Exemplar	Above	Above	Below	Below
Frequent thread	Above	Above	Above	Above
Infrequent thread	Below	Below	Above	Above
Outlier	Below	Below	Below	Below

frame is highlighted in these months. An exemplar month represents the total timeframe well, and the total timeframe also represents the testing month well.

Outlier months do not have features in common with other months. These months highlight events that are short-lived sensational sub-events.

Thread months represent a subset of months well. In general, infrequent-thread months represent a small (i.e., much smaller than the number of months in the timeframe) number of months while frequent-thread months represent a large percentage of the months in the timeframe. All thread months have variance metrics larger than the medians. The accuracy metrics differentiate the infrequent-thread from the frequent-thread months.

Figure 7 displays the results for the training metrics while Figure 8 displays the results for the testing metrics for each month in the minimized-sized collection.

The months October 2003, December 2003, and July 2004 are classified as outlier months. These results are plausible since all three months contained short-lived highly sensational events. In October 2003 the Madrid conference raised funds to rebuild Iraq, and December 2003 Saddam Hussein was captured. July 2004 is the month the Australian Flood Report, the British Butler Report and the US Senate Intelligence Committee all released statements stating the pre-war intelligence exaggerated Hussein's threat.

The months March 2003, April 2003, May 2003, August 2004, September 2004 and March 2005 are classified as infrequent-thread months. March 2005 is the only thread month not adjacent to an outlier or another thread month. One thread identified was from March 2003 – May 2003, the combat months of the war and the celebratory "Mission Accomplished" speech. This beginning segment of the timeframe is very different from the rest of the timeframe in sentiment and in events. The other infrequent thread identified was August 2004 – September 2004. Only these months contain the Battle of Najaf and the death toll reaching 1000.

Our only frequent-thread month identified is October 2004; it chronologically follows the previous infrequent thread. This month contains the release of the Duelfer report that states there were no weapons of mass destruction in pre-war Iraq. Weapons of mass destruction are a frequent sub-topic over our two-year time period, November 2004 best represents this subtopic.

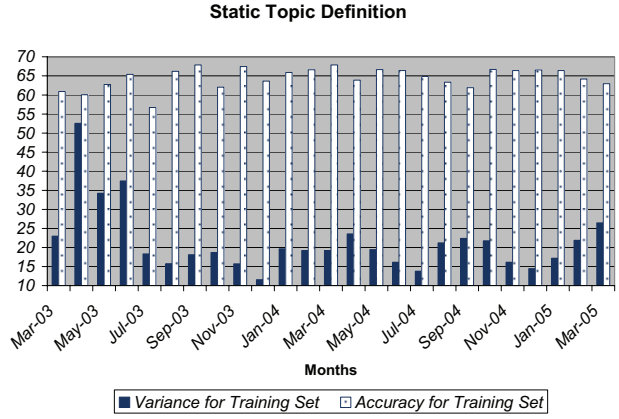

Figure 7 The training accuracy and variance for each month.

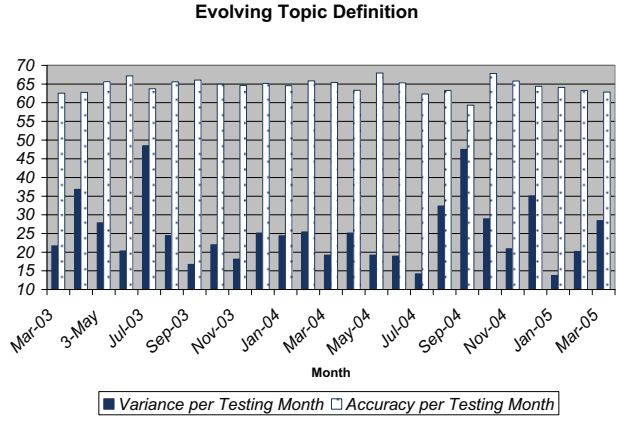

Figure 8 The testing accuracy and variance for each month.

The eight months August 2003, September 2003, November 2003, March 2004, May 2004, June 2004, November 2004 and January 2005 were identified as exemplar months for our topic.

We next investigate the impact of testing-training difference on our four categories. Our categories are quite statistically different; yet Figures 9-12 illustrate that examples from all four categories still exhibit effects of the testing-training difference. The figures display a general decline in accuracies as we move away from the origin for all our classes.

Conclusions and Future Research

Our research has explored the effects of time on sentiment classification accuracy. We showed that sentiment classification models are indeed time dependent, something not previously shown rigorously. As the duration of time between the testing and the training data increases the accuracy results decrease. We also showed there is a disparity in the testing-training difference for future and past data. Models trained on future data produce higher accuracy results than models trained on past data, at a 0.1%

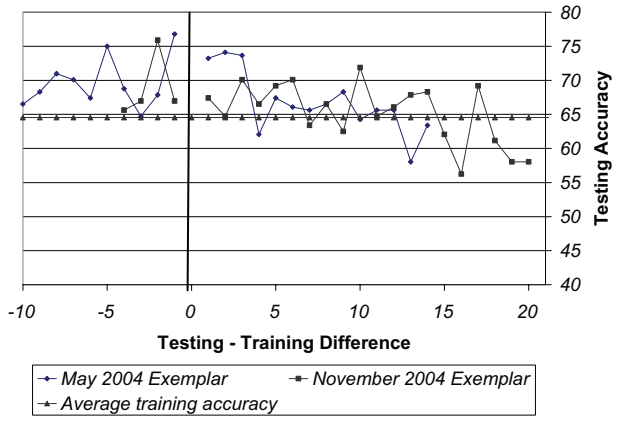

Figure 9 The testing accuracy for two exemplar months (May 2004 and November 2004). These months are successful on all months in the timeframe.

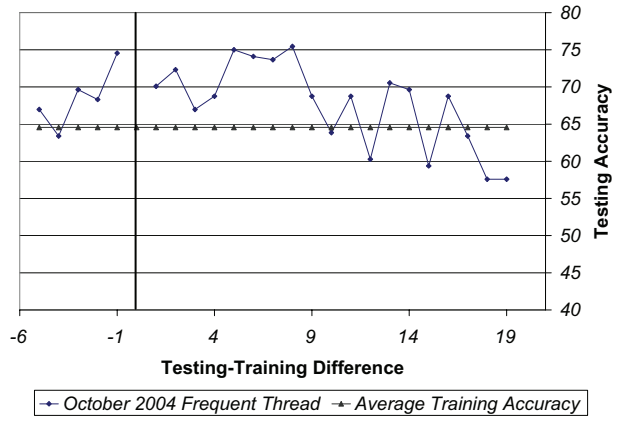

Figure 10 The testing accuracy for frequent-thread month (October 2004). It generates testing accuracies higher than the average on most months. This month is successful on a large subset of the timeframe.

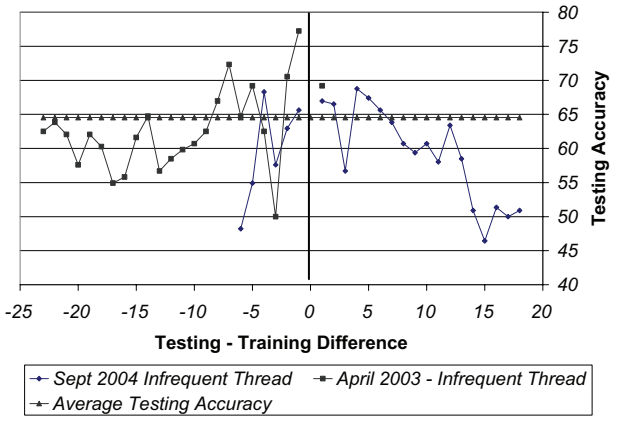

Figure 11 The testing accuracy for two infrequent thread months (September 2004 and April 2003). These months are successful on a few months chronologically close to each testing month.

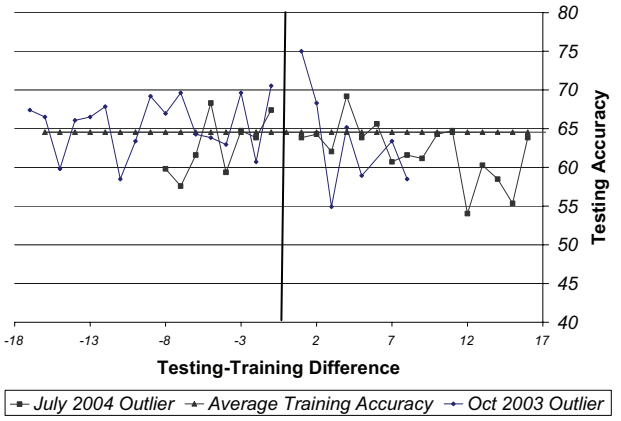

Figure 12 The testing accuracy for two outlier months (July 2004 and December 2003). These months, on average, do a poor job predicting the other months in the timeframe.

significance α-level; meaning our finding has a 99.9% chance of being true. Our finding is sensible since a timeframe in the future is aware of all potential discussion points of the past; however the past is not aware of all potential discussion points in the future.

We have also defined testing accuracy, training accuracy, testing variance and training variance; concepts that data. With these definitions we were able to identify outlier allow us to label, at a gross level, the utility of a specific time segment's data for predicting other time segment's months that contained short-lived sensational events, two reasonable threads and eight exemplar months. We showed even though these categories are statistically different all categories are affected by the testing-training difference.

Our research uses the total two-year timeframe to identify outlier, thread and exemplar months. We would like to analyze the impact of the timeframe length on the assigned classes. We would like to test our classes by comparing the testing accuracy of future time segments on models created from our exemplar, thread and outlier classes.

In our month classification system we only considered months where both accuracies and both variances are either above or below the median. However, there may be interesting aspects in the time segments when the two accuracies and two variances diverge. We would like to extend our classification system to these other permutations of accuracies and variances.

At its current definition our month classification method is most useful in analyzing a collection of topic-specific data sources and identifying the months containing unordinary events, typical events and repeating sub-topics. We would like to add a real-time aspect to our system so that months can be classified while they are happening.

The classification technique's success in identifying thread and exemplar months is limited. Our method does not indicate a specific subset of months to which the thread is useful. We do not consider cyclical threads, threads that may repeat within the timeframe. We need more sophisticated methods and precise metrics to identify all threads and their duration. We would like to extend our system to address all of these issues. We would also like to be able to

automatically determine the most influential time segmentation of data given a collection of topic-specific, time-ordered data. If we can capture a metric of thread sentiment, we can use it to predict existing sentiment that also varies with time, such as polling results.

Acknowledgements

This research was supported in part by a gift from the Google Corporation.

References

Beineke, Philip; Hastie, Trevor; and Vaithyanathan, Shivakumar. 2004. The Sentimental Factor: Improving Review Classification via Human-Provided Information. In *Proceedings of the Association of Computational Linguistics.* Barcelona, Spain.

Conover, W.J. 1980. *Practical Nonparametric Statistics*. Wiley Publications.

Das, Sanjiv, and Chen, Mike. 2001. Yahoo! for Amazon: Extracting Marketing Sentiment from Stock Message Boards. *Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA 2001).*

Dave, K.; Lawrence, S.; and Pennock, D.. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *Proceedings of the Twelfth International World Wide Conferences (WWW'03).*

Dietterich, T. G. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation, 10 (7),* pp. 1895-1924.

Durant, Kathleen T., and Smith, Michael D. 2006. Mining Sentiment Classification from Political Web Logs. *In Proc. of WebKDD 2006: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), August 20-23 2006, Philadelphia, PA.*.

Durant, Kathleen T., and Smith, Michael D. 2007. Predicting the Political Sentiment of Web Log Posts using Supervised Machine Learning Techniques Coupled with Feature Selection. To appear in *Advances in Web Mining and Web Usage Analysis 2006 selected revised papers from the 8th International Workshop on Web Mining and Web Usage Analysis,* O. Nasraoui, M. Spiliopoulou, L. Srivastava, B. Mobasher, B. Masand, Eds. Springer Lecture Notes in Artificial Intelligence, 2007.

Engström, Charlotta. 2004. Topic Dependence in Sentiment Classification. Master's Thesis. St. Edmund's College, University of Cambridge.

Hatzivassiloglou, Vasileios, and McKeown, Kathleen. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of the ACL-EACL'97 Joint Conference: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics,* pp. 174-181.

Hu, Minqing, and Liu, Bing. 2004. Mining and Summarizing Customer Reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining KDD 2004.*

Huettner, Allison, and Subasic, Pero. 2000. Fuzzy typing for Document Management. *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pp. 26-27.

Kohavi, R.. 1995. A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* pp. 1137-1143.

Mozer, Michael C.; Dodier, Robert; Guerra, Cesar; Wolniewicz, Richard; Yan, Lian; Colagrosso, Michael; and Grimes, David. 2000. Prediction and Classification: Pittfalls for the Unwary. University of Colorado Technical Report.

Nasukawa, Tetsuya, and Yi, Jeonghee. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture*, pp. 70-77.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, March, 2006.

Pang, Bo; Lee, Lillian; and Vaithyanathan, Shivakumar. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79-86.

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

Read, Jonathon. 2005. Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL Student Research Workshop*, Ann Arbor, MI, USA.

Sack, Warren. 1994. On the Computation of Point of View. *Proceedings of the Twelfth American Association of Artificial Intelligence (AAAI),* page 1488. Student Abstract.

Tong, Richard, M. 2001. An Operational System for Detecting and Tracking Opinions in On-line Discussion. *Workshop note, SIGIR 2001 Workshop on Operational Text Classification.*

Witten, Ian, H. and Frank, Eibe. 2000. Data Mining Practical Learning Tools and Techniques with Java Implementation. Academic Press, San Diego, CA.