Delta TFIDF: An Improved Feature Space for Sentiment Analysis

Justin Martineau, and Tim Finin
University of Maryland, Baltimore County 1000 Hilltop Circle, Baltimore, MD 21250 410-455-1000 extension 6338
{jm1, finin1}@cs.umbc.edu

Abstract

Mining opinions and sentiment from social networking sites is a popular application for social media systems. Common approaches use a machine learning system with a bag of words feature set. We present Delta TFIDF, an intuitive general purpose technique to efficiently weight word scores before classification. Delta TFIDF is easy to compute, implement, and understand. We use Support Vector Machines to show that Delta TFIDF significantly improves accuracy for sentiment analysis problems using three well known data sets.

Introduction

By gathering and automatically determining an author's feelings based on the text they've written, we can solve multiple problems in both the public and private sectors. Governmental usage of textual sentiment analysis in blogs can help identify potential suicide victims and terrorists. Textual sentiment analysis can also provide business intelligence for market research, financial investments, and politics.

To capitalize upon these opportunities we must mitigate blog noise. Blogs are often informally written, poorly structured, rife with spelling and grammatical errors, and contain factually incorrect or contradictory information. This makes techniques like parsing, simple pattern matching, complex grammars, and knowledge reasoning using the semantic web difficult to apply.

Due to Joachims' success dealing with these problems using SVMs in the bag of words vector space model, many researchers have adopted a similar approach. In (Joachims 1997) each dimension measures the count of a specific word or ngram word pair. Alternatively, words can be counted as booleans as shown in (Pang, Lee, & Vaithyanathan 2002) and (Whitelaw, Garg, & Argamon 2005), or weighted by their IDF score like (Kim *et al.* 2006). We introduce a novel way to weigh words using the difference of their TFIDF scores in the positive and negative training corpora and show how this improves accuracy.

Related Work

Determining movie, book, and product review sentiment is a well studied problem. We provide a brief summary of similar approaches that we later compare our results against.

Support Vector Machines using bag of words feature sets provide a strong baseline accuracy of 82.7% (Pang, Lee, & Vaithyanathan 2002) for movie reviews. SVMs are an appropriate tool because they are resistant to blog noise, can handle large feature sets consisting of bag of unigram and bigram words feature sets, and are traditionally good at similar tasks like topic based classification. (Joachims 1997) Techniques like augmenting the training sets with human supplied annotator rationales (Zaidan, Eisner, & Piatko 2007), using appraisal groups (Whitelaw, Garg, & Argamon 2005), and casting the problem into a graph and using minimum cuts (Pang & Lee 2004) have pushed these results up to 90%.

Zaidan et al. used rational annotation to augment the training set with near duplicate documents by copying the raw document and removing from it the best features (as supplied by human annotators). These new support vectors cut down the margin size to improve classification accuracy.

Appraisal groups are phrases level text snippets centered on adjectives containing markup that expresses the type and strength of the appraisal. Whitelaw et al. used a semantic taxonomy of manually verified appraisal groups automatically generated from a seed set using WordNet and other similar resources in conjunction with a standard bag of words to achieve 90.1% accuracy on movie reviews.

In (Pang & Lee 2004) they trimmed out objective content from movie reviews and used an SVM bag of words classifier to determine review polarity. To determine subjective sentences, they cast the task as a graph problem, and used the minimum cut between the subjective node and the objective node to form a classifier. First, they created an SVM subjectivity classifier and trained it with objective and subjective sentences from a different set of movie reviews. Then they broke reviews into sentences and inserted them into the graph as nodes. They also inserted a positive node and a negative node. Next, they weighted the edges between sentence nodes and the positive and negative pole nodes using the distance of those sentences from the margin of their subjectivity classifier. Finally, they assigned scores to edges between sentences by their proximity within the review. Second, they used the minimum cut on this graph to remove the objective content from their reviews. Third, they trained and tested another SVM bag of words classifier on their trimmed reviews.

Approach

In a bag of words each word or ngram word pair is associated with a value. These values are commonly their word count in the document. Sometimes these values are further weighted by metrics measuring how rare these terms are in other documents. Instead, we weight these values by how biased they are to one corpus.

We assign feature values for a document by calculating the difference of that word's TFIDF scores in the positive and negative training corpora. Given that:

1. $C_{t,d}$ is the number of times term t occurs in document d
2. P_t is the number of documents in the positively labeled training set with term t
3. $|P|$ is the number of documents in the positively labeled training set.
4. N_t is the number of documents in the negatively labeled training set with term t
5. $|N|$ is the number of documents in the negatively labeled training set.
6. $V_{t,d}$ is the feature value for term t in document d.

Since our training sets are balanced:

$$\begin{aligned} V_{t,d} &= C_{t,d} * \log_2\left(\frac{|P|}{P_t}\right) - C_{t,d} * \log_2\left(\frac{|N|}{N_t}\right) \\ &= C_{t,d} * \log_2\left(\frac{|P|}{P_t}\frac{N_t}{|N|}\right) \\ &= C_{t,d} * \log_2\left(\frac{N_t}{P_t}\right) \end{aligned}$$

Our term frequency transformation boosts the importance of words that are unevenly distributed between the positive and negative classes and discounts evenly distributed words. This better represents their true importance within the document for sentiment classification. The value of an evenly distributed feature is zero. The more uneven the distribution the more important a feature should be. Features that are more prominent in the negative training set than the positive training set will have a positive score, and features that are more prominent in the positive training set than the negative training set will have a negative score. This makes a clean linear division between positive sentiment features and negative sentiment features.

Consider the example in Table 1. Delta TFIDF's top scoring features are clearly more sentimental than either TFIDF or plain term frequencies. TFIDF's top scoring features appear to be the topics of the review. The top raw terms are dominated by stop words. Delta TFIDF places a much greater weight on sentimental words than either of the alternatives.

Evaluation

We test our approach on Pang and Lee's movie review, subjectivity, and congressional debates transcripts data-sets. We compare our results against the standard bag of unigram and bigram words representation using 10 fold cross validation and two tailed t-tests to prove a statistically significance improvement in classification accuracy.

Delta TFIDF	TFIDF	Raw Term Count
, city	angels	,
cage is	angels is	the
mediocrity	, city	.
criticized	of angels	to
exhilarating	maggie ,	of
well worth	city of	a
out well	maggie	and
should know	angel who	is
really enjoyed	movie goers	that
maggie ,	cage is	it
it's nice	seth ,	who
is beautifully	goers	in
wonderfully	angels ,	more
of angels	us with	you
underneath the	city	but

Table 1: The top 15 features for a positive movie review of the City of Angels.

Movie Review Data	10 Fold CV Acc	Variance
SVM DeltaTFIDF	**88.1%**	17.88
SVM Term Count Baseline	84.65%	3.94
SVM TFIDF baseline	82.85	9.17
Mincuts with subjectivity detection	87.2%	Unknown

Table 2: Sentiment polarity classification on full text movie reviews: Documents are labeled as positive sentiment or negative sentiment.

We ran our own baseline to control how the words were parsed, counted, and stop worded between different experiments and to ensure experimental uniformity and validity. We represented documents as sets of both single words, and ordered word pairs. We removed any word that did not occur in at least two documents from the feature set, but did not remove stop words. All our tests used svm_perf with a linear kernel as described in (Joachims 1999).

We used the linear kernel because it was fast, so we could compare our results with other researchers, because linear kernels yield higher accuracy in (Leopold & Kindermann 2002) for most variations on the bag of words feature sets, and because we deem sentiment classification to be a linearly separable problem. We did not stem or lemmatize words because (Leopold & Kindermann 2002) shows that these expensive steps are detrimental to accuracy.

Movie review sentiment classification

For the full text movie reviews in Table 2 Delta TFIDF outperforms the baseline with a statistical significance of 95% on a two tailed t-test. Our results are higher than the data-set's creators using their more complex minimum cuts approach. Their approach requires an additional trained SVM subjectivity classifier which requires even more labeled data.

Subjectivity detection in movie reviews

If our subjectivity detector is more accurate than their subjectivity detector then using our subjectivity detector should improve their movie review results. Using their subjectivity data-set we created our own subjectivity detector and a

Subjectivity	10 Fold CV Acc	Variance
SVM Difference of TFIDFs	**91.26%**	.47
SVM Term Count Baseline	89.4%	.74

Table 3: Sentence level subjectivity detection in movie reviews: Sentences are labeled as objective or subjective.

Congressional Debates	10 Fold CV Acc	Variance
SVM Delta TFIDF	**72.47%**	13.84
SVM Term Count Baseline	66.84%	7.36

Table 4: Congressional Debate Transcripts: Speech segments are labeled by supporting if the congressman voted for that bill.

baseline subjectivity detector matching their approach. Table 3 shows that our transformation yields a clear improvement with a 99.9% confidence interval over the baseline bag of words. Consequently, we can improve the results of the minimum cuts approach by using Delta TFIDF.

This test proves that Delta TFIDF works on both subjectivity detection and sentiment polarity classification, as well as documents of varying sizes.

Congressional approval for bills

Our third experiment involves determining if a congressman's speeches support the bill they are discussing. This test is designed to measure how well our term frequency transformation generalizes to other domains, therefore we did not use party affiliation information. Nor did we join together speech segments from the same people. Our baseline results support those show in (Thomas, Pang, & Lee 2006) for speech segment only SVM classification on their test set. Their results on their development set, which is presumably tainted by development on it, were higher. Our improved feature set is clearly better at classifying a segment of text as supporting the bill than the baseline with a 99.9% confidence interval.

Our results are higher than their test set results when they use their speaker agreement links, (although they were significantly lower than their results using their development set). To create these speaker agreement links they did manual co-refference resolution on the named entities in the text. Our approach does not require this extra annotation step.

Discussion

Delta TFIDF produces significantly better results than flat term frequencies and TFIDF weights. The TFIDF measure boosts the value of very frequent terms in the document that occur in very few other documents. Since our data-sets are composed of sentimental documents, sentimental words like "love", "hate", "good", "bad", "great", and "terrible", tend to be used in a large number of these documents giving poor IDF scores. Additionally, these words tend to have very low frequency counts in any given document because authors spice up their reviews using synonyms to avoid boring their readers, resulting in low TF scores. In practice many sentiment words are generic and tend to have low TFIDF scores.

Terms in a document should have a greater weight if they occur more often in that text, and if they are comparatively

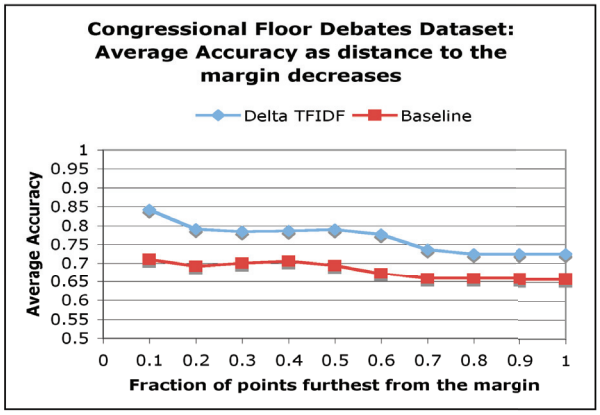

Figure 1: Uses congressional floor debate transcripts.

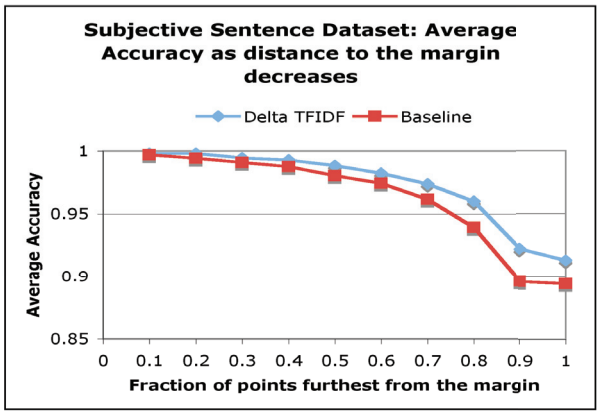

Figure 2: Uses subjective sentence data-set.

rare in oppositely label documents. Our feature weighting scheme does this by weighting that feature's term count by the log of the ratio of positive and negative training documents using this term.

Distance to Margin Implies Confidence

An SVM can provide the distance of a test point from the margin. A good classifier should have higher classification accuracy for points that are farther from the margin. The graphs in Figures 1, 3, and 2 show the running average accuracy of our judgments as data points get closer to the margin.

The curve in Figure 1 shows that our term frequency transformation is better than using the raw counts. In this case, distance to the margin is a weak estimator of confidence. Even the tenth farthest points from the margin don't have very high accuracy.

Figure 2 shows that Delta TFIDF's judgments on the furthest 20% of points from the margin are 99.8% accurate for subjectivity classification. The gradual falloff shows that the distance from the margin acts as a very strong indicator of confidence, and that there are relatively few hard to classify but easy to identify data points. Most of our performance gain comes from an increased accuracy with challenging data points implying a much sharper margin than the

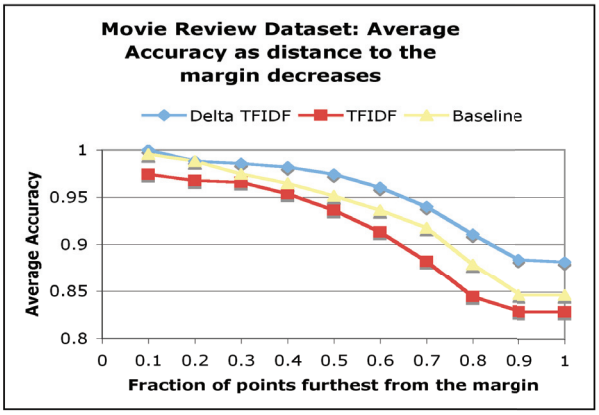

Figure 3: For movie review data Delta TFIDF's advantage over the baseline grows as points get closer to the margin. TFIDF consistently underperforms the baseline.

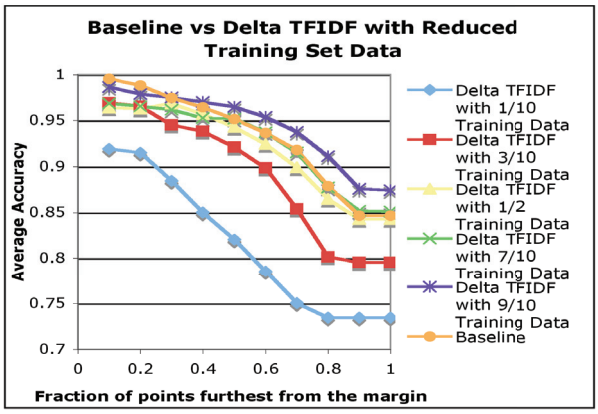

Figure 4: Delta TFIDF using half the training data achieves comparable results to the baseline for full text movie reviews.

baseline.

Figure 3 shows that Delta TFIDF's judgments on the furthest 10% of points from the margin are 100% accurate for movie reviews. The curve drops off earlier than the subjectivity curve and dips lower indicating that there are a greater number of hard to classify points which are harder to identify. About two thirds of our performance gain comes from an increased accuracy with data points that are a moderate to high distance from the margin, and the rest comes from better accuracy with very close data points. This indicates that our transformation not only yields a sharper margin, it allows better spreads points away from the margin based upon how well they represent their sentiment.

Advantages with limited data

Since sentiment is highly domain dependent each problem requires a hand annotated data-set resulting in small training set sizes and exacerbating accuracy issues. Figure 4 shows that our approach yields comparable results to the baseline approach using only half the data.

Future Work

Our research raises three key questions. How well does our term frequency transformation work with existing bag of words based sentiment analysis techniques such as earlier work on applying graph based minimum cuts, using linguistic appraisal groups, and creating rational annotations? How well does our technique generalize to non-sentiment based classification tasks? Given that redundancy is more effective than IDF weights (Leopold & Kindermann 2002), how can we improve our technique using redundancy? We expect Delta TFIDF to work well with existing techniques and generalize to other textual classification tasks. In the future we plan to test this hypothesis and work on improving accuracy using redundancy.

Conclusion

Delta TFIDF statistically outperforms raw term counts and TFIDF feature weights for documents of all sizes for subjectivity detection, sentiment polarity classification, and detecting congressional support for bills. Delta TFIDF is the first feature weighting scheme to identify and boost the importance of discriminative terms using the observed uneven distribution of features between the two classes before classification. This transformation should work with character level n-grams, on other domains, on other languages, and with any of the techniques we've compared ourselves to.

References

Joachims, T. 1997. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Springer.

Joachims, T. 1999. *Making large-scale support vector machine learning practical*. MIT Press Cambridge, MA, USA.

Kim, S.; Pantel, P.; Chklovski, T.; and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *Proc. of EMNLP*, 423–430.

Leopold, E., and Kindermann, J. 2002. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning* 46(1):423–444.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proc. of the ACL* 271–278.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of EMNLP 2002*.

Thomas, M.; Pang, B.; and Lee, L. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proc. of EMNLP*, 327–335.

Whitelaw, C.; Garg, N.; and Argamon, S. 2005. Using appraisal groups for sentiment analysis. *Proc. of the 14th ACM international conf. on Information and knowledge management* 625–631.

Zaidan, O.; Eisner, J.; and Piatko, C. 2007. Using Annotator Rationales to Improve Machine Learning for Text Categorization. *Proc. of NAACL HLT* 260–267.