Improving Binary Classification on Text Problems using Differential Word Features

Justin Martineau, Tim Finin, Anupam Joshi and Shamit Patel
University of Maryland, Baltimore County, Baltimore, MD 21250
{jm1, finin, joshi, shamitp1}@umbc.edu

18 August 2009

ABSTRACT

We describe an efficient technique to weigh word-based features in binary classification tasks and show that it significantly improves classification accuracy on a range of problems. The most common text classification approach uses a document's ngrams (words and short phrases) as its features and assigns feature values equal to their frequency or TFIDF score relative to the training corpus. Our approach uses values computed as the product of an ngram's document frequency and the difference of its inverse document frequencies in the positive and negative training sets. While this technique is remarkably easy to implement, it gives a statistically significant improvement over the standard bag-of-words approaches using support vector machines on a range of classification tasks. Our results show that our technique is robust and broadly applicable. We provide an analysis of why the approach works and how it can generalize to other domains and problems.

This is a preprint of a short (poster) paper to appear in the Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, 2-6 November 2009.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation*; I.5.4 [**Pattern Recognition**]: Applications—*text processing*

Keywords

Text classification, support vector machine, SVM, sentiment.

1. INTRODUCTION

With the vast amounts of content being authored on the Web, text analysis problems such as sentiment search have become increasingly popular. Sentiment search involves finding documents that express sentiment about a given topic.

Sentiment Feature	X^{th} **Lowest IDF Score**	**Subjectivity Feature**	X^{th} **Lowest IDF Score**
like	41	you	62
good	83	love	65
best	149	like	67
bad	171	good	99
great	198	very	129
better	203	i	137
love	221	us	179
funny	315	we	185
interesting	331	funny	197
the best	343	your	210
more than	348	bad	216
a good	357	better	246
original	390	fun	324
fun	394	beautiful	328
pretty	424	the best	369
unfortunately	506	our	374
a great	512	entertaining	392
humor	624	want	394
entertaining	710	humor	404
obvious	728	interesting	410
perfect	730	evil	416
beautiful	839	feels	417
boring	862	feel	428
obviously	926	emotional	446
worse	928	you're	495
interest	930	sense of	498
enjoy	956	seem	522
stupid	992	pretty	543

Table 1: The common technique of weighing word features with their IDF scores performs poorly for classifying movie reviews for subjectivity and sentiment. This table shows surprising examples of low scoring baseline IDF features from over 300,000 features in two movie review datasets.

CIKM '09 Hong Kong, China

This type of search is in great demand in both the public and private sectors. Governmental use 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.

Many researchers have adopted the vector space model and the bag-of-words machine learning approach. Joachims [2] demonstrated that support vector machines (SVMs) with a bag-of-words vector space are resistant to noise such as spelling and grammatical errors when determining document topic. Pang [7] demonstrated that the same general technique provides a strong baseline accuracy of 82.7% for sentiment classification in movie reviews. In the bag-of-words model each dimension of the vector space is typically weighted by the count of a specific word or ngram word pair. Later researchers have produced many variations on this basic scheme in an attempt to improve classification accuracy further. Common variants include counting the Boolean presence of words, or weighting the numerical word counts by their inverse document frequency (IDF) scores.

With the exception of IDF, these methods value every feature occurrence equally even though not all features are equal. IDF weights, however, are a poor choice for domain specific datasets. Examining the movie review dataset reveals that obviously sentimental words like those shown in Table 1 rank in the bottom 0.3% of 300,000 features. Words that express clear value judgments are some of the lowest scoring IDF features for subjectivity detection. Similarly, personal pronouns signal an opinion, but score very low under the standard IDF approach. We show how Delta TFIDF fixes these problems and further explain why the approach described in [5] works on a broad range of classification tasks.

In analyzing political speeches, [8] exploited the argument structure found in speaker reference links to help determine how a members of congress would vote given their congressional floor speeches. The speakers' votes were used to determine ground truth class labels for the dataset. Manual annotations provide links between the various speakers.

In classifying movie reviews, [6] recognized that reviews often start with predominantly objective plot summary before expressing opinions. They trimmed out such objective content from movie reviews and used an SVM bag of words classifier to determine the sentiment polarity of the remainder of the review. In determining objective 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. To do this they constructed a graph of review sentences cast as nodes and inserted nodes representing a positive negative pole. They distance between sentence nodes and the poles was calculated as the margin of a new trained SVM subjectivity classifier from a different set of movie reviews. Then they assigned scores to edges between sentences by their proximity within the review. Next, they found the minimum cut between the positive and negative poles, and threw away the sentences on the objective side. Finally, they trained and tested another SVM bag-of-words classifier on their trimmed reviews.

However, some research take a different approach, instead of trying to classify the documents as a whole they determined sentiment about the features of a product, like a camera's picture quality or size. The data they presented in [1] allowed us to test how well our technique picks out features that humans would choose. In addition to this we choose to create labels for this dataset's reviews by summing the sentiment about a product's features.

Sentiment Classification Movie Review Data

IDF Baseline	Positive	Negative
1700s	is excellent	this mess
seems muted	. cameron	this turkey
viewer's imagination	harris)	terrible .
metal at	most powerful	worst movie
only semi-serious	, great	a stupid
and astrophysicist	very effective	dull ,
paid big	lovingly	is terrible
fiction spectaculars	characters with	lame ,
, 133	fargo ,	falls flat
latifah as	melancholy	bland and
real alien	. spielberg	anyway ?
compelling performances	is terrific	degenerates
crichton science	gattaca	(scott
; has	ideals	not funny
norman's less	one which	4 on

Subjectivity Classification Movie Review Data

IDF Baseline	Subjective	Objective
apartments and	but it's	discovers
unfolding hidden	. it's	decide
personal belongings	. but	he finds
grifter	me	his father
hidden secrets	movie's	where he
and rummages	laughs	falls in
thus unfolding	the movie's	year old
breaks into	the screen	his mother
rummages their	it doesn't	) who
their apartments	. the	boyfriend
rummages	if you're	help of
grifter breaks	flick	the help
a grifter	it does	government
the all	entertaining	her to
ultimately provides	. this	discovers that

Table 2: Top 15 highest scoring features for movie review datasets for both sentiment detection and sentiment polarity classification. Top IDF scoring words versus the top class-specific features found using the Delta IDF technique.

2. APPROACH

In a bag-of-words feature model, each term (i.e., word or phrase) is assigned a numeric value. Choosing the best way to compute this value can be crucial to obtaining good performance. A term's value is often just its frequency in the document. Sometimes these values are further weighted by metrics measuring how rare the terms are in the documents in the corpus. Our approach treats the positive and negative training points as two different corpora and weights term counts by how biased they are to one corpus using the difference of their TFIDF scores in the two corpora. To avoid problems caused by words outside of the training set that occur in the test set, we treat each document as if it is a member of both classes when we calculate its feature values. This also prevents potential errors caused by dividing by zero if a feature exists in only one corpus.

Given the following definitions:

1. $C_{t,d}$ is the frequency of term t 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

we can simplify the formula for feature values if we can assume that the training set is balanced, i.e. has approximately the same number of positive and negative examples.

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

The first equation assigns evenly divided features zero weight, but prefers words that are increasingly unevenly distributed between the positive and negative classes using inverse document frequency (IDF) values. Prominent or high IDF scoring features in a given class are rarer in that class, their presence in a document indicates that the document does not belong to that class.

Features that are more common in the negative training set than the positive one receive negative scores, perfectly balanced features receive a zero, and predominantly positive features receive positive scores. Regular TFIDF lacks this capability, which allow us to display the top positive and negative features to verify how effective our technique will be for a domain. As Tables 2 and 3 show, the best IDF scoring words for each domain are much less useful than the class specific features determined by our technique.

Delta TFIDF is very accurate for determining positive and negative sentiment words in movie reviews. Not only are the top scoring positive and negative features clearly more sentimental than the features valued by IDF, they are also correctly oriented. Most of our top features are either obvious complements, insults, sentimentally expressive words, or sentimental phrases. Mentions of very popular films, such as seven-time Academy Award winner *Fargo*, correlate with positive sentiment, while mentions of unpopular films are, no surprisingly, just rare. The rest of the top 1000 positive and negative features using Delta TFIDF are just as intuitive and powerful.

Delta IDF is also very effective for subjectivity detection. Many objective features identified are story related because the reviewer must summarize the plot, this involves talking about how the main character discovers something about his past, or falls in love with some other character, or where the main character receives the help of other characters and defeats a villain. Subjective features such as *"entertaining"* and *"laughs"* express a clear value judgment. Other top subjective features indicate a change of expectation such as *"but"*, or prime the reader for a value judgment with references to the author, the reader, and generic mentions like *"the movie's"*.

Expressing political support is complex. While obvious features like *"looking forward"* and *"not oppose"* exist, many

Congressional Debates Transcripts

IDF Baseline	**Support Bill**	**Oppose Bill**
one program	look forward	no child
their optimum	competition	less likely
between mentoring	. hastings	is supposed
developed .	order against	proponents
after school	representation act	struggling
15 hours	in representation	african
preschoolers from	july	proponents of
start graduates	property rights	votes for
nor would	not oppose	to recruit
optimum	elections to	recruit
offer mentoring	divided and	not discriminate
even become	to working	to amend
themselves some	general debate	rights protections
family services	him to	separation
a qualified	sponsor of	separation of

Enron Email Spam Classification

IDF Baseline	**Spam**	**Not Spam**
milind	meds	enron
name)	viagra	hpl
use vacation	paliourg	daren
x 39247	pain	=- forwarded
mountains	php	forwarded by
of bummed	. php	/ ect
like july	cialis	hou /
have transcended	drugs	/ hou
the keyboard	in compliance	/ enron
patil	spam	@ enron
as dave	biz	@ ect
they miss	xp	: subject
plateau	sex	ect @
39247	. biz	meter
transcended	dealer	ect cc

Sentiment Analysis on Products

IDF Baseline	**Positive**	**Negative**
company does	is easy	Symantec
fried	very easy	busy
they asked	solid	you pay
repaing	a camera	n't play
stopped supporting	camera for	neither
they stopped	is really	refund
os .	I like	to contact
happy man	great camera	mistake
recharged ,	very pleased	to avoid
sold for	her	of junk
mac .	i like	not buy
the damned	are easy	a refund
within months	beautiful	freezing
soundblaster	megapixel	of Norton
damned	6610	Security 2004

Table 3: Examining the 15 highest scoring features for three different binary classification datasets shows that the Delta TFIDF technique is better at identifying useful features when compared to the IDF baseline.

features are more complicated. For example, *"as amended"* indicates support because it shows that both sides have had a chance to compromise, or at least buy votes with pork, and come to an agreement. Given the nature of politics mentioning inflammatory issues pertaining to race, religion, discrimination, sex, and party affiliation is a quick way to close down real debate and compromise. Talking about party affiliation is a sure sign that partisan politics are in play. Even mentions of partisanship and bipartisanship are toxic, features such as *"party-line vote"*, and *"bipartisan spirit"* were predominantly used by opponents of the bill. If you have to talk about a bipartisan spirit you certainly don't have it. These types of features show up in the top 1000 out of over 300000 features ranked by Delta TFIDF. Furthermore, these features are ranked much higher by Delta TFIDF than by the IDF baseline.

Spam classification features as shown in Table 3 are easy to understand. Spammers advertise medications and products. Our top 1000 spam features include a long list of investment related terms, pain relief related terms, and terms relating to deals or free stuff. Top not spam features include terms related to Enron and HPL, which was acquired by Enron. Real business communications frequently involve forwarding messages, communicating with shared coworkers, and attaching documents, many of which are spreadsheets. The 27^{th} highest not-spam feature is *".xls"*. Popular names for the current generation of workers also feature prominently in emails that are not spam.

Many of the top scoring Delta IDF features are dominated by product sentiment imbalance in the training set. People love their digital cameras, but hate their anti-virus software: camera reviews are positive by at least a nine to one ratio while reviews for anti-virus software are three to one negative. Features like *"very easy"*, *"mistake"*, *"to avoid"*, and *"a refund"* are present in greater numbers when using Delta TFIDF than when using regular IDF weights.

Our technique finds features that correspond to human judgments. We choose to evaluate our approach on Liu's product data [1] because this dataset was annotated with sentiment polarity scores for product-specific attributes such as size and picture quality. Our top discovered positive features for camera size include *"so small"*, *"is small"*, *"very small"*, *"small"*, *"small ,"*, *"small and"*, and *"a small"*. As you would expect for cameras, none of the top 1000 negative features contained the word small. Top positive features for the camera picture attribute include *"quality pictures"* and *"great pictures"* while the top negative features include *"picture after"* and *"no picture"*.

The difference of the IDF scores must be multiplied by the number of occurrences for that feature to produce its Delta TFIDF score. The movie review used in Table 4 shows that Delta TFIDF's top scoring features are clearly more sentimental than either TFIDF or plain term frequencies when used to represent a document. TFIDF's top scoring features appear to be the topics of the review. The top raw terms are dominated by stop words. In this example Delta TFIDF places a much greater weight on sentimental words than either of the alternatives.

Delta TFIDF	TFIDF	Term Frequency
, 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 4: The three feature-value metrics (Delta TFIDF, TFIDF and raw frequency), emphasize different features. Compare the 15 highest ranked features for a positive review of the film *City of Angels*. Delta TFIDF has promoted features that evidence positive sentiment.

3. EVALUATION

Tables 2 and 3 show evidence for the discriminative power of our top features. Using these weighted features to represent data points provides a statistically significant improvement to state-of-the-art machine classification accuracy.

Our evaluation uses several datasets including Pang and Lee's movie review, subjectivity, and congressional debates transcripts data-sets, along with the Enron email spam corpus, and Liu's product review dataset. By using a variety of datasets, labeled for multiple different classification tasks, and with data points ranging from sentences to full documents we show that our technique is robust and versatile. We compare our method against a baseline bag of unigram and bigram words using 10-fold cross validation and paired two tailed t-tests to prove statistical significance.

We ran our own baseline to ensure experimental uniformity and validity. Our feature sets included both single words, and bigrams (i.e., ordered word pairs). We removed words that occurred in only one document from the feature set and retained stop words. All our tests used svm_perf with a linear kernel [3]. 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 [4] for most variations on the bag-of-words feature sets, and because we deem our problems to be a linearly separable. We did not stem or lemmatize words because [4] shows that these expensive steps are detrimental to accuracy.

Table 5 shows that Delta TFIDF outperforms the raw bag-of-words baseline for each dataset. Our movie review sentiment classification results are higher than the dataset's creators using their more complex and computationally more expensive minimum cuts approach. They use an additional trained sentence level SVM subjectivity classifier, which requires an additional set of subjectivity labeled sentences. The subjectivity entry in Table 5 shows that when Delta TFIDF is used on their subjectivity dataset it outperforms the type of subjectivity classifier they used in [6] with a P value of .000106. We conclude that using Delta TFIDF will even further improve their movie review results. These two datasets prove that Delta TFIDF works on both subjectivity detection and documents of varying sizes.

Our technique also works for other kinds of sentimental datasets such as product reviews. We used Liu's data for

Movie Review Data	10-fold Acc	Variance
SVM DeltaTFIDF	**88.1%**	17.88
SVM Term Count Baseline	84.65%	3.94
SVM TFIDF baseline	82.85	9.17
Mincuts + subj. detection	87.2%	Unknown
Subjectivity		
SVM Difference of TFIDFs	**91.26%**	.47
SVM Term Count Baseline	89.4%	.74
Product Reviews		
SVM Delta TFIDF	81.41%	.00306
SVM Term Count Baseline	79.242%	.00205
Congressional Debates		
SVM Delta TFIDF	**72.47%**	13.84
SVM Term Count Baseline	66.84%	7.36
Spam Detection		
SVM Delta TFIDF	**98.917%**	$2.5 * 10^{-5}$
SVM Term Count Baseline	96.617%	$6.8 * 10^{-5}$

Table 5: This table summarizes the accuracy of Delta TFIDF and term frequency for binary classification on our five datasets, showing the 10 fold cross validation accuracy and variance. Boldface results are significant to the 98% confidence level.

nine products [1]. Since reviews in this dataset were not annotated for overall product opinion, we labeled them for sentiment using the sum of product feature scores. While this method does not account for the importance of different product feature to the reviewer, we believe it is a good approximation for human labels. This could be one reason why the results in Table 5 fall just short of a 95% confidence interval. With a P value of .0509 we can still be reasonably confident that our results represent a modest gain.

We suspect that a cause for this weaker result is an imbalance of class labels for each product. The dataset consists of 399 positive documents and 198 negative documents, but the actual imbalance for any given product type is much worse. Canon cameras had 75 positive but only eight negative reviews – not even enough negative Canon reviews to put one in each fold! Reviewer's are also overwhelmingly positive about routers. Norton AntiVirus has 32 negative reviews to nine positive ones. This distorts the distribution of product specific features.

Our technique produces significant improvements over the baseline for other kinds of binary classification problems as well, including spam detection and predicting a congress member's vote on a bill given their comments about it. Our baseline congressional classifier matches the method described in [8] for SVM speech classification and produces equivalent results, but Delta TFIDF improves on the baseline with a P value of .000582. Our Enron email spam classifier out-performs the baseline with a P value of $1.7 * 10^{-5}$.

4. DISCUSSION AND CONCLUSIONS

Most bag-of-words approaches weight features using only a function of their occurrence count in the document. TFIDF is a notable exception where the term frequency of a feature is multiplied by its pre-computed IDF score in the corpus. However, IDF weights are a bad choice for domain specific datasets because they prefer rare features. Because the documents are all from the same domain, the most descriptive features for that domain will be much more frequent than normal. This results in some of the best features having the worst IDF scores.

This is especially noticeable for sentiment classification tasks. When detecting subjective versus objective speech for film reviews, IDF ranks the word *"funny"* and the bigram *"funny ,"* along with sentences that start with the word *"but"* among the lowest scoring features. Additionally, sentiment words tend to have very low frequency counts in any given document because authors, especially professional writers, often add linguistic variety to their reviews using synonyms, resulting in lowering TF scores. In practice many sentiment words are generic and tend to have low TFIDF scores. Delta TFIDF not only ranks these example words and other similar words as some of the most useful features, but also correctly identifies the polarity of the sentiment or subjective speech they indicate.

For support vector machines, Delta TFIDF statistically outperforms raw term counts and TFIDF feature weights for documents of all sizes for subjectivity detection, sentiment polarity classification, detecting congressional support for bills, and spam classification. Delta TFIDF provides the orientation of a term to the class label and ranks these terms by their strength. 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. We believe that this transformation will improve performance with character level ngrams, on other domains, on other languages, with any binary classification technique that uses a bag-of-words.

5. REFERENCES

[1] X. Ding, B. Liu, and P. Yu. A holistic lexicon-based approach to opinion mining. In *Proc. of the Int. Conf. on Web Search and Web Data Mining*, pp. 231–240. ACM, New York, NY, 2008.

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

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

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

[5] J. Martineau and T. Finin. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Proc. of the Third AAAI Int. Conf. on Weblogs and Social Media.* AAAI Press, May 2009.

[6] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 271–278, July 2004.

[7] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pp. 79–86, July 2002.

[8] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pp. 327–335, July 2006.