Cite as: Justin Martineau, Doreen Cheng and Tim Finin. 2013. TISA: topic independence scoring algorithm. In Proc. 9th Int. Conf. on Machine Learning and Data Mining (MLDM'13). Springer-Verlag, 555-570.

TISA: Topic Independence Scoring Algorithm

Justin Martineau[1], Doreen Cheng[1], and Tim Finin[2]

[1] Samsung Information Systems North America
[2] University of Maryland Baltimore County

Abstract. Textual analysis using machine learning is in high demand for a wide range of applications including recommender systems, business intelligence tools, and electronic personal assistants. Some of these applications need to operate over a wide and unpredictable array of topic areas, but current in-domain, domain adaptation, and multi-domain approaches cannot adequately support this need, due to their low accuracy on topic areas that they are not trained for, slow adaptation speed, or high implementation and maintenance costs.

To create a true domain-independent solution, we introduce the Topic Independence Scoring Algorithm (TISA) and demonstrate how to build a domain-independent bag-of-words model for sentiment analysis. This model is the best preforming sentiment model published on the popular 25 category Amazon product reviews dataset. The model is on average 89.6% accurate as measured on 20 held-out test topic areas. This compares very favorably with the 82.28% average accuracy of the 20 baseline in-domain models. Moreover, the TISA model is highly uniformly accurate, with a variance of 5 percentage points, which provides strong assurance that the model will be just as accurate on new topic areas. Consequently, TISAs models are truly domain independent. In other words, they require no changes or human intervention to accurately classify documents in never before seen topic areas.

1 Introduction

Text analysis techniques, such as sentiment analysis, are valuable tools for business intelligence, predicting market trends, and targeting advertisements. This technology is especially salient because written works include tweets, Facebook posts, blog posts, news articles, forum comments, or any other sample of electronic text that has become prevalent due to the grow of the web.

Textual analysis applications need to operate over a wide and unpredictable array of topic areas, often in real-time. However, current approaches are unable to reliably and accurately operate in real-time for new domains.

Text analysis on a wide array of topic areas is difficult because word meaning is context sensitive. Word sense disambiguation issues are one reason why classifiers trained for one topic area do poorly in other topic areas. The linguistic community has spent a great deal of effort trying to understand the differences between word senses by build linguistic resources such as WordNet [5], WordNet Affect [12]and Senti-WordNet [1]. Word sense disambiguation is still challenging.

Fortunately, word sense disambiguation issues can be side stepped for specific problems. Consider sentiment polarity classification, which is the binary classification task where either the author approves of, or the author disapproves of the specific topic of interest. For sentiment polarity classification knowing word meaning is irrelevant, but knowing word connotation is crucial. In the following example, "I proudly wore my new shirt to the bank." It is irrelevant whether the bank is a financial institution or a river bank because both senses of the word bank have no sentimental connotation for apparel. Thus, the word sense disambiguation problem can be simplified into a word connotation calculation. By extension to text classification: knowing the word's sense is irrelevant, but knowing it's class bias for a topic area is sufficient.

We introduce a method to determine topic independent class bias scores for words. These words can be used to build bag-of-words models that operate well in a wide area of diverse topics. Creating topic independent word scores is simple when there exists labeled data from multiple domains. Bias scores for a word can be calculated in each topic area using your machine learning algorithm of choice. A function can then be applied to these scores to determine a topic independent class bias score for the word. Intuitively, to measure topic independence, it makes sense to observe the variance of a word's class bias in multiple topic areas. We introduce our Topic Independence Scoring Algorithm as a method to calculate topic independent class bias scores from a set of existing topic area specific class bias scores.

Since our Topic Independence Scoring Algorithm uses only bias scores produced by another supporting machine learning algorithm, it has several useful properties. First, the supporting machine learning can be swapped out. Machine learning experts can use our algorithm with the most appropriate algorithm for the task at hand. Second, our algorithm works on models not training data. This is very valuable in industrial settings when the training data may be lost or inaccessible due to business reasons. Alternatively, this is useful when the expertise to tune the original algorithm may no longer be available, but the model still remains. Finally, the topic independence scoring algorithm can be evaluated against the algorithm that produced the topic area specific scores. This allows us to more effectively evaluate the value of topic independence scoring.

As a use case and for evaluation purposes we build a topic independent model for sentiment analysis that is highly accurate across 20 never before seen test topic areas. Our topic independent model is even more accurate than the supporting machine learning algorithm in the test domains using 10 fold CV. Using our algorithm, we built a domain-independent sentiment model from five product review categories in the Amazon product reviews dataset [2] and evaluated it upon 20 additional product categories. Our classifier significantly outperforms the classifiers built specifically for each of the 20 product review categories. The baseline classifiers built specifically for the 20 test domains were almost twice as likely to make an error as our domain independent model.

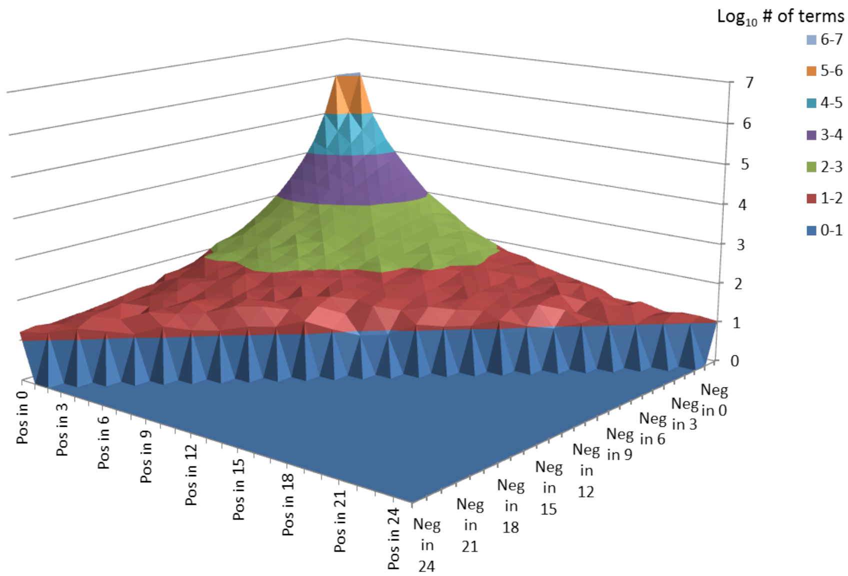

Fig. 1. Distribution of topic independence by positive vs. negative bias across 25 topic areas.

2 Understanding Topic Independence

Our approach introduces the ground breaking concept of term level topic independence, which is the degree to which a terms orientation to a class remains the same when measured across multiple topics [7]. Poly-synonymous words are one reason why classifiers trained in a single domain do poorly in other domains. However, even when the sense of the word remains the same, the usage of that word implies different things in different topic areas. These problems lack a clear mathematical definition upon which machines can compute. Our Topic Independence Scoring Algorithm provides a clearly defined mathematical counting problem to eliminate word sense disambiguation issues when doing textual machine learning. The afore mentioned counting problem counts the different orientations a term has across multiple topic areas. This concept enables simple and fast computation for topic independent text analysis, and is therefore a very useful and important new concept.

We shall further explain topic independence using sentiment analysis as an example. A term can have either a positive, a negative, or a neutral connotation when it is used in context. Framed as a binary classification problem, the presence of any term is either an indicator of positive sentiment, an indicator of negative sentiment, or it has no class bias. This bias can be determined in context by determining if documents in that context (aka domain or topic area) are more likely to be positive or negative when that term is present. Given a set of different contexts we can count the number of contexts where the term is positive and the number of contexts where the term is negative. In Figure 1 we chart these values along the x and y axis for every term in the popular 25 category Amazon Product Reviews Dataset [2]. This chart shows why our Topic Independence Scoring Algorithm is so important.

Sentimental topic independence is a matter of degree: there is almost always some situation where a normally positive word or phrase will have a negative connotation. The topic independent sentiment bias of a term should be based not only upon its sentimental strength in most situations, it must also be weighted by it‘s reliability and uniformity. Put another way, the exceptions are so frequent that they must be accounted for in the general rule.

Figure 1 shows that there are only 11 terms that have a positive sentimental orientation in all 25 product review categories, while 16 terms have a negative sentiment orientation. For example, the 11 most topic independent positive terms: “excellent”, “highly recommend” ”, “the best”, “best”, “an excellent”, “I love”, “love”, “wonderful”, “a great”, “always”, and “recommend” occur at. For example, the 11 most domain independent positive terms: excellent, highly recommend , the best, best, an excellent, I love, love, wonderful, a great, always, and recommend. The most topic independent negative terms include: “don’t waste”, “ your money”, “waste”, “waste your”, “would not”, “money”, “disappointed”, “worse”. These terms are very revealing, but are not enough to cover a representative sample of any given text document.

The vast majority of all terms, over two million, are unique to exactly one product category in our dataset. From that peak, the total volume of terms falls off very rapidly according to the degree of topic independence. This implies that we need to properly scale the sentiment strength scores for terms with their degree of sentimental topic independence in-order to use the less topic independent terms without overpowering the more topic independent terms.

3 Approach

The unique idea of our approach is to build a topic independent model by scoring terms based upon how much their class bias shifts *as observed across many topics*. By doing this *irrespective of the target topic area* where the model will be applied we can be reasonably confident that the model will work well for any topic area. This contrasts quite sharply with domain adaptation methods that seek to adapt a model build for one domain into a model that will better fit another specific domain. Using domain adaptation thus ensures that you will need to domain adaptation again for the next domain. Furthermore, this kind of custom fitting to a single dataset is more likely to overfit artifacts in those datasets than a model that must fit multiple different domains since artifacts can be cross-checked with other domains. Domain independent models are much more useful than single domain models because they are more broadly applicable and less susceptible to artifacts and other noise.

Training a classifier with out-of-domain data can be accurately preformed if you can answer two key questions:

1. For any term, what is that term's class bias in the source domain(s)?
2. From this bias what can be concluded about its bias in the target domain?

The first question is fairly straight forward and easy to answer using standard techniques for supervised machine learning. Delta TFIDF [8] weights works particularly well for this task [9], but they can easily be replaced as the state-of-the-art advances.

Answering the second question is difficult for current domain adaptation approaches because they model the situation as a relationship between a predetermined training topic area and the target topic area. This setup assumes a different relationship between each pair of topics.

Question two is very difficult to answer with that assumption, so let us instead *assume that the class bias of a term is equally likely to shift between any randomly selected pair of topics.* This implies that we can predict how likely any given term's class bias is to shift when applied to another arbitrary topic area simply by observing how frequently it actually shifts class bias between multiple topic areas. Similarly, we can observe the magnitude of these class bias shifts to determine the likely magnitude of the class bias for other arbitrary topic areas.

Term level topic independence class bias scores need to measure and unify the following semantics:

– *Sensible orientation* : A term's class orientation should agree with its overall orientation in the set of topic areas.
– *Strength* : Terms with higher scores in the topic areas should have higher scores.
– *Broad applicability* : Terms that are used in more topic areas should have higher scores.
– *Uniform meaning* : Terms with more uniform topic area scores should have higher scores.

To measure these semantics we introduce our Topic Independence Scoring function in Equation 1. *Strength* can be measured with a simple average. This average should also give us a *sensible orientation.* A strongly oriented term is more valuable as it becomes more *broadly applicable* so multiplying the strength score and the applicability score makes sense. The *uniform meaning* metric is difficult. Variance is not a good choice since variance scores increase with dis-uniformity, have an undefined range, and when all the values are multiplied by a constant the variance goes up by the square of the constant. Attempting to address the dis-uniformity problem by dividing the other scores by the variance is not a good solution because this can cause divide by zero problems and because of the rate at which the variance score changes. A good way to score uniformity is to use the geometric mean of the topic area scores. The geometric mean is a good choice because it has a predefined range with a maximum equal to the arithmetic mean when the values are totally uniform and with scores dropping as uniformity decreases. This final uniformity term should be multiplied with the earlier calculations because, a strong broadly applicable term is more valuable when the strong scores are more uniform.

Given that:

D_t is the number of topics that term t occurs in.
$S_{d,t}$ is the class bias score for term t in topic area d.
$TIS(t)$ is the feature value for term t.
We calculate Topic Independence Scores with the following formula:

$$TIS(t) = \sum_{d}^{D_t} S_{d,t} \left(\prod_{d}^{D_t} |S_{d,t}| \right)^{1/D_t} \tag{1}$$

The Topic Independence Scoring Algorithm creates a topic independent model from a set of existing topic dependent models using the TIS function on each term. Algorithms, such as SVMs and Logistic Regression use weight vectors to produce judgments. With a set of topic area specific models built by such algorithms TISA can produce a new topic independent weight vector covering all the terms in the source models. This new weight vector can be used to do topic independent classification using the same classification algorithm that produced the original topic area specific weight vectors. Our topic independent classification can be easily used with a wide variety of popular machine learning algorithms: There is a very low adoption barrier.

4 Evaluation

In this evaluation we demonstrate how to build topic independent sentiment models using our topic independence scoring algorithm. We demonstrate that:

1. Topic independent sentiment models outperform in-topic models.
2. Topic independent models use additional out-of-topic training data more effectively than alternative techniques including:
   (a) Weighted voting with multiple models.
   (b) Building a single model on the union of multiple topic area datasets.
3. Topic independent sentiment models can be used to find revealing and informative topic specific vocabulary.

Our topic independent sentiment model is 89.6% accurate when measured over 20 additional held-out test topic areas with a low variance of 5.05 percentage points. Our approach is the most accurate approach published on this dataset.

4.1 Test 1: TISA vs. In-topic Models

This test evaluates our topic independence scoring algorithm as a method for domain independent sentiment classification using 20 different held-out test topic areas.

For our baseline we used the standard 10-fold cross-validation methodology in each of the 20 test topic areas. For this baseline we choose to use the Delta IDF [7] classification algorithm, which is a slight modification on the Delta TFIDF document feature weighting algorithm [8]. To train a Delta IDF model calculate each feature in the bag-of-words as shown below and add them to a weight vector. Given that:

$|P_t|$ is the number of positively labeled training documents with term t.
$|P|$ is the number of positively labeled training documents.
$|N_t|$ is the number of negatively labeled training documents with term t.
$|N|$ is the number of negatively labeled training documents.
V_t is the feature value for term t.

$$V_t = \log_2\left(\frac{(|N|+1)\,(|P_t|+1)}{(|N_t|+1)\,(|P|+1)}\right) \tag{2}$$

We need to balance the positive vs. the negative bias because we know that the datasets have been class balanced by the original author of the dataset. Follow the procedure described below.

Bias Balancing Procedure:

1. Create a copy of the weight vector and call it the positive vector. Call the original vector the negative vector.
2. For every feature in the positive vector, if the feature value is less than zero set the value to zero.
3. For every feature in the negative vector, if the feature value is greater than zero set the value to zero.
4. L^2 normalize the positive vector
5. L^2 normalize the negative vector
6. Add the positive and negative vectors together and return the answer.

For our classification function we use the dot product of the document with the weight vector. Data points with a dot product greater than or equal to zero are positive, otherwise the point is negative.

To keep our comparison uniform and meaningful we apply the same bias balancing procedure and use the same classification function for both TISA and Delta IDF. Bias balancing is a good idea for TISA because the overall class balance is topic area dependent. For example, while most people love their digital cameras they absolutely hate their anti-virus software.

To further eliminate external factors we use the Delta IDF algorithm to produce the set of domain specific feature scores used by TISA in equation 1.

Since Delta IDF does not have any tunable parameters, no one can claim that the input models for TISA were better tuned that the baseline models. These choices remove potential confounding factors that could have been responsible for TISA's better performance.

We built our topic independent model from a set of five topic dependent models using TISA as described in our approach. The five source models were built using Delta IDF on a different set of topic areas than the 20 held out test topic areas. The five source models were built on the books, dvds, electronics, kitchen appliances, and music topic areas because these are the most popular domains. This matches real world situations where there exists more labeled data for popular topic areas and far less labeled data for other areas.

Target Category	**In-Dom Model**	**TISA Model**
Apparel	89.17	89.90
Automotive	80.92	85.99
Baby	89.41	90.32
Beauty	85.38	90.62
Camera	86.54	91.56
Cell Phone	83.66	83.82
Comp Games	72.77	88.04
Food	76.41	88.86
Grocery	84.25	89.14
Health	87.36	89.31
Instruments	84.28	90.32
Jewelry	85.32	89.44
Magazines	85.40	89.68
Office	76.32	89.91
Outdoor	84.13	92.41
Software	79.44	87.43
Sports	87.09	90.24
Tools	56.67	94.74
Toys	86.87	90.40
Video	84.19	89.46
Average	82.28	89.60
Variance	55.70	5.05

Table 1. A general model built form using TISA to combine Delta IDF scores on data about books, DVDs, electronics, kitchen appliances, and music does very well on 20 different product categories when compared to in-domain models built using Delta IDF on each of the categories.

On average our topic area independent model is 89.60% accurate, which is a statistically significant improvement over the 82.28% accurate product area specific Delta IDF baselines to the 99.9% confidence interval. Table 1 shows the accuracy of our TISA model compared to the baseline for each of our 20 test product review categories. Please note that the low accuracy of the tools baseline is not a mistake. We will discuss it in greater detail in the next section.

Unlike other algorithms, TISA is highly accurate on every topic area with very low variance. Even though many of the topic areas are substantially different from TISA's training data our TISA model is more accurate and nearly 11 times more stable in terms of variance than the domain specific models. While domain adaptation algorithms try to exploit the relationship between topic areas, TISA attempts to minimize the effects of these relationships. This decouples TISA's training topic areas with its testing topic areas. This has the added benefit of allowing researchers working with TISA to use labeled data from any topic area. This can allow researchers to avoid using low quality topic area datasets, such as topic areas with very little data, harder data points to classify, or low inter-annotator label agreement.

Table 2 illustrates the difference between topic independent term scores produced by TISA and topic dependent scores that were used as input to TISA. This table shows the top 50 most negative and most positive words or pairs of words for TISA and the baseline models. The terms highlighted in the figure show that TISA's most important terms are very general purpose, while the terms in the input books model are very specific to the books topic area. These example terms support our argument that TISA favors topic independent bias terms.

The product specific baselines built using Delta IDF make for an excellent comparison. These product specific baselines are not straw men; they have been shown to outperform Support Vector Machines on this dataset [7]. By correctly setting up our experiment we have eliminated confounding factors and can conclude that the quality of the models is responsible for the difference between the two algorithms. By evaluating TISA against the Delta IDF algorithm used to create its constituent sub-models we negate any potential objections that our improvement was due to the difference between the baseline algorithm and the algorithm used to create the sub-models. Thus the difference between the two models comes from either the intelligent combination of models using TISA, or the amount and quality of the training data. Both of these are good points for TISA, since TISA allows the researcher to freely select dataset without respect to the topic area that the model will be used on.

4.2 Test 2: TISA vs. Ensemble Methods

Skeptical readers might object to comparing TISA against in-domain Delta IDF because TISA is using more total labeled data. In machine learning, it is well known that using more training data will improve accuracy, but it is also well known that using training data that is not similar to the test data will hurt accuracy. One of TISA's main benefits is that it allows machine learning practitioners to leverage large amounts of dis-similar data by reducing the impact of the dis-similarity. The tools entry in Table 1 is a clear example of why this approach is so important: using more training data is the entire point of domain adaptation.

One reason why TISA is very accurate is that it preserves and intelligently uses the information captured by splitting the document pool into different domains or topic areas. Consider building a Delta IDF dot product classifier using the union of all the data that the topic independent model was trained on. That

TISA Identifies General Terms by Decreasing the Score of Topic Specific Terms

Positive Terms Books Domain	Positive Terms TISA	Negative Terms Books Domain	Negative Terms TISA
must for	*highly recommended*	waste your	waste your
magnificent	only complaint	not worth	very disappointed
only complaint	must for	two stars	two stars
worth every	worth every	very disappointing	a refund
excellent read	great addition	**worst book**	refund
a must	delighted	uninteresting	don't waste
wonderful book	a must	very disappointed	waste of
delighted	*great buy*	sorry but	*not recommend*
definitely worth	every penny	don't waste	your money
great resource	excellent for	a joke	not worth
excellent overview	an outstanding	waste of	zero stars
excellent reference	must-have	not waste	very disappointing
a delight	*well worth*	very poorly	complete waste
essential reading	*another great*	is poorly	save your
must-have for	definitely worth	save your	very poorly
my clients	a must-have	a disappointment	*avoid this*
weaves	and allows	poor quality	not waste
great addition	my only	no new	waste
detailed account	great way	refund	a waste
a magnificent	*highly recommend*	a poorly	*buyer beware*
great fun	*are amazing*	excuse for	total waste
offers an	*is superb*	wasted my	wasted my
pleasantly surprised	*excellent condition*	complete waste	big disappointment
every penny	exceeded my	skip this	a disappointment
great introduction	superb	big disappointment	*of junk*
pleasure to	*pleasantly surprised*	zero stars	hard earned
and accessible	*is awesome*	**terrible book**	really disappointed
be required	*great condition*	**worst books**	a joke
really helped	*great product*	**poorly organized**	*don't buy*
be missed	not disappoint	**good reviews**	*stinks*
not disappoint	i highly	your money	*money back*
top notch	*excellent choice*	disappointing	*a poorly*
terrific book	*best ever*	**boring book**	poor quality
beautifully written	excellent i	a refund	*returned this*
excellent resource	loves this	unfortunately this	*insult to*
transcends	outstanding	**poorly written**	or money
renewed	delighted with	**factual errors**	*extremely disappointed*
great collection	recomend it	**glowing reviews**	*is terrible*
fabulous book	gem	new here	disappointment
must-have	*loves it*	disappointment i	*not buy*
first rate	*very pleased*	total waste	*not recommended*
an outstanding	*definitely recommend*	am disappointed	stay away
refreshing and	no nonsense	was boring	don't bother
you wanting	also great	irritated	worthless
a pleasure	*can't beat*	**even finish**	*i regret*
developing a	the raw	disappointing i	huge disappointment
teaches us	great look	had hoped	*never buy*
from home	thumbs up	disappointment	*dud*
poems and	she loves	**drivel**	disappointing
very comprehensive	*love this*	a waste	the trash

Table 2. Top 50 most positive and negative terms for the Books domain as determined by in-domain Delta IDF vs. Top Most positive and negative terms as determined by TISA using Books, DVDs, and Electronics. All terms shown have the correct sentimental orientation and are strongly oriented. However, in-domain Delta IDF identifies many features, shown in bold and highlighted in **red** , that will not generalize well to non-book data. Instead, TISA placed more importance on terms, shown in italics and highlighted in *green* , that should generalize very well to other domains.

process ignores the information provided by the subdivision in the dataset between different domains. Table 3 shows that the TISA model is more accurate than a Delta IDF classifier created from the union of the same set of documents at an accuracy of 89.6% to 86.3%. This difference is significant to the 99.5% confidence level. Clearly it is better to use the information provided by domain membership than to ignore it.

Target Category	**Dom Size**	**In-Dom Model**	**TISA Model**	**Union Model**	**Weighted Voting**
Tools	19	56.67	94.74	73.68	84.21
Instruments	93	84.28	90.32	88.17	87.10
Office	109	76.32	89.91	88.07	87.16
Automotive	314	80.92	85.99	81.85	80.57
Food	377	76.41	88.86	86.21	87.27
Computer Games	485	72.77	88.04	85.98	84.33
Outdoor	593	84.13	92.41	90.22	89.71
Jewelry	606	85.32	89.44	88.45	88.61
Grocery	654	84.25	89.14	88.23	88.69
Cell Phone	692	83.66	83.82	78.90	79.05
Beauty	821	85.38	90.62	87.33	88.67
Magazines	1124	85.40	89.70	86.39	87.82
Software	1551	79.44	87.43	84.53	83.75
Camera	1718	86.54	91.56	88.07	88.53
Baby	1756	89.41	90.32	89.07	89.46
Sports	2029	87.09	90.24	87.83	88.37
Apparel	2603	89.16	89.90	88.21	89.44
Health	2713	87.36	89.31	85.51	86.10
Video	4726	84.19	89.46	90.12	88.30
Toys	4929	86.87	90.40	89.06	89.53
Average	2317	82.28	**89.60**	86.30	86.83

Table 3. General TISA "BDEKM" model built from the Books, DVDs, Electronics, Kitchen Appliances, and Music Delta IDF models vs. Weighted Voting with these models vs. a single Delta IDF model built on the union of all the Books, DVDs, Electronics, Kitchen Appliances, and Music data. Results have been sorted by size. The 10-fold in-domain accuracies for each test domain are displayed for reference.

A popular alternative technique to leverage more out of domain data is to use multiple classifiers under a weighted voting approach. Delta IDF dot product classification is particularly well suited to this approach because, when both the documents and the weight vectors are normalized to unit length, the magnitude of the dot product can serve as the vote's weight. Weighted voting using the books, DVDs, electronics, kitchen appliances, and music domains over the test domains is 86.83% accurate. The difference between weighted voting and the TISA method using the same training and test points is significant to the 99.9% confidence level. The weighted voting approach is statistically no different than the union model as indicated by a p-value of .3555 . These results are displayed in detail in Table 3.

4.3 Sentiment Feature Mining

In many case it is valuable to know what the important domain specific bias features are. For example, someone who is shopping for clothes may want to know why a specific article of clothing was rated poorly by users. While reporting to the shopper the highest scoring topic independent features for the product will clearly show that people did not like the product, it will not do a good job of showing why people did not like the article of clothing because topic independent features are very generic. To solve this sentiment mining problem we must report to the shopper the topic specific reasons why people did not like the article of clothing.

Fortunately, the topic-independent model can be used to automatically generate topic-specific sentiment models. These topic specific models can then be used to report specific reasons why people liked or disliked the topic.

Positive Terms		**Negative Terms**	
compliments on	toe is	returned them	poor customer
great quality	hubby	holes	received a
is comfortable	thick as	defective	credited
are soft	so soft	cheaply made	disappointed when
great item	wanted a	the return	recieved the
confortable	tons of	policy	post
great shoes	best bra	make sure	return shipping
ones and	locally	charged	cancelled
and comfortable	monday	the photo	i emailed
with jeans	great	to remove	never order
great bag	really great	sent the	ears
fit very	definitely buy	so thin	wont
them very	best shoes	send the	item back
khaki	are exactly	the ankle	top and
were exactly	sleek	off my	tore
comfortable they	walking shoe	ordered <num>	too wide
is slightly	good shoe	known	i see
he really	ride up	times and	the seam
love em	last forever	holes in	just about
feels great	things and	shrunk	pay to
reasonable price	under jeans	so tight	pants were
many different	very confortable	<num> sizes	thin that
bra ever	as thick	big and	opened
comfortable from	wanted something	thin and	ordered a
even in	tons	torn	uncomfortable the

Table 4. Top 50 most positive and negative terms mined for the apparel topic area using the topic independent model built by TISA on books, DVDs, electronics, kitchen appliances, and music data. The terms are strongly sentimental and are correctly oriented for apparel. The terms tend to be very specific to the apparel topic area.

This takes 3 steps: (1) gather a set of documents about the topic the user is interested in, (2) classify every document using the topic-independent model and label them as positive or negative with the classifiers decision, (3) compute ΔIDF(t) scores for terms in the set of documents that were mechanically labeled in the previous step. The top most features of this model are the strongest reasons why people liked or disliked the product. Table 4 shows the top 50 strongest sentimental terms for the clothing topic computed using this method.

The words and phrases shown in Table 4 are good apparel specific indicators of sentiment that help explain why a user liked or disliked the piece of apparel under review. Many of these phrases express an opinion about an apparel specific product feature. For example, "Feels great" indicates a positive opinion about the feel of the clothing. Likewise, "Great quality" expresses a positive opinion about the item's quality. Other phrases assert a good, or bad, property of apparel for the item under review. Examples include "Is comfortable" and "Are soft" both of which are desirable aspects of many apparel items. Other stop words, or near stop words, are informative components of strong apparel specific sentiment indicators. Sentiment amplifiers such as "So" ,"Very" , and "Really" are important stop words because they amplify the strength of the rest of the phrase. The presence of these phrases can indicate why a user gave a positive or negative rating to a piece of apparel.

5 Related Work

Supervised machine learning is a common approach for sentiment analysis. Normally, a classifier is trained on a hand labeled dataset for the specific topic area of interest. Training these classifiers generally takes a long time, but once they are trained they can rapidly make accurate judgments of the type they were trained to make, on the type of things they were exposed to during the training process. Using Support Vector Machines [6] with a bag-of-words feature space is one of the most popular examples of this approach, including the seminal work on sentiment analysis for movies [11].

While these in-domain methods work well in a predefined topic area with a sufficient amount of labeled data they do not work well when used outside of the predefined topic area. As a result these methods do not work well for important applications, such as personal assistants, that need to provide answers for any domain, or topic area, that the user is interested in at the moment.

Current domain-adaptation approaches such as CODA [4], SCL-MI [2], SFA [10], and Couple Spaces [3] build a model for a domain, which has no labeled data, using labeled data from a different domain. This is unacceptable because it is infeasible to train a new model in real-time whenever an electronic personal assistant encounters a question about a new domain.

To address these challenges and enable personal assistants to succeed in unexpected topic areas we took a strikingly different approach to re-score sentiment features using their domain-independence. Our work alone has been designed to build models that remain highly accurate even when they are used on unfamiliar topics that may be vastly different.

In a business setting it is highly desirable to be able to deploy trained models on new topic areas that they were not designed for. Training these models should not require any special changes for the topic area. Furthermore, these models should be highly accurate in every topic area that they will be used upon even if the list of topic areas they will be used upon is unknown. Unlike state-of-the-art Domain Adaptation approaches, our TISA fulfills these demands as summarized in Table 5.

Our approach is highly accurate across 20 never before seen test domains. Surprisingly, our algorithm is even more accurate than models that were custom tailored to the test domains.

Comparison Criteria	**TISA**	In-Dom ΔIDF	SCL-MI	SFA-DI	CODA with 0 Target	CODA with 1600 Target
Situations Modeled	20*	20**	12***	12***	12***	12***
Requires Labeled Data from Other Domains	Yes	No	Yes	Yes	Yes	Yes
Requires In-domain Labeled Data	**No**	Yes	No	No	No	Yes
Requires Unlabeled In-domain Data	**No**	No	Yes	Yes	Yes	Yes
Average Accuracy	**89.6**	82.28	77.97	78.66	83.23	86.46
Variance	5.05	55.70	25.38	17.29	11.54	2.89

Table 5. TISA has the easiest to satisfy training data requirements, is simple, fast, highly accurate, and reliable. Caution should be taken when directly comparing the average accuracy and variance numbers of TISA and our ΔIDF baseline to other published approaches due to the different training environments described.

6 Conclusion

In this paper we showed that topic-independent sentiment analysis is highly important for a wide array of applications. We pointed out how state-of-the-art domain-adaptation approaches do not address these problems. To address these problems, we designed our approach with the core goal of accurate sentiment classification for unforeseen topic areas.

Our algorithm has several advantages over other approaches because it does not require any information about the topic area, including labeled or unlabeled data from the topic area. First, machine learning experts can use our scoring algorithm with the most appropriate algorithm for the task at hand. Second, even if the training data has been lost, is inaccessible due to business reasons, or the expertise to tune the original algorithm is no longer available, existing models can still be used with TISA to produce topic independent models. Third, training time is substantially reduced for super-linear training algorithms by cutting the number of documents down into multiple smaller pools. Fourth, TISA

* Each modeled situation corresponds to a product review category since each is a held-out test set.

** Each product review category is a topic area and is treated as a test situation. Although 10-fold cross-validation is used in each product review category folds are not counted as a test situation. Average and variance scores are computed over test situations. Please note that the average and variance reported in this table for Δ IDF includes domains that TISA was trained on.

*** Each unique source/target product review category pair is being treated as a modeled situation. Every domain adaptation source/target pair for the Books, DVDs, Electronics, and Kitchen product review categories were modeled.

can leverage existing labeled data in any number of topic areas. We speculate that this reduces overfitting and leads to our demonstrated better results.

TISA is the only true scalable topic-independent sentiment analysis solution for real world problems. A single topic-independent model built using TISA is vastly preferable to using multiple models domain specific models for the following reasons: One, a single model is much easier and less costly to create and maintain. Two, topic independent models do not require topic detection to determine which domain specific model to use. Three, topic-independent models created using TISA are even more accurate than topic-specific models due to their ability to leverage more data and reduce the affects of noisy features. Four, our topic-independent models are 11 times more reliable than domain specific models. Five, TISA models require no changes to work well on a new topic area. These factors make TISA the best choice for practical real world sentiment analysis.

References

1. S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the Seventh conf. on LREC, Valletta, Malta, May. ELRA*, 2010.
2. J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting of Association For Computational Linguistics*, volume 45, pages 440–447, 2007.
3. J. Blitzer, S. Kakade, and D. P. Foster. Domain adaptation with coupled subspaces. *Journal of Machine Learning Research - Proceedings Track*, 15:173–181, 2011.
4. M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *NIPS*, pages 2456–2464, 2011.
5. C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.
6. T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998.
7. J. Martineau. *Identifying and Isolating Text Classification Signals from Domain and Genre Noise for Sentiment Analysis*. PhD thesis, University of Maryland, Baltimore County, Computer Science and Electrical Engineering, December 2011.
8. J. Martineau, T. Finin, A. Joshi, and S. Patel. Improving binary classification on text problems using differential word features. In *Proceeding of the 18th ACM CIKM*, pages 2019–2024. ACM, 2009.
9. G. Paltoglou and M. Thelwall. A study of Information Retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 1386–1395. ACL, 2010.
10. S. Pan, X. Ni, J. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760. ACM, 2010.
11. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on EMNLP Volume 10*, pages 79–86. ACL, 2002.
12. C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of LREC*, volume 4, pages 1083–1086, 2004.