SAND2011-0102C

Multilingual Sentiment Analysis Using Latent Semantic Indexing and Machine Learning

Brett W. Bader
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1318
bwbader@sandia.gov

W. Philip Kegelmeyer
Sandia National Laboratories
P.O. Box 969, MS 9159
Livermore, CA 94551-0969
wpk@sandia.gov

Peter A. Chew
Moss Adams LLP
6100 Uptown Blvd, Suite 400
Albuquerque, NM 87110-4189
pchew@mossadams.com

Abstract

We present a novel approach to predicting the sentiment of documents in multiple languages, without translation. The only prerequisites are an (unannotated) multilingual parallel corpus and a list of words, each characterized by normative emotional ratings, in just one language. An example of such a list is the Affective Norms for English Words (ANEW) dataset (Bradley and Lang, 1999). Latent Semantic Indexing (LSI) converts the multilingual corpus into a multilingual 'concept space'. New documents are then projected into that space, allowing cross-lingual semantic comparisons between the documents without the need for translation. Meanwhile, the single language affective-norms wordlist is used to determine sentiment polarity for a set of documents in its language. Those documents, with known sentiment, then train a machine learning algorithm which can, because of the multilingual nature of the document projections, be used to predict sentiment in all the other languages. We evaluate the accuracy of this approach, and also quantify the extent to which topic and sentiment separately contribute to classification accuracy in a way which we believe sheds some light on whether topic and sentiment can be sensibly teased apart.

1 Introduction

In this paper, we explore whether sentiment analysis and a multilingual vector-space approach to information retrieval can be integrated (in what, to the best of our knowledge, is a novel way) to address the problem of multilingual sentiment analysis.

Perhaps because sentiment analysis deals with what is inherently subjective, and perhaps because it is still a relatively new (decade-old) field, it remains hard to formulate sentiment-analysis problems in a computationally well-defined manner (Liu, 2010). Yet the growth of the World Wide Web has driven a great deal of practical interest in this area, as grassroots opinions and reviews on everything from consumer products to government foreign policy actions are posted online. For anyone involved in marketing or public relations, there is a wealth of free information available just waiting to be mined. Early forays into this field were made by Turney (2002) and Pang et al. (2002), who investigated the application of machine-learning techniques to product and movie reviews, treating the reviews as bags-of-words.

In the last couple of years, interest has grown in application of sentiment analysis to multilingual text; companies and governments are increasingly interested in how their products or actions are perceived worldwide. Research in this area is still fairly sparse, but approaches include first translating documents from source languages into English (for which sentiment-tagged data is more readily available), then applying sentiment analysis (Denecke, 2008), or the reverse: translating sentiment-tagged lists from English and then using these with to classify documents in the other language (Mihalcea et al., 2007).

Here, we take a novel approach that avoids the need for translation altogether, building upon the framework of multilingual Latent Semantic Indexing (LSI) (Berry et al., 1994; Young, 1994). LSI is a particular instantiation of the vector-space view of language. Under this view, the terms in documents are counted and each document is then represented as a vector. If n is the number of items in the vocabulary, the vector is n-dimensional. In fact, this is precisely the approach taken by Pang et al. (2002), though they describe the terms as 'features', not 'dimensions'. But this is essentially a terminological difference between the fields of machine learning and vector-space information retrieval.

In addition to enabling one to characterize documents and terms using a very limited number of features (which can greatly reduce the overhead for machine learning problems), LSI also has the proven ability to bring documents into a single 'language-blind' concept space, enabling documents in one language to be compared directly to those in another with high precision (Chew et al., 2010).

To motivate the integration of multilingual LSI and machine learning, consider that by using terms as features in a nonreduced space, documents can be classified by sentiment with accuracy well above the baseline (Pang et al., 2002). We also know that the multilingual LSI approach achieves precision of up to 95% in a non-trivial multilingual clustering task (Chew et al., 2010) — and therefore that multilingual LSI can capture the overall semantics of documents, even abstracting away from the particular languages they are in and even though the reduced space of LSI only approximates the non-reduced space pre-LSI. While sentiment may be hard to define computationally, our specific hypothesis here is simply that sentiment forms a part of a document's overall semantics, and that like topic, sentiment is at least somewhat independent of the language a document is in. If this is the case, it should be possible to detect (predict) sentiment in the LSI vectors for documents using machine learning techniques.

Related to the question of whether sentiment is detectable in the document vectors, and just as important, is that of whether sentiment and topic are separable in the document vectors. It could be that sentiment is hard to define simply because it is fully bound up with topic; in simple terms, sentiment might be a type of topic. If true, this could have major implications for the field of sentiment analysis. LSI, however, provides a way we can investigate this part of the problem too. One of the by-products of LSI is a ranked list of the most important topics in a corpus. By ensuring that the documents we use for machine learning come from a *mix* of topics, we can test our hypothesis that sentiment is not just a type of topic: evidence for this hypothesis is the predictability of sentiment even when topics are mixed.

To explore these ideas, our first step, in Section 2, will be to introduce the different datasets (Europarl, the Bible, and ANEW) that support our analysis. We propose a method for attaching sentiment features (based on ANEW) to untagged English-language documents from the Bible; these are then subjectively evaluated to confirm their reasonableness. The sentiment features are extended to the translations of the English documents on the weak assumption that whatever sentiment is present in an English document will also be present in translations of that document in other languages. This provides the 'ground truth' for our evaluation of cross-language sentiment analysis. In Section 3, we briefly review the mechanics of LSI, explaining why it works for cross-language information retrieval, how we apply it to the Europarl parallel corpus to construct a cross-language semantic space, and outlining the specific LSI settings which have been found to work best for this kind of application. In Section 4, we show specifically how the output of LSI can be used to characterize terms and documents (both those from Europarl and new documents) in feature space in a way which allows a natural formulation of our multilingual sentiment analysis task as a machine learning problem. In Section 5, we test whether sentiment is indeed predictable by mixing the vectors for documents in different languages, both with and without an enforced mix of topics. We conclude and suggest further research in Section 6.

2 Datasets

To set up and enable empirical evaluation of our approach to cross-language sentiment analysis, only two elements are absolutely necessary: (a) a parallel (multilingual) corpus, and (b) documents in different languages which are tagged according to sentiment (however that is defined). The tagged documents in (b) could be from the multilingual corpus in (a), or they could be a collection of non-parallel documents which happen to be in multiple languages. We use the documents in (b) as our 'ground truth', and attempt to predict their sentiment, using some portion of (b) to train a machine learning algorithm and holding the other portion out for testing.

For our research, we have chosen to use Europarl (Koehn, 2005) as our base multi-parallel corpus. Europarl is a parallel corpus extracted from the proceedings of the European Parliament; our copy of the dataset contains speeches in 11 languages, aligned at roughly sentence level, in several hundred thousand parallel chunks.

We choose to set up the source data in a slightly more elaborate way than outlined previously by using a separate multi-parallel corpus for (b). The second multi-parallel corpus is the Bible, which we have in 54 languages.[1] We use two separate parallel corpora for the following reasons:

- Use of a parallel corpus in (b) is a control to ensure that exactly the same sentiment which is present in documents of one language, is present in documents of all languages.

- Use of *separate* corpora in (a) and (b) ensures that the data used to train LSI are kept separate from data used for machine learning. Fundamentally, we do this because we recognize that in most real-life applications, we will not want to predict sentiment in documents from a parallel corpus. Whether or not (b) is a parallel corpus, its documents can be 'projected' into the semantic space of (a) through a matrix multiplication operation which will be described in section 4. Effectively, we will therefore test whether the sentiment of the parallel Bible documents is still present after those documents are projected into the Europarl semantic space.

Since the Bible is, as far as we know, untagged for sentiment, we had to prepare those sentiment tags ourselves, at least to the extent needed for training a machine learning tool. To accelerate that process, we drew upon the Affective Norms for English Words (ANEW) dataset (Bradley and Lang, 1999), which provides normative emotional ratings for just over 1,000 English words, based on aggregating responses from subjects in psychophysiological experiments.

There are 1,189 chapters in the Bible, and we chose to chunk the Bible at the chapter level and obtain sentiment labels for roughly 200 of them. To compute a rough score for the chapter valence, we use a simple weighted average of the ANEW valence[2]. For each chapter, we look up the ANEW valence value for each token in that chunk and sum the valence scores weighted by the token frequency in each chapter, and finally divide by the total number of ANEW words in the chapter. The resultant score will be in the same range as the ANEW valence (between 1 and 9). For example, if a particular document has 9 mentions of "happy" (8.20) and 1 of "unhappy" (1.57), then the overall score would be 7.5. Using this method, the chapter valence scores range from 4.25 for Ezekiel 5 to 7.8 for Psalm 117.

Admittedly, this is a crude approach that may miss simple semantics (e.g., negation), but to a first approximation, it serves our need for a way to sort the chapters to ease manual inspection for label assignment. Moreover, use of the ANEW dataset in this way allows us to initially

[1] Use of the Bible as a parallel corpus was first proposed by Resnik et al. (1999).

[2] High valence in ANEW corresponds to 'pleasant'; low valence corresponds to 'unpleasant'. The three words with the highest valence in ANEW are 'triumphant', 'paradise', and 'love'; those with the lowest valence are 'rape', 'suicide', and 'funeral'.

base our hand tagging on independent psychological research.

Once we had the initial per-chapter valence scores, we sorted them to find the chapters with the top 100 and bottom 100 values for valence, and then manually inspected those, to weed out the small handful of chapters whose sentiment was mis-estimated by this process. An example of a 'positive' Bible chapter produced by this process is Psalm 126 (the theme of which is thanksgiving), and examples of 'negative' chapters are 1 Samuel 31 (about the defeat and slaying of Saul and his sons) and Revelation 9 (describing a demonic plague of locusts).

3 Using LSI and Europarl to construct a semantic space

In the standard LSI framework (Deerwester et al., 1990) a term-by-document matrix X is factorized by the singular value decomposition (SVD),

$$X = USV^T. \tag{1}$$

Each column vector in U maps the terms in the corpus to a single arbitrary concept, such that semantically related terms will tend to group together with similar values in columns of U.

Typically, however, a truncated SVD is computed: if R indicates the reduced number of concept dimensions in LSI, only the R largest singular values in S are kept, and the rest discarded. Similarly, only the first R vectors of U and V are retained. This means that equality in (1) no longer holds; USV^T (after truncation) represents the best rank-R least-squares approximation to matrix X. Here, based on prior work with multilingual clustering, we choose to set $R = 300$.

It should be noted that the best results are obtained in LSI when care is taken over the preparation of the X matrix before SVD. Each entry X_{ij} in X represents the weight of a particular term i in a particular document j. Typically, the 'weight' is not the raw frequency of i in j, but rather a weighted frequency. A popular weighting scheme is log-entropy (Dumais, 1991), but it has been shown that pointwise mutual information (PMI) weighting is simpler and sometimes yields considerably better results(Chew et al., 2010). According to PMI (which we use in this paper), entry X_{ij} is

$$\log\left(\frac{p(i,j)}{p(i) \times p(j)}\right) = \log\left(\frac{p(i|j)}{p(i)}\right) \tag{2}$$

If the input is a multilingual parallel corpus (for example, Europarl), LSI can still be used, and with interesting results. Assume now that X is a five-language multi-parallel corpus (meaning X contains translations of each document into all five languages). X will then have a structure as shown in Figure 1. Here, the rows still represent terms; the horizontal bands in X show that the terms are grouped in languages, and the fact that the bands are of different heights corresponds to the fact that different languages have different vocabulary sizes. The columns also still represent documents, but the documents here are 'cross-language' (parallel text chunks), so the intersection of column j with the band for language k contains the weighting of terms in the translation of j into language k. The SVD of X is then as shown in Figure 1. Here, U now maps the terms (in specific languages) to multilingual concepts.

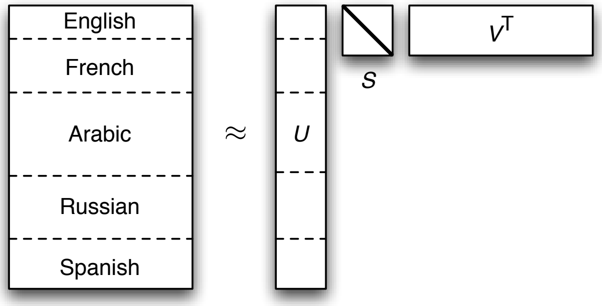

Figure 1: SVD of multi-parallel corpus

Again, in practice we take care to prepare X appropriately. We adapt (2) for the multilingual case by simply making each probability conditional upon language, as follows:

$$\log\left(\frac{p(i,j|k)}{p(i|k) \times p(j|k)}\right) \tag{3}$$

The result, when we take the Europarl matrix, form a weighted term-by-document matrix, and apply SVD, is that we have all the terms used in Europarl mapped to a cross-language concept space based on Europarl.

4 Characterizing terms and documents in Europarl semantic space

For applications in information retrieval — and for the present case — we are usually more interested in document vectors than in term vectors. Specifically in most applications of multilingual sentiment analysis that we envisage, the sentiment of a document is more important than that of a term. As explained above, the U matrix projects terms into the LSI semantic space. Likewise, V projects documents into the same space; however, it does so only for the documents in the input. If the latter is a parallel corpus, V will contain vectors only for documents in the parallel corpus, but as stated above, the documents we are ultimately interested in analyzing are very unlikely to be parallel.

Fortunately, LSI allows new documents (in our case, those not in Europarl) to be projected into the Europarl semantic space. This is achieved by multiplying term vectors for the new documents by the product US^{-1}, to yield concept vectors for these documents. These vectors encode the semantics of the non-Europarl documents just as the vectors in V encode the semantics of the Europarl documents.

In our case, then, we form a term-by-chapter matrix for the Bible and apply weighting using the expression in (3). If we call the resulting matrix B, then we can obtain Europarl concept vectors for each chapter in the Bible from the product $US^{-1}B$. As described earlier, our hypothesis is that these 'chapter vectors' sufficiently encode the semantics of each chapter such that, if treated as input variables in a machine learning task, they will enable the prediction of the known output variable 'sentiment' (obtained from the tagging procedure described in section 2).

5 Results and Discussion

5.1 The Sentiment Training Data

Since the ANEW lexicon pertains to English, we assigned sentiment to a subset of the Bible chapters in English using the process described in Section 2. After manual inspection and then reseeding with additional chapters to replace those weeded out, the result was 115 chapters labeled positive and 78 chapters labeled negative; again, all in English.

Projecting each of the 193 English Bible chapters with known sentiment labels into the multilingual concept space results in 300 features for each chapter. To build some sense of the interplay of sentiment and topic in this data, consider Figure 2, which shows a visualization of the 115+78=193 chapters embedded in two-dimensional space using multidimensional scaling (MDS) (Borg and Groenen, 2005) of a pairwise distance matrix (computed as one minus the cosine similarity of *concept* feature vectors for chapters i and j).[3] This layout has three branches, and several books of the Bible cluster in different regions. For example, Psalm is in the upper left; Proverbs and Ecclesiastes are in the lower left; and Numbers, Chronicles, and Joshua are on the right. The chapters from the New Testament tend to be in the center left.

Because the positive and negative chapters are intermixed in this layout, we have reason to believe that the chapters have topics mixed with sentiment. That is, it appears that sentiment is not highly correlated with topic. Still, this is something that we try to mitigate in our experiments.

Our nominal training set is 193 chapters with 300 features each. To expand the size of our training set, we found five English translations that appeared to have enough variation with respect to one other in terms of MDS visualizations. Table 1 shows the five English versions that we used along with their term counts and overlap with Europarl and ANEW. We created

[3] MDS finds an embedding of the chapters in n dimensions such that the distances between chapters in the higher dimensional term space are preserved as best as possible.

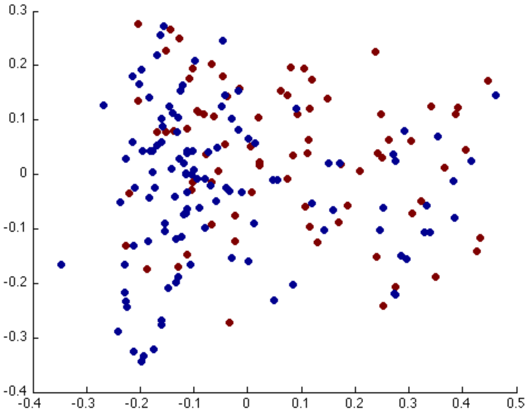

Figure 2: Two-dimensional MDS embedding of English chapters described by concept vectors. Red dots are negative chapters; blue dots are positive chapters.

individual training sets from each version and also collected them all together into one large training set for five times the amount of data.

5.2 Sentiment Prediction Accuracy

The concept feature vectors trained a nonparametric statistical prediction model (specifically, an ensemble of decision trees generated by the Avatar software package (Banfield et al., 2007)), and that model was used to predict the sentiment labels for the exact same chapters in three test languages: Spanish, French, and German.

We performed an experiment whereby we trained on actual chapters from the five English versions collected together in one single training set and then tested on actual chapters from Spanish, French, and German. The average accuracy over ten runs was 77.1%, which is better than the 59.6% baseline accuracy one could obtain by guessing that all chapters were positive.

Table 1: English translations used as training data

	Unique	Europarl	ANEW
	terms	overlap	overlap
King James	12,335	6,188	522
Young's Literal	12,192	6,149	504
Webster's Bible	12,312	6,150	505
World English	12,210	7,250	578
Basic English	5,985	2,639	276

5.3 Separating Sentiment and Topic

The accuracy results of the previous section are encouraging, but they also raise a concern: we have interpreted them as accuracy in predicting sentiment, but perhaps sentiment is so entwined with topic that all we have actually done is show, again, that one can predict topic properties.

To address this, we conducted a series of experiments on a randomized dataset that we believe minimizes this possibility. Specifically, we shuffled the verses (which are about a sentence or two in length) to make new chapters. To be precise, we separately enumerated all of the verses in each class (positive or negative) and then did a random permutation of those verses so that the original chapter length (in terms of verses) was held constant but the content was scrambled within a chapter.

The idea here is that shuffling should decouple concepts from sentiment, or at least generate largely new concepts on each run. For instance, there might be 8 chapters with 30 sentences mentioning taxes, corruption, evasion, and such, which might be enough to give rise to a "taxes protest" concept. But when those 30 sentences are doled out at random to 100 documents, that concept will likely be broken up. So, if performance trained on actual chapters is close to average performance trained on the shuffled chapters, this suggests that the performance really is due to finding sentiment.

Figure 3 shows an MDS layout of the shuffled chapters. The positive and negative chapters now cleanly separate into two groups. This is an important result for our purposes; the fact that sentiment can be a dominant characteristic in the data once topic has been "averaged away" suggest that sentiment is not necessarily inextricably intertwined with topic.

We looked at the average chapter valence for the actual chapters versus the shuffled chapters. The shuffling tends to tighten the distribution around the mean. So instead of seeing positivity / negativity piling up in a select few chapters, we tend to see a tighter cluster about the mean.

Figures 4 and 5 show histograms of the chap-

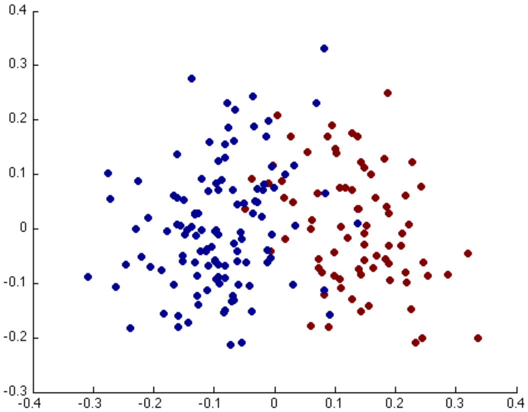

Figure 3: Two-dimensional MDS embedding of shuffled English chapters described by concept vectors.

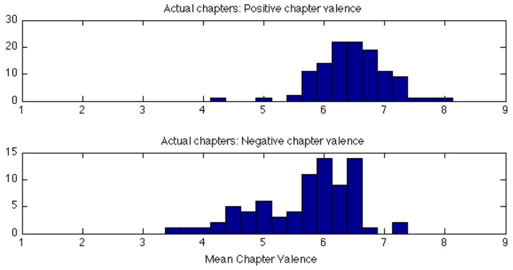

Figure 4: Valence histogram for actual chapters.

ter valence scores.[4] One can see that the shuffling tends to tighten the distribution, especially in the case of the negative chapters, which originally had a bimodal distribution.

We determined that shuffling only once (instead of every time before training a predictive model) was sufficient. We tested this by training on the King James version and testing on the three foreign languages. The results favored shuffling every time by 0.6%, but they were not statistically significant.

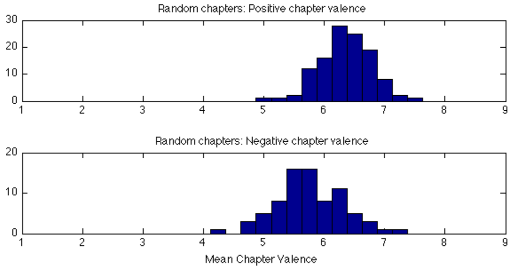

Figure 5: Valence histogram for shuffled chapters.

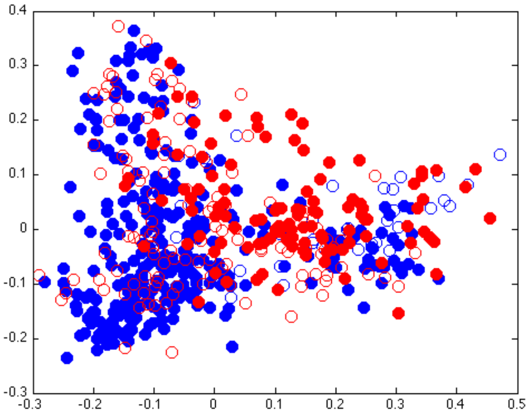

Figure 6: Two-dimensional MDS embedding of the multilingual test set. The solid dots indicate a correct prediction, and the unfilled circles indicate a missed prediction.

To test the shuffling idea, we trained on shuffled chapters from five English versions and then tested on the actual chapters from Spanish, French, and German. The average accuracy was 72.0%, which is lower than the 77.2% achieved by training on the unshuffled chapters. This suggests that there was indeed *some* sentiment predictive value in the original topics.

Figure 6 shows a plot of the two-dimensional MDS embedding of the test set labeled by sentiment and accuracy of prediction for one of the runs on shuffled training data (accuracy 71.2%). One can see that the missed predictions are not isolated to a single region, which indicates that topic has not been a factor.

To test whether individual English translations are a factor, we created a combined version that merges all terms from 5 English versions on a chapter by chapter basis (instead of keeping the 5 English versions separate for 5 times as much training data). We trained on this single version by itself and also added it to the other five for a total of six times as much training data.

[4] Because the term 'God' appears frequently in the Bible and has a high valence at 8.15, we removed this term from the chapter valence computation. Otherwise, the chapter valence scores would be skewed more positive.

Table 2 shows that the separate versions help when training on actual chapters, but not when training on shuffled chapters. This seems to suggest that with more data from the actual chapters, a predictive model is also learning topics, but the shuffling is breaking that association.

Finally, we trained on all English versions separately to see if there were particular translations that were affecting the results or if the better results were due to having additional training data. Table 3 shows that, while the results are mixed, some versions are better than others, but not to the extent that would explain the results in Table 2.

5.4 Does Shuffling Preserve Sentiment?

We assumed in the previous section that shuffling sentences across documents of like valence would preserve the valence of those documents. That seems reasonable, but by no means certain, and worthy of investigation. So to double check that shuffling does indeed preserve valence, we performed another experiment. We trained a predictive model on the original positive and negative English chapters from one of the translations, and tested it on the *shuffled* English chapters from the same translation. So here we are using the shuffled data as the test, not the training. If testing on the shuffled English data gives a result materially worse than the average accuracy derived just from the original English chapters, then that would indicate that shuffling does not necessarily preserve valence. According to our results in Table 4, all of these are higher than the opposite experiment, which suggests that shuffling preserves valence.

Table 2: Training with varying amounts of data (original/shuffled chapters).

	Actual	Shuffled
Combined version	70.9%	72.0%
5 separate versions	74.9%	72.0%
5 separate & combined	77.1%	72.7%

6 Conclusion

We have described a machine learning approach for detecting positive/negative sentiment in multilingual documents. We used only a parallel corpus and a single-language sentiment lexicon. Our experiments showed an average accuracy of about 72% for detecting sentiment. To prevent the predictive model from learning topic, a key step was to shuffle the sentences in each class, which we found helps break any topic/sentiment association.

While our investigation here centered on valence/sentiment, we see no reason why this approach could not be extended to other emotional dimensions contained in ANEW (or elsewhere), or to other meta-properties of the language only peripherally related to topic. As an example, we will soon be applying these methods to find "framing language" (Lakoff, 2004) in text, as a means of intuiting the perspective of the author.

Acknowledgments

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. 2007. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):173–180, January.

Michael W. Berry, Susan T. Dumais, and G. W.

Table 3: Training on individual versions (original/shuffled chapters).

	Actual	Shuffled
King James	69.9%	68.4%
Young's Literal	68.2%	70.9%
World English	68.5%	71.2%
Basic English	71.9%	70.9%
Webster	69.9%	69.3%
All 5 merged	70.8%	72.0%

OBrien. 1994. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595.

I. Borg and P. Groenen. 2005. *Modern Multidimensional Scaling, 2nd ed.* Springer-Verlag, New York.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. *Technical Report C-1, The Center for Research in Psychophysiology University.*

Peter A. Chew, Brett W. Bader, Stephen Helmreich, Ahmed Abdelali, and Stephen J. Verzi. 2010. An information-theoretic, vector-space model approach to cross-language information retrieval. *Journal of Natural Language Engineering.*

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *ICDE Workshops*, pages 507–512. IEEE Computer Society.

S. Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers*, 23:229–236.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005.*

George Lakoff. 2004. *Don't Think of an Elephant: Know Your Values and Frame the Debate.* Chelsea Green Publishing. ISBN 978-1931498715.

Bing Liu. 2010. Sentiment analysis and subjectivity. In Indurkhya and Damerau, editors, *Handbook of Natural Language Processing, Second Edition.*

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

P. Resnik, M. Broman Olsen, and M. Diab. 1999. The bible as a parallel corpus: Annotating the "book of 2000 tongues". *Computers and the Humanities*, 33:129–153.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association of Computational Linguistics*, pages 417–424.

P. Young. 1994. Cross language information retrieval using latent semantic indexing. Master's thesis, University of Knoxville, Tennessee, Knoxville, TN.

Table 4: Training on original chapters and testing on shuffled chapters.

English version	Average accuracy
King James	84.6%
Young's Literal	81.7%
World English	85.6%
Basic English	90.7%
Websters	88.3%