The Importance of Neutral Examples for Learning Sentiment

Moshe Koppel and **Jonathan Schler**

Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, Israel
koppel, schlerj@cs.biu.ac.il

Abstract

Most research on learning to identify sentiment ignores "neutral" examples, learning only from examples of significant (positive or negative) polarity. We show that it is crucial to use neutral examples in learning polarity for a variety of reasons. Learning from negative and positive examples alone will not permit accurate classification of neutral examples. Moreover, the use of neutral training examples in learning facilitates better distinction between positive and negative examples.

1 Introduction

The problem of how to exploit a labeled corpus to learn models for sentiment analysis has attracted a good deal of interest in recent years (Dave et al., 2003), (Pang et al., 2002), (Shanahan et al., 2005), (Turney, 2002). One common characteristic of almost all this work has been the tendency to define the task as a two-category problem: positive versus negative. In almost all actual polarity problems, including sentiment analysis, there are, however, at least three categories that must be distinguished: positive, negative and neutral.[1] Not every comment on a product or experience expresses purely positive or negative sentiment. Some – in many cases, most – comments might report objective facts without expressing any sentiment, while others might express mixed or conflicting sentiment.

Researchers are aware, of course, of the existence of neutral documents. The rationale for ignoring them has been a reliance on two tacit assumptions:

- Solving the binary positive vs. negative problem automatically solves the three-category problem since neutral documents will simply lie near the boundary of the binary model
- There is less to learn from neutral documents than from documents with clearly defined sentiment

The purpose of this paper is to show that there is no basis for either of those myths and that neutrals can be exploited in interesting ways to great effect. The outline of the paper is as follows: In Section 2, we will introduce two test corpora corresponding to different types of neutral documents. In Section 3, we will show that neutral documents do not necessarily lie close to the learned positive-negative boundary. In Section 4, we will show that using neutral training documents and standard multi-class learning methods leads to some improvement in classification accuracy but is still sub-optimal. In Section 5, we will show that properly combining the respective models obtained by learning from

[1] As we were completing the final version of this paper, Lillian Lee and Bo Pang kindly made available to us a preprint of their forthcoming paper [8], which considers a number of the problems raised here. In particular, they deal with documents that might be classified according to various degrees of positive or negative sentiment, including neutrality.

pairwise coupling of classes (that is, positive vs. negative, positive vs. neutral, and negative vs. neutral) (Dietterich and Bakiri, 1995), (Fuernkranz, 2002), (Hastie and Tibshirani, 1998) can potentially yield extremely significant improvement in overall classification accuracy.

2 Varieties of Neutrality

We consider two different types of labeled corpora. The first, which we will call the TV corpus, is a collection of posts to chat groups devoted to popular U.S. television shows. These posts have been manually labeled as positive, negative or neutral. We work with 1974 posts equally distributed among positive, negative and neutral documents.

The second corpus consists of 4017 posts to shopping.com's product evaluation pages (http://www.shopping.com)[2] in the areas of digital cameras, strollers and printers. Contributors to these pages have the option of assigning a rating of 1 to 5 to a product under review. We labeled reviews that assigned ratings below 3, exactly 3, and above 3 as negative, neutral, and positive, respectively. The corpus was chosen so that it consists of an equal number of positive, negative and neutral documents.

The neutral documents that appear in the two corpora are of two fundamentally different types. The neutral television chat group posts are generally reports of upcoming or just-seen plots, scheduling announcements or other objective information. The neutral product reviews are generally mixed reviews highlighting both positive and negative features of a given product. As we shall see, the difference between the two different types of neutrality must be borne in mind in exploiting this material.

3 Neutrality and Boundary Distance

First, to establish a baseline for later experiments, we run five fold cross-validation experiments on each corpus, training and testing a linear SVM (using Weka's implementation of SMO (Witten and Frank, 2000)) on positive and negative examples and ignoring neutral examples entirely. The feature set in this experiment, as well as in all experiments in this paper, is the set of all words that appear in the corpus at least 3 times. We obtain accuracy of 67.3% for the TV corpus and 82.7% for the shopping.com corpus.

Is it in fact the case that neutral documents lie near the boundary of a learned model that distinguishes positive and negative examples? To test this, we trained a linear SVM on all positive and negative documents in the TV corpus. In Figure 1a, we show the signed distance from the boundary of the positive and negative training examples (in descending order from left to right). This SVM correctly classifies 79.1% of the training examples. We also show the signed distance from the boundary of all neutral examples. We make several observations:

- The neutral curve does lie between the positive and negative curves (for clarity, this indicates only that the signed distance from the boundary of the neutral document with the n^{th} lowest signed distance from the boundary is in between that of the signed distance from the boundary of the corresponding positive and negative examples)
- The neutral curve is generally closer to the negative curve and most neutral examples are below the boundary. This indicates that neutral documents are more similar to negative documents than to positive documents. One insight into why this is so can be gained by examining the features that distinguish most sharply (using infogain) between positives and negatives: they are almost all positive features. Negative documents, like neutral documents, are distinguished mainly by the absence of positive features.

[2] The TV corpus is property of Trendum Corporation and has not been made publicly available. The shopping.com corpus has been made available to researchers by request. Our thanks to Amir Ashkenazi for his generosity.

- There is no band near the boundary in which the preponderance of examples is neutral. We indicate the band around the boundary that is optimal in terms of overall classification accuracy (positive, negative, or neutral) when all examples in the band are classed as neutral. Even using this optimal band, we attain accuracy of only 54.8%. Note that using an empty band (that is, simply using the SVM boundary to distinguish positive from negative and not classifying any examples as neutral) would yield accuracy of 52.7%. To avoid confusion, it should be understood that these accuracy numbers refer to the training data itself (with the addition of neutrals), and not to a separate test set.

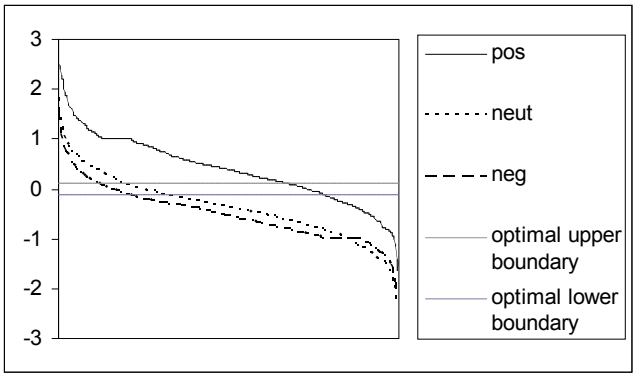

Fig. 1a. Distance from boundary in the TV shows corpus

In Figure 1b we show the results of the same experiment on the shopping.com corpus. In this experiment, the SVM correctly classifies 90.1% of the (positive and negative) training examples. The neutral curve is better centered in this experiment but even choosing the band that maximizes overall accuracy, we obtain accuracy of only 63.0%. Note that using the empty band would yield accuracy of 60.0%.

All in all, there is clear evidence from both corpora that neutral documents cannot be isolated from positive and negative documents simply by using signed distance from the learned positive-negative SVM boundary.

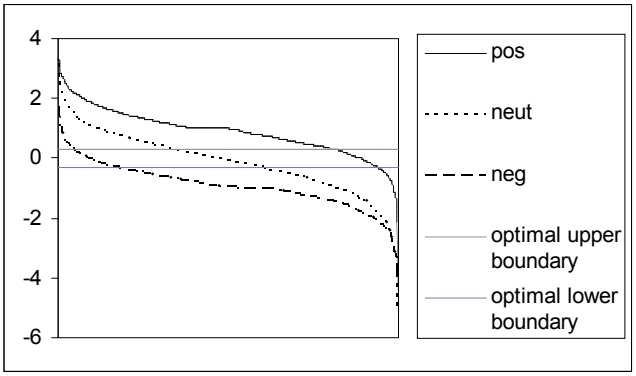

Fig. 1b. Distance from boundary in the shopping.com corpus

4 Learning from Neutrals – Preliminary Attempts

It is evident from the above that if we wish to classify documents as positive, negative or neutral, we will need to use neutral training documents. In this section we consider two straightforward methods for doing so:

1. Multi-class SVMs, treating the three classes as unordered (specifically, Weka's implementation (Witten and Frank, 2000) of the Hastie-Tibshirani algorithm using one-vs-all and one-vs-one models, respectively, as the basis for multi-class learning).

2. Linear regression, treating neutrals as intermediate between positives and negatives

We ran five-fold cross-validation experiments using each of these methods on each of our two corpora. For the TV corpus, multi-class learning based on one-vs-all SVMs yields accuracy of only 52.5%, multi-class learning based on one-vs-one SVMs yields accuracy of 56.4% and linear regression yields accuracy of 69.0%. Note that the latter two results are better than even the optimal results attainable using the boundary method of the previous section. Similarly, for the shopping.com corpus, multi-class learning based on one-vs-all SVMs yields accuracy of 55.1%, multi-class learning based on one-vs-one SVMs yields accuracy of 63.8% and linear regression yields accuracy of 66.3%. Again the latter two results are better than that obtained using the optimal band for isolating neutrals.

It should be noted that this improvement is not attributable simply to the fact of having available more training examples. Even if we use only two thirds of our training examples (so that the total number of training examples is the same as in the previous experiment), we obtain essentially the same results, which are better than the optimum of the boundary method. This is simply the result of the use of neutral examples.

While these results show some improvement over ignoring neutral examples, we shall see that they still do not make optimal use of the neutrals.

5 Optimal Stacks of Binary Classifiers

Let's reflect for a moment on why both regression and multi-class SVM might not properly leverage the neutral examples. In the case of regression, we assume that neutrals are merely intermediate between positives and negatives. In some sense, then, we are not leveraging those aspects of neutral examples that are distinct from positives and negatives but not intermediate.

In the case of multi-class SVM, we need to consider the algorithm used in this experiment for extending a binary algorithm to handle multiple classes, namely, pairwise coupling (Dietterich and Bakiri, 1995), (Fuernkranz, 2002), (Hastie and Tibshirani, 1998). In this approach, a model is learned for each pair of classes (positive-negative, positive-neutral, negative-neutral) and these models are then combined. Note that this method treats the three constituent pairwise problems identically. That is, no allowance is made for the particular relationships that positive, negative and neutral examples stand in to each other.

In this section, we will see that it is crucial to take these special relationships into account. We begin by running the following experiment. For each of the pairs, negative-positive, negative-neutral, and positive-neutral, we ran five-fold cross-validation experiments. For each example, we recorded how it was classed in the holdout set in each of the three experiments.

5.1 The TV Corpus

Table 1a shows the actual class distribution of examples in the TV corpus assigned to each of the eight possible outcomes.

Table 1a. Class distribution of examples per pairwise outcomes in TV corpus

Pos Vs Neg	Pos Vs Neut	Neut Vs neg	original category		
			neg	**neut**	**pos**
Neg	Neut	Neg	**354**	52	
Neg	Neut	Neut	117	**154**	148
Neg	Pos	Neg		**47**	
Neg	Pos	Neut		9	**108**
Pos	Neut	Neg	**145**	69	
Pos	Neut	Neut	42	**225**	46
Pos	Pos	Neg		**90**	
Pos	Pos	Neut		12	**356**

As can easily be computed from the table, the accuracies of the pairwise models in five-fold cross-validation trials on their respective category pairs are: positive-negative, 67.3%; positive-neutral, 73.7%; negative-neutral, 68.5%. Let us consider how we could, in principle, parlay these pairwise models into the best possible three-class model. To do this, let us define a *stack* (Wolpert, 1992) as a mapping from each of the eight possible outcomes to some class. Let an *optimal stack* be the mapping from each of the eight possible outcomes to the majority class of the examples with that outcome.

Savicky and Fuernkranz (2003) have considered when such optimal stacks (determined using holdout data) might permit optimal use of pairwise coupling. They concluded that this kind of stacking is only occasionally effective. We will see that for the polarity problems we consider here, these methods can potentially be quite effective.

For a given example, let's use the shorthand Class1 > Class2 to mean that the learned model of Class1 vs. Class2 classed the example as Class1. The optimal stack for this data can be neatly summarized as follows:

> If positive > neutral > negative then class=positive
> If negative > neutral > positive then class=negative
> Else class=neutral

This simple stack yields accuracy of 74.9% on the three-class problem, which is, somewhat surprisingly, actually better than that obtained for any of the constituent two-class problems. This illustrates that the best way to distinguish positive examples from negative ones is by leveraging the neutrals.

In fact, this stack not only leverages neutral data, it completely ignores the positive-negative model. Any stack that uses the positive-negative will do worse than this stack. One interesting aspect of this stack is that it deviates considerably from majority vote. For example, if both positive and negative defeat neutral, the example is classed as neutral. In this context, that makes some perverse sense: the example likely expresses some mixed sentiment. It is not classed as neutral by either learned model since in this corpus most neutral examples are not mixed but simply express no sentiment.

What is most astonishing about this table is the following: When, according to our model for positive vs. neutral, a test example is classified as positive, it is not necessarily positive, but we *can assert with certainty that it is not negative* (despite not a single negative example being used in training.) Likewise, when, according to our model for negative vs. neutral, a test example is classified as negative, it is not necessarily negative, but we *can assert with certainty that it is not positive* (despite not a single positive example being used in training.)

Of course, we have chosen the optimal stack post hoc. We still need to show that we can use training data to determine a stack that will work well for an out-of-sample test set. To do so we run the following experiment. We run five-fold cross-validation in which for each fold the training data is used twice:

1. Models are learned for positive-negative, positive-neutral and negative-neutral, respectively.

2. Five-fold cross-validation is run within the training set and used to find optimal stacks as described above.

Test examples are then classified by combining the models learned in step 1 according to the stack learned in step 2.

This method yields accuracy of 74.1% which is significantly better than the methods considered above, as illustrated in Figure 2a.

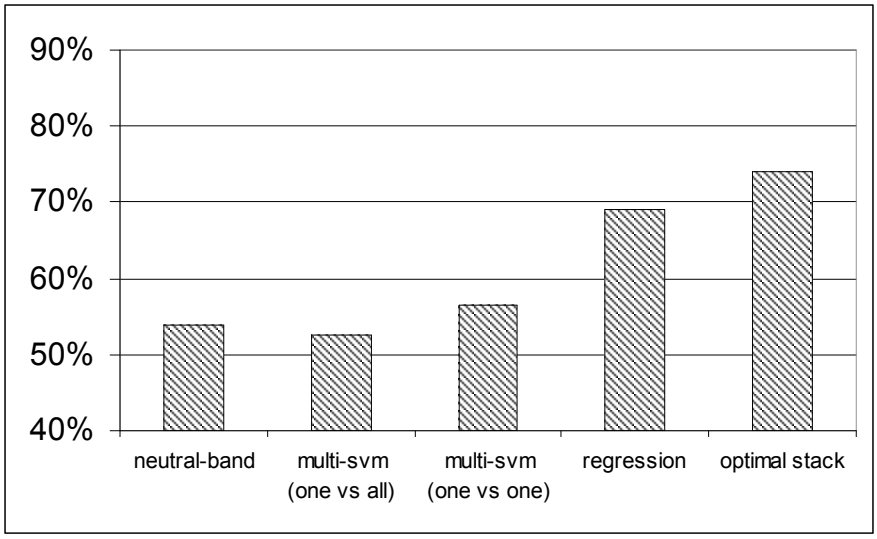

Fig. 2a. Five-fold cross-validation results on the TV corpus using a variety of methods. Optimal neutral band is also shown for comparison.

Moreover, when used in this way, neutral examples also improve results for the problem considered by previous researchers in which all test examples are known to be either positive or negative, but not neutral. We simply adapt our method for choosing the optimal stack so that for each of the eight outcome rows, we choose the class with most examples from among positive and negative only. This method classifies positive and negative test examples (in five-fold cross-validation experiments) with accuracy of 75.1%, which is considerably better than the accuracy of 67.3% obtained by learning SVMs

directly from positive and negative training examples (as seen in Section 3 above). Moreover, this increase is not attributable to the fact that the neutral examples provide us with 50% more training examples. Even when we randomly eliminate one third of the training examples, accuracy on the test set of positives and negatives is 74.3%. We can only conclude that we are better off with a mix of positive, negative and neutral training examples than with only positive and negative training examples, even when our test set is known to contain only positive or negative examples.

It is interesting to speculate that it may be a general property of polarity problems that pairwise coupling ought to be done in a non-standard way: symmetric methods such as simple majority vote may be sub-optimal. We will see that an analogous principle holds in the shopping.com corpus.

5.2 The shopping.com Corpus

Now let us consider the same experiment for the shopping.com corpus (Table 1b).

Table 1b. Class distribution of examples per pairwise outcomes in shopping.com corpus

Pos Vs Neg	**Pos Vs Neut**	**Neut Vs neg**	**original category**		
			neg	**neut**	**pos**
Neg	Neut	Neg	**1043**	243	114
Neg	Neut	Neut	201	**825**	
Neg	Pos	Neg		**60**	59
Neg	Pos	Neut		**211**	
Pos	Neut	Neg	30		**132**
Pos	Neut	Neut	**65**		
Pos	Pos	Neg			
Pos	Pos	Neut			**1034**

As can be computed from the table, the accuracies of the pairwise models in five-fold cross-validation trials on their respective category pairs are: positive-negative, 82.7%; positive-neutral, 71.8%; negative-neutral, 71.0%. The optimal stack for this corpus yields accuracy of 82.3% for the three-class problem.

It is evident, though, that the optimal stack in this case is entirely counter-intuitive. For example, in the case where neutral > positive > negative and neutral > negative, the majority class is the highly unexpected negative. What is astonishing in this table is that we obtain a result oddly analogous to the result obtained on the TV corpus: The best indication that an example is not negative is that positive > neutral. (This is identical to the rule above.) The best indication that an example is not neutral is the fact that positive > negative. As above, in both cases, there are no exceptions.

The optimal stacks in our two corpora are indeed different from each other, a fact that no doubt reflects the differing nature of the neutral examples. But what these optimal stacks have in common is more telling: each is asymmetric in the way it handles the different pairwise models, each displays certain firm, if counter-intuitive, rules, and each classifies with significantly higher accuracy than methods that treat the pairwise models symmetrically or in some fixed relation.

We now run the experiment described in Section 5.1 to determine if we can learn optimal stacks on training data and then apply them successfully to test data. We find that this method yields accuracy of 80.1%, which is far better than all the methods we considered earlier, as shown in Figure 2b.

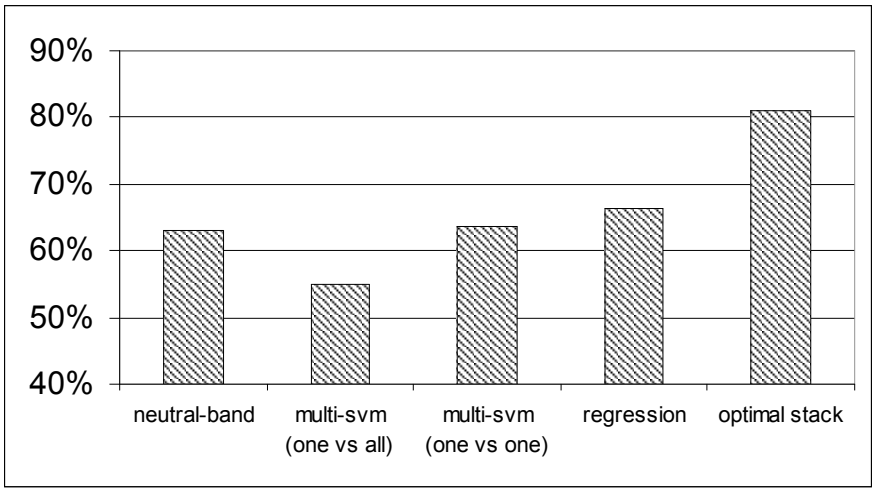

Fig. 2b. Five-fold cross-validation results on the shopping.com corpus using a variety of methods. Optimal neutral band is also shown for comparison.

When this method is applied, with adjustments as described above in the discussion of the TV corpus, for test examples known to be either positive or negative, we obtain accuracy of 85.5%. This is better than accuracy of 82.7% obtained when training on positives and negatives only. Using only 2/3 of the training data, we achieve 84.6% accuracy. Thus, once again we find that a mix of training examples including neutrals is superior to a training set of the same size that consists solely of positives and negatives. The results on this experiment for both corpora are summarized in Figure 3.

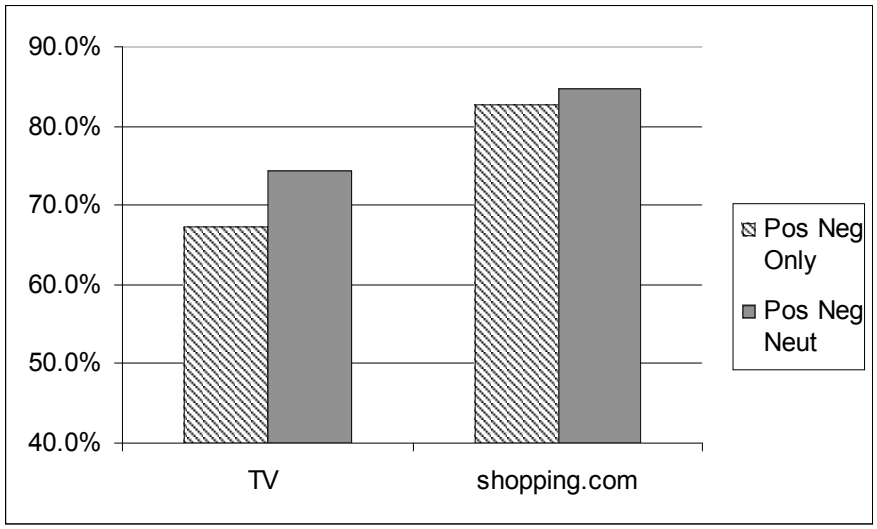

Fig 3 Accuracy on positive and negative test examples, using a training set consisting of positive and negative examples only versus using a training set (of equal size) consisting of positive, negative and neutral examples.

6 Discussion

We have seen that in learning polarity, neutral examples cannot be ignored. Learning from negative and positive examples alone will not permit accurate classification of neutral examples. Moreover, the use of neutral training examples in learning facilitates better distinction between positive and negative examples.

For the case of sentiment analysis, we find that properly combining pairwise learned models leads to extremely significant improvement in overall classification accuracy. The particular method of combination that is most appropriate depends on the nature of the neutral documents in the corpus as well as other considerations. We have found that in one corpus, in which most neutral documents express no sentiment, such neutral documents can be conveniently used as a foil for testing both for negativeness or positiveness and direct positive vs. negative testing can be ignored. When most neutral documents are of mixed sentiment, other stacks might be superior.

More broadly, these results suggest that polarity problems might be best handled as three-class problems using pairwise coupling but combining results in interesting ways. Although (Savicky and Fuernkranz, 2003) found stacking of pairwise couples to provide uneven results, it appears to be just the right approach for polarity problems. Specifically, there may often be optimal counter-intuitive stacks that yield results considerably better than those achievable through voting or related multi-class methods. It remains to be explored if more sophisticated multi-class methods (Crammer and Singer, 2001) might achieve comparable results.

References

Crammer, K. and Y. Singer On the Algorithmic Implementation of Multi-class SVMs, Journal of Machine Learning Research, 2001.

Dave, K., Lawrence, S., and Pennock, D. M., 2003 Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the Twelfth International World Wide Web Conference WWW-2003.

Dietterich, T. G. and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263-286, 1995.

Fuernkranz J. Round Robin Classification, *Journal of Machine Learning Research*, 2:721-747, 2002.

Hastie, T. and R. Tibshirani, Classification by pairwise coupling in M. I. Jordan, M. J. Kearns, and S. A. Solla (eds.) *Advances in Neural Information Processing Systems* 10 (NIPS-97), pp. 507-513. MIT Press, 1998.

Pang, B., Lee, L. and Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002*

Pang, B. and Lee, L., Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with respect to Rating Scales. In *Proceedings of the ACL*, 2005 (to appear).

Savicky, P. And Fuernkranz, J., Combining Pairwise Classifiers with Stacking. in Advances in Intelligent Data Analysis V. (Ed.: Berthold M.R., Lenz H.J., Bradley E., Kruse R., Borgelt Ch.) - Berlin, Springer 2003, pp. 219-229.

Shanahan, James G., Yan Qu, Janyce Wiebe (Eds.) *Computing Attitude and Affect in Text*, Springer, Dordrecht, The Netherlands, 2005

Turney, P. D., Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of ACL 2002,* 417-424.

Witten I. H. and Frank E. "*Data Mining: Practical machine learning tools with Java implementations*" Morgan Kaufmann, San Francisco, 2000

Wolpert, D.H., Stacked Generalization, *Neural Networks*, Vol. 5, pp. 241-259, Pergamon Press, 1992