Performance Measures in Classification of Human Communications

Marina Sokolova and Guy Lapalme

Département d'informatique et de recherche opérationnelle
Université de Montréal
sokolovm@iro.umontreal.ca
lapalme@iro.umontreal.ca

Abstract. This study emphasizes the importance of using appropriate measures in particular text classification settings. We focus on methods that evaluate how well a classifier performs. The effect of transformations on the confusion matrix are considered for eleven well-known and recently introduced classification measures. We analyze the measure's ability to retain its value under changes in a confusion matrix. We discuss benefits from the use of the invariant and non-invariant measures with respect to characteristics of data classes.
Key words: Machine Learning, Evaluation Measures, Text Classification, Human Communication.

1 Introduction

Machine Learning has recently benefited from attention to the *performance measures* used in classification. The interest is supported by the development of new methods and their application in different domains. Evaluation of learning algorithms concentrates on two goals: comparison of algorithms and the applicability of algorithms on a specific domain. Empirical comparison is often done by applying algorithms on one or many data sets and then ranking the performance of the classifiers the algorithms have produced [1].

We focus on measures that evaluate how well a classifier identifies classes, without reference to computational costs or time. Specifically, we address the problem of performance measures for *new types of text classification*. The amount of web-posted texts necessarily invited applications of Data Mining (DM) and Text Data Mining (TDM), Machine Learning (ML) and Natural Language Processing (NLP) techniques. The data became a popular subject of DM, TDM, ML and NLP research through text classification (for a detailed review of the field refer to [2]). However, current studies on text classification undistinguish between texts in general and the data obtained in *human-to-human communication*. This led to leaving characteristics, specific to communications, out of the research scope. As a result, the same measures are used to evaluate classification performance on documents and on records of political debates, e.g., [3] and [4].

In this work we show the that problem of classification of communication records differs from the problem of general text classification. Thus, standard

performance measures for text classification should be re-evaluated with respect to the characteristics of new problems. For this purpose we seek ways to compare various evaluation measures. We suggest a set of changes in a confusion matrix that correspond to specific characteristics of communication textual data. We analyze under what changes a measure retains its value, and therefore preserves the classifier's rank. This is called the measure's invariance under a change. We analyze *measure invariance* for several measures with respect to a transformation of a confusion matrix. Invariance properties identify measure applicability to particular learning settings. The presented analysis is supported by *examples of applications* where invariance properties of measures lead to good ranking of classifiers.

Establishing measure's invariance is one of the main goals of the measurement theory. If a measure is not invariant under the permissible transformations then statistical inference can be applied only to the measure values, but not to the measured attribute [5]. Data Mining has successfully exploited the invariant properties of interestingness measures for comparison of association and classification rules [6–8]. The invariant properties of classification measures recently had been discussed in [9], without specifically referencing them to text classification problems.

2 Human Communications in Text Classification

In recent years NLP and ML communities have turned their attention to studies of *opinions, subjective statements, and sentiments*. Data for these studies are found on chart-boards, blogs, product and movie reviews, and in email, records of phone conversations and political debates, electronic negotiations, etc. These sources represent records of *human communications*. Communication, through the variety of forms, conveys meanings sent by a speaker and received by a hearer. These meanings can be complex and subtly expressed and made up from what is said and what is implied [10].

Success of communication depends on the speaker's ability to produce a message and on the hearer's ability to understand it. Pragmatics, the study of language use, accepts that to be able to infer the meaning of the speaker's message, the hearer expects that the message satisfies standards of the Grice Maxims [11]: Quantity (informativeness), Quality (truthfulness), Relation (relevance) and Manner (clarity). These require the message to be as informative as the situation requires, trustworthy, relevant, clear and brief [10].

Not all communications satisfy Grice Maxims. Sometimes a hidden context interferes with the correct understanding of a message. We present examples of situations where communications and actions come in sharp contradiction. Seemingly successful negotiation given by Table 1 fails because the participant refuses to sign the agreement. Praise for a camera reported in Table 1 is also misleading because the user labels the camera as negative.

Language plays an important role in communication. The language role is critical in a situation when people communicate only verbally, e.g., by phone. In

Communication	Action
Dear XXX, I am YYY, a representative of Such and Such Company. Our company is interested in your [products] . . . Dear XXX, I like your last offer and accept it. Thank you very much for your cooperation.	The participant refuses to sign the agreement.
A great camera! . . . easy to use, viewfinder, flash are good . . . it's a best buy!	The user labels the camera as negative.

Table 1. Examples of situations where communicated meaning contradict actions of interlocutors.

Means	Interaction types		
	one-to-one	one-to-few	one-to-many
Written	Letter	List email	Chart-board message
Verbal	Phone talk	Local radio announcement	Radio broadcast
Visual & verbal	Videophone talk	Video presentation	YouTube video
Face-to-face	Conversation	Lecture	Rally address

Table 2. Examples from communication categories.

exclusively written communication language is the only tool to deliver a message. However, delivery of a message depends on many factors, including

- means, e.g., face-to-face meeting, email;
- topic of discussion, e.g., business, personal;
- time mode, i.e., synchronous or asynchronous;
- interaction mode, determined by the speaker-hearer ratio, e.g., one-to-one, one-to-many;
- speaker-hearer roles, e.g. doctor-patient, buyer-seller, presenter-audience; etc.

We suggest to use a two-dimensional Interaction-Means taxonomy that allows to distinguish between different types of interactions and mediations:

– one-to-many written: chart-boards, blogs, web-posted product and movie reviews;
– one-to-few face-to-face: political debates in the US Congress;
– one-to-few written: list email;
– one-to-one written: electronic negotiations.

Table 2 presents examples of different communication categories. Columns could be added with "few-to-one", ..., "many-to-many" types.

Records of one-to-one and one-to-few communications, e.g., electronic negotiations and email discussions, are used in studies of individual behavior. The aim of such studies is to find what factors influence behavior of a person in a specific situation. Classification of texts depends on the problem statement, e.g. [12, 13].

Transcripts of the US Congress debates are used as a part of fast-growing studies of social networking. Here a common task is to define important influence factors and predict future behavior of members of a social group. In this case, records are classified according to actions of speakers, e.g. [4].

So far, records of one-to-many communications attract more attention and produce more volume of research than other types of communication. These records are studied as evaluative texts, i.e. delivering the author's opinion on the discussed subject. Movie reviews, blogs are often used in sentiment analysis to find whether texts reflect positive or negative opinion of the author on certain products or events. In this case, texts are classified according to opinion/sentiment labels, e.g. [14, 15]. Another popular learning task is to establish strength of the author's opinion, e.g. [16].

3 Text Classification and performance measures

Quality of classification can be assessed using a confusion matrix, i.e., records of correctly and incorrectly recognized examples for each class. Table 3 reports on binary classification, where *tp* are true positive, *fp* – false positive, *fn* – false negative, and *tn* – true negative counts.

Class	Classified	
	as *pos*	as *neg*
pos	*tp*	*fn*
neg	*fp*	*tn*

Table 3. A confusion matrix for binary classification

In text classification, an input text needs to be classified into one (and only one) of j classes (or groups) $C_1, \ldots, C_j$. The existence of the classes is known a priori. Work by Gabrilovich and Markovitch, e.g. [17], exemplifies characteristics of traditional text classification:

1. this is essentially classification of documents, e.g, research papers, technical reports, magazine articles, etc.
2. the main task is topic classification, e.g, identification of documents about Dallas, Texas, or documents about bands and artists, etc.
3. classes are built as relevant vs irrelevant documents, i.e., documents about Dallas, Texas, are distinguished from all other documents; hence, classes are built as positive vs "everything else";
4. retrieval of relevant documents, or a positive class, is the most important task, thus focus is on *tp* classification.

Importance of retrieval of positive examples is reflected by the choice of performance measures for text classification:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$Fscore = \frac{(\beta^2 + 1)tp}{(\beta^2 + 1)tp + \beta^2 fn + fp} \tag{3}$$

$$BreakEvenPoint = \frac{tp}{tp + fp} = \frac{tp}{tp + fn} \tag{4}$$

Three measures evaluate the classifier performance by calculating the ratio of correctly classified positive examples to examples labeled as positives (*Precision*), positive examples in data (*Recall*), and total positive examples, labeled and from data, (*Fscore*). *BreakEvenPoint* essentially estimates when disagreement between data and algorithm labeling of positive examples is balanced ($fp = fn$). All these measures omit *tn* in their formulas, thus do not consider correct classification of negative examples.

Work by Lee *et al*, e.g., [4], concentrates on records of communications and presents direction in text classification started by [14] in 2002:

1. this is classification of political debates, web postings, phone calls, etc., i.e., records of human communications;
2. the main task is non-topic classification, e.g, vote classification, gender classification, mood classification, etc.
3. classes often have distinct features, e.g., male and female, success and failure, etc.; in this case positive and negative classes are both well-defined;
4. retrieval of a positive class, discrimination between classes, balance between retrieval of both classes are possible tasks whose importance depends on the problem at hand.

So far, there is no common understanding on the choice of measures used to evaluate performance of classifiers in opinion, subjectivity, and sentiment analysis. Employed performance measures are either

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn}, \tag{5}$$

which is used in [14, 4] and other works by this group, or *Precision*, *Recall*, *Fscore*, e.g., [18], or correspondence between

$$Sensitivity = \frac{tp}{tp + fn} = Recall \tag{6}$$

and

$$Specificity = \frac{tn}{fp + tn} \tag{7}$$

reported in [13].

With different measures in use, it is important to know how performance evaluations, produced by those measures, relate to each other. Experimental evidence shows that disagreement happens quite often [13].

4 Invariance properties of measures

Finding appropriate measure is possible by establishing *how comparable are the involved measures*. Following [9], we focus on the ability of a measure to preserve its value under a change in a confusion matrix. The invariance of a measure signals that it does not detect this change. Depending on the learning goals, non detection can be beneficial or adverse.

For instance, text classification extensively uses *Precision* and *Recall* (*Sensitivity*). These measures do not detect changes in tn, when all other matrix entries remain the same. In document classification, a large number of unrelated documents constitutes a negative class that lacks unifying characteristics (a multimodal negative class). The criterion for the performance of the classifier is its *performance on related documents* (a well-defined, unimodal, positive class) and may not depend on tn. *Precision* and *Recall* depend on tp, which shows agreement between data and algorithm labeling of positive examples, and fp and fn, which show disagreement between data and algorithm labeling of positive examples. Thus these measures provide the most important perspective on classifiers' performance for document classification. Another emerging application of text classification, classification of consumer reviews, works with highly related documents constituting unimodal positive and negative classes. Thus the evaluation measure may depend on classification of negative examples and reflect the tn change, when other matrix elements stay the same.

We examine the invariance properties with respect to basic changes of a matrix. Our claim is that the following invariance properties affect the measure's applicability and trustworthiness:

Exchange of tp with tn and fn with fp (t1) Table 4 shows the confusion matrix after the changes to the confusion matrix reported in Table 3. A measure is invariant if

$$m(tp, fn, tn, fp) = m(tn, fp, tp, fn) \tag{8}$$

This shows measure permanence with respect to classification results distribution. If the measure is invariant, then it does not distinguish tp from tn and fn from fp and may not recognize asymmetry of classification results. Thus it may not be trustworthy when classifiers are compared on data sets with different and/or unbalanced class distributions. For example, invariant measures may be more appropriate for assessment of classification of consumer reviews then for document classification.

Change of true negative count (t2) Table 5 presents the resulting confusion matrix. A measure is invariant if

$$m(tp, fn, tn, fp) = m(tp, fn, tn', fp) \tag{9}$$

This measure does not recognize specifying ability of classifiers. Such evaluation may be more applicable to domains with a multi-modal negative class, built as "everything not positive". If the measure is non-invariant, has $\overline{t2}$, then it acknowledges ability of classifiers correctly identify negative examples. If the measure is able to do this, it may be reliable for comparison

Class	Classified	
	as *pos*	as *neg*
as *pos*	*tn*	*fp*
as *neg*	*fn*	*tp*

Table 4. Confusion matrix after the exchange of *tp* with *tn* and *fn* with *fp*.

Class	Classified	
	as *pos*	as *neg*
pos	*tp*	*fn*
neg	*fp*	tn'

Table 5. Confusion matrix after a change in true negative count.

in domains with a well-defined, unimodal, negative class. In case of text classification, these invariant measures are suitable for evaluation of document classification and non-invariant measures are preferable for evaluation of such communications where criteria exist for positive as well as for negative results.

Change of a false count (t3) Table 6 reports the confusion matrix. A measure is invariant if

$$m(tp, fn, tn, fp) = m(tp, fn, tn, fp') \tag{10}$$

t3 indicates measure constancy if disagreement increases between the data and classifier labels. An invariant measure shows preference for data labels. In case of unreliable data labeling such measure may give misleading results. A non-invariant measure may not be suitable for data with many counter examples. If classifier ranking improves when *fp* increases, the measure may favor a classifier prone to faux positives. In case of t3, the use of invariant and non-invariant measures might be decided based on problem and data characteristics. This is especially important for problems in sentiment classification of blogs, charts, consumer reviews, where some data do not have consistent labels because of the absence of rigorous labeling rules, and in classification of records of long-term communications, where some data have a substantial number of counter-examples.

Classification scaling (t4) Table 7 presents the confusion matrix. A measure is invariant if

$$m(tp, fn, tn, fp) = m(k_1 tp, k_2 fn, k_2 tn, k_1 fp) \tag{11}$$

This shows measure uniformity with respect to proportional changes of classification results. If the measure is non-invariant, then its applicability may depend on class sizes. If we expect that for different data sizes the same portion of examples exhibits positive (negative) characteristics, then the invariant measure may be a better choice for classifiers' evaluation. The non-invariant measures may be more reliable if we do not know how representative is the data sample in terms of proportion positive/negative examples (which is might be the case in web-posted consumer reviews).

Class	Classified	
	as *pos*	as *neg*
pos	*tp*	*fn*
neg	fp'	*tn*

Table 6. Confusion matrix after a change in false positive count.

Class	Classified	
	as *pos*	as *neg*
pos	$k_1 tp$	$k_2 fn$
neg	$k_1 fp$	$k_2 tn$

Table 7. Confusion matrix after scaling.

5 Empirical evidence

Application on "real life" communication data supports our claim on necessity of measure comparison. The data are the records of human-to-human electronic negotiations, where a buyer and a seller try to reach an agreement on virtual purchase of commercial products. A negotiation is successful when agreement is reached, otherwise it is failed. Support Vector Machine(SVM) and Naive Bayes (NB) have been applied on the same data set. Tables 8 and 9, adapted from [13], report their confusion matrices. The matrices are representative in a sense that changes in data representation do not statistically affect SVM and NB performance and, consequently, *tp*, *fn*, *fp*,*tn*.

Class	Classified	
	as *pos*	as *neg*
pos	1242	189
neg	390	740

Table 8. Confusion matrix for SVM.

Class	Classified	
	as *pos*	as *neg*
pos	1108	323
neg	272	858

Table 9. Confusion matrix for NB.

We apply several measures to rank the classifiers, starting with the listed in Section 3 evaluators. Except these evaluators, the employed measures include the Area Under Curve (AUC), calculated for one run of the Receiver Operating Characteristic

$$AUC = \frac{1}{2}\left(\frac{tp}{tp+fn} + \frac{tn}{tn+fp}\right) \tag{12}$$

likelihoods ρ_+, ρ_- that are frequently used for comparison of diagnostic tests [19],

$$\rho_+ = \frac{tp(tn+fp)}{fp(tp+fn)} \tag{13}$$

$$\rho_- = \frac{fn(tn+fp)}{tn(tp+fn)}. \tag{14}$$

Huang and Ling [20] newly introduced a combined measure that they denote as $AUC:acc$

$$AUC:acc = \frac{Sensitivity + Specificity}{2 \cdot Accuracy} \tag{15}$$

First four columns of Table 10 report measure values and the classifier ranks. *Recall* is omitted because of co-linearity with *Sensitivity*. *Fscore* is used with $\beta = 1$. In four "Invariance" columns "+" and "-" denote invariance and non-invariance respectively on our data.

Measure	Empirical evidence				Invariance				Cluster		
	SVM		NB		t1	t2	t3	t4	1	2	3
	%	rank	%	rank							
Accuracy	77.4	1	76.8	2	+	-	-	-			√
Sensitivity	86.8	1	77.5	2	-	+	+	-		√	
Specificity	65.4	2	75.9	1	-	-	-	-	√		
Precision	76.0	2	80.1	1	-	+	-	+		√	
Fscore	81.2	1	78.9	2	-	+	-	-		√	
BreakEvenPoint	74.0	1	72.4	2	-	+	-	-		√	
AUC	52.3	2	53.4	1	-	-	-	-	√		
ρ_+	2.51	2	3.22	1	-	-	-	-	√		
ρ_-	0.20	1	0.30	2	-	-	-	-	√		
$AUC : acc$	98.3	2	99.8	1	-	-	-	-	√		

Table 10. Empirical comparison, invariance properties and clustering of performance measures.

We also emphasize that measure's focus on *tp* does allow to evaluate how well a classifier deals with the specific problems of human communication data (refer to examples given by Table 1). Most probably, those examples could be falsely classified as positives.

6 Analysis of results

The invariant properties, introduced in Section 4, divide the measures into *three clusters*. One cluster is constructed from measures non-invariant under the four matrix transformations. *Specificity*, AUC, ρ_+, ρ_- and $AUC : acc$ change their values under all the considered changes in a confusion matrix.

- The first non-invariance, $\overline{t1}$, means that the measures are sensitive to asymmetry of classification. This is a well-known characteristic for *Specificity*, but not for the other four measures that have been recently introduced to classification. The non-invariance may explain why $AUC : acc$ is more reliable than *Accuracy* when used for classifiers' assessment on imbalanced data [20].
- The second non-invariance, $\overline{t2}$, signals that the use of the measures is more appropriate on data with a unimodal negative class than with a multi-modal one. This implication is more important for AUC and $AUC : acc$ than for *Specificity* and ρ_+, ρ_-. The latter are usually used in combinations with other measures, whereas the former might be applied separately.

– The third non-invariance, $\overline{t3}$, shows that the measures may be resistant to unreliable data labeling. To find out whether the measures may favor a classifier with a poor ability of detecting counterexamples, we have to check if ranking increases when *fp* increases. This is not true for ranking produced by *Specificity*. Rankings produced by the other four measures are not monotonic under the property assumptions.
– The last non-invariance, $\overline{t4}$, indicates that the measures may not be comparable when used on data with considerably different sizes.

The remaining five measures can be naturally categorized into *Accuracy* and the *Fscore* group (*Precision*, *Recall* (*Sensitivity*), *Fscore* and *BreakEvenPoint*). All the *Fscore* group measures are invariant under the change of *tn*. This well-known property have made them a tool of choice for evaluations of document classification. Within this group, *Precision* is invariant under scaling, *Recall* (*Sensitivity*) – under the change of *fp* and *Fscore* and *BreakEvenPoint* have identical invariance properties ($\overline{t1}$, $t2$, $\overline{t3}$, $\overline{t4}$). The *Accuracy*'s only invariance, $t1$, has been much discussed in Machine Learning community. The last three columns of Table 10 represent three clusters containing the measures that correspond to the check marks ($\surd$) in the lines.

Invariance with respect to the matrix transformations is especially important because it connects evaluation measures to particular learning settings. We summarize *applicability of measures* to subfields of text classification: document classification and classification of human communications. The initial assumption would be to apply *Fscore* measures as the most suitable for text classification evaluation. However, subfields' classification problems exhibit different characteristics. That may require applications of different evaluation measures. Based on the analysis of invariance properties of measures we propose the following:

– Document classification data are usually highly imbalanced. Relevant documents construct a small well-defined positive class, a populous negative class is built from non-relevant documents as "everything non-positive". Presence of a multi-modal negative class favors the use of the *Fscore* measures.
– Classification of human communications often is mostly represented by sentiment classification where data are collections of free form texts of product evaluations. Proportion of positive and negative examples depends on the popularity of a product. Positive and negative classes are well-defined. Due to presence of a unimodal negative class, *Sensitivity* and *Specificity* may provide more reliable classifier ranking than *Precision* and *Recall* (*Sensitivity*). $AUC : acc$ may be preferable over *Accuracy* if there is a class imbalance. However, other measures might be suitable for classification of communications in social activities, such as political debates or electronic negotiations. If data have a unimodal negative class and a large number of counter examples, as in records of electronic negotiations, *Accuracy*, *Precision*, *Recall* (*Sensitivity*), and *Specificity* may be used for reliable classification ranking.

7 Conclusions and future work

We have analyzed applicability of performance measures to different subfields of text classification. We have shown that document classification differs from classification of human communications, thus that these two types of text classification may require different set of performance measures.

We have shown that the results of the classifier comparison depend on a number of factors, including invariant properties of the measures. We have considered effects of various transformations of the confusion matrix on several well-known performance measures. The invariance properties have lead to fine distinctions of relations between the measures and the data characteristics. One way to insure reliable evaluation is to employ a measure corresponding to the learning setting. The next step would be to expand the list of connections between learning settings and evaluation measures.

This approach opens new directions for future work. First, we built a framework for the *two-dimensional* relations "measure vs invariance" and omitted decision theory relations. Note that the listed measures evaluate different decision aspects of the classifier performance. Given below is a condensed description from [19, 1]:

- *Accuracy*, *Recall* (*Sensitivity*), *Specificity* show how effectively a classifier identifies the data labels;
- *Precision* estimates the class agreement of the data labels with the labels given by the classifier;
- AUC indicates the classifier's ability to avoid false classification;
- ρ_+ and ρ_- assess prediction ability on positive and negative classes respectively.

Combining the decision aspects with the existing framework leads to constructing a *three-dimensional* "measure vs invariance vs decision aspect" taxonomy of measures.

Next, this study focuses on binary classification. A natural way to extend it is to apply similar systematization to *multi-class classification*. Multi-class extension is desirable because many Machine Learning applications switch from binary "positive vs everything else" to finer grained problems. For example, in sentiment analysis, opinion mining and other subfields of subjectivity analysis three-class classification problems gradually substitute binary classification problems.

Next, our study concentrates on text classification, but it can be expanded to other language applications of Machine Learning. Machine Translation and Natural Language Processing are other examples of the fields where the discussed measures, e.g., *Fscore*, are used for comparison of classifiers.

Acknowledgments This work has been funded by the Natural Sciences and Engineering Research Council of Canada.

References

1. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30
2. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1) (2002) 1–47
3. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proc 21st International Conf on Machine Learning ICML'04. (2004) 489–495
4. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. (2006) 327–335
5. Sarle, W.S.: Measurement theory: Frequently asked questions. In: the Disseminations of the International Statistical Applications Institute. ACG Press (1996) 61–66
6. Geng, L., Hamilton, H.: Interestingness measures for data mining: A survey. ACM Computing Surveys **38**(3) (2006)
7. Lallich, S., Teytaud, O., Prudhomme, E.: Association rules interestingness: measure and validation. In F., G., J., H.H., eds.: Quality Measures in Data Mining. Springer (2006) –
8. Tan, P., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems **29**(4) (2004) 293–313
9. Sokolova, M.: Assessing invariance properties of evaluation measures. In: Proceedings of the Workshop on Testing of Deployable Learning and Decision Systems, the 19th Neural Information Processing Systems Conference (NIPS 2006). (2006)
10. Leech, G.: Principles of Pragmatics. second edn. Longman (1991)
11. Grice, P.: Studies in the Way of Words. Harvard University Press (1989)
12. Boparai, J., Kay, J.: Supporting user task based conversations via email. In: Proc 7th Australasian Document Computing Symposium. (2002)
13. Sokolova, M.: Learning from Communication Data: Language in Electronic Business Negotiations. Ph.D. dissertation. (2006)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proc Empirical Methods of Natural Language Processing EMNLP'02. (2002) 79–86
15. Mishne, G.: Experiments with mood classification in blog posts. In: Proc 1st Workshop on Stylistic Analysis of Text for Information Access (Style2005). (2005) staff.science.uva.nl/gilad/pubs/style2005-blogmoods.pdf.
16. Wilson, T., Wiebe, J., Hwa, R.: Recognizing strong and weak opinion clauses. Computational Intelligence **22**(2) (2006) 7399
17. Gabrilovich, E., Markovitch, S.: Text categorization with many redundant features: Using aggressive feature selection to make svms competitive with c4.5. In: Proc 21st International Conf on Machine Learning ICML'04. (2004) 321–328
18. Gamon, M.: Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004). (2004) 841–847
19. Biggerstaff, B.: Comparing diagnostic tests: a simple graphic using likelihood ratios. Statistics in Medicine **19**(5) (2000) 649–663
20. Huang, J., Ling, C.: Constructing new and better evaluation measures for machine learning. In: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'2007). (2007) –