Information Processing and Management 45 (2009) 427–437

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A systematic analysis of performance measures for classification tasks

Marina Sokolova [a,*], Guy Lapalme [b]

[a] Electronic Health Information Lab, Children's Hospital of Eastern Ontario, Ottawa, Canada
[b] Département d'informatique et de recherche opérationnelle Université de Montréal, Montréal, Canada

ARTICLE INFO

Article history:
Received 14 February 2008
Received in revised form 21 November 2008
Accepted 6 March 2009
Available online 8 May 2009

Keywords:
Performance evaluation
Machine Learning
Text classification

ABSTRACT

This paper presents a systematic analysis of twenty four performance measures used in the complete spectrum of Machine Learning classification tasks, i.e., binary, multi-class, multi-labelled, and hierarchical. For each classification task, the study relates a set of changes in a confusion matrix to specific characteristics of data. Then the analysis concentrates on the type of changes to a confusion matrix that do not change a measure, therefore, preserve a classifier's evaluation (*measure invariance*). The result is the measure invariance taxonomy with respect to all relevant label distribution changes in a classification problem. This formal analysis is supported by examples of applications where invariance properties of measures lead to a more reliable evaluation of classifiers. Text classification supplements the discussion with several case studies.

1. Motivation

Machine Learning (ML) divides classification onto binary, multi-class, multi-labelled, and hierarchical tasks. In this work we present a systematic analysis of twenty four performance measures used in these classification subfields. We focus on how well classes are identified without reference to computation cost or time. We consider a set of changes in a confusion matrix that correspond to specific characteristics of data. We then analyze the type of changes that do not change a measure's value and therefore preserve a classifier's evaluation. This is what we call *measure invariance*. As a result, we build the measure invariance taxonomy with respect to all relevant label distribution changes in a classification problem. We supplement the formal analysis by examples of applications where invariance properties of measures lead to a more reliable evaluation of classifiers; examples are taken from text classification. Note, that we focus on recent ML developments; more details on ML measures can be found, for example, in Sokolova, Japkowicz, and Szpakowicz (2006) which looks into relations between the measures and assessment of medical trials. To the best of our knowledge, our current study is the first reviews of ML measures which comprehensively evaluates the invariant properties of measures. Preliminary results on binary classification appear in (Sokolova & Lapalme, 2007). This study expands the results two-fold, with discussion of new invariant properties, in some cases, adding monotonicity properties, and consideration of multi-class, multi-labelled, and hierarchical measures.

Empirical evaluation remains the most used approach for the algorithm assessment, although ML algorithms can be evaluated through empirical assessment or theory or both, e.g., derived generalized bounds and empirical results (Marchand & Shawe-Taylor, 2002). Evaluation techniques based on multiple experiments are considered in Dietterich (1998), one of the most cited work on empirical evaluation of ML algorithms. An extensive critique of ML evaluation practice can be found in Salzberg (1999). The author analyzes the currently used methods and their statistical validity. The paper distinguishes two goals of evaluation: a comparison of algorithms, and the feasibility of algorithms on a specific domain. Demsar (2006) surveys how classifiers are compared over multiple data sets. Empirical comparison is most often done by applying algorithms on various data sets and then evaluating the performance of the classifiers that the algorithms have produced; accuracy

* Corresponding author.
E-mail addresses: msokolova@ehealthinformation.ca (M. Sokolova), lapalme@iro.umontreal.ca (G. Lapalme).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.03.002

being the most often used measure. In all these assessment approaches, the algorithm and the output classifiers take the central stage.

We take an alternative route looking how characteristics affect the objectivity of measures. Our formal discussion of ML performance measures complements popular statistical and empirical comparisons such as the ones presented in Goutte and Gaussier (2005). We show that, in some learning settings, the correct identification of positive examples may be important whereas in others, the correct identification of negative examples or disagreement between data and classifier labels may be more significant. Thus, standard performance measures should be re-evaluated with respect to those scenarios. Previously, ML studies of performance measures have primarily focused on binary classification. For a complete review, we add multi-class, multi-topic and hierarchical classification measures. The current study can be useful for measure design. So far, the ML community did not consider measures' invariance when new ones were introduced (Bengio, Mariéthoz, & Keller, 2005; Huang & Ling, 2007) or suggested for adoption from other disciplines (Sokolova et al., 2006).

2. Overview of classification tasks

Supervised ML allows access to the data labels during the algorithm's training and testing stages. Consider categorical labels when data entries $x_1, \ldots, x_n$ have to be assigned into predefined classes $C_1, \ldots, C_l$. Then classification falls into one of the following tasks:

Binary: the input is to be classified into one, and only one, of two non-overlapping classes (C_1, C_2); Binary classification is the most popular classification task. Assigned categories can be objective, independent of manual evaluation (e.g, republican or democrat in the votes data of the UCI repository (Asuncion & Newman, 2007)) or subjective, dependent on manual evaluation (e.g., positive or negative reviews in Amazon.com (Blitzer et al., 2007)). Classes can be well-defined (e.g., the votes labels), ambiguous (e.g., the review opinion labels), or both (e.g., medical vs. other texts in the Newsgroups collection[1]).

Multi-class: the input is to be classified into one, and only one, of l non-overlapping classes. Multiclass problems include the identification of the iris type in a three-class data set popular in pattern recognition (Duda & Hart, 1973), in the learning the original 135 categories in the benchmark Reuters collection,[2] or in tagging utterances as objective, subjective, or neutral (Wilson, Wiebe, & Hwa, 2006). As for the binary case, multi-class categorization can be objective or subjective, well-defined or ambiguous.

Multi-labelled: the input is to be classified into several of l non-overlapping C_j. Examples include classification of functions of yeast genes (Mewes, Albermann, Heumann, Lieb, & Pfeiffer, 1997), identifying scenes from image data (Li, Zhang, & Zhu, 2006) or text-database alignment and word alignment in machine translation (Snyder & Barzilay, 2007). In text mining of medical information, multi-label classification methods are often evaluated on OHSUMED, a collection of medical references (Hersh, Buckley, Leone, & Hickam, 1997). When the learning task is document topic classification, multi-labelling is often referred as multi-topic classification such as for clinical texts that are assigned multiple disease codes from ICD-9-CM (Sasaki, Rea, & Ananiadou, 2007). Binary, multi-class, and multi-labelled problems form flat classification (Yang, 1999), in which categories are isolated and their relations are not considered important. The next, hierarchical, problem addresses relations among categories and includes their structure into learning targets.

Hierarchical: the input is to be classified into one, and only one, C_j which are be divided into subclasses or grouped into superclasses. The hierarchy is defined and cannot be changed during classification. Text classification and bioinformatics supply many examples, e.g., protein function prediction (Eisner, Poulin, Szafron, Lu, & Greiner, 2005). Hierarchical classification can be transformed into flat classification. For example, the Reuters collection classification can be multi-class (Bobicev & Sokolova, 2008), multi-labelled (Tikk & Biró, 2003), and hierarchical (Sun, Lim, & Ng, 2003).

A frequent appearance of language and text problems among the listed above examples can be explained by a special role Natural Language Processing (NLP) holds in ML applications. The richness of language characteristics and the fast-increasing volume of readily available digital texts make texts not only a nearly inexhaustible research area, but also one of the most important data formats for ML applications (Shawe-Taylor & Christianini, 2004). Text Classification has achieved a prominent place among ML applications to NLP problems. It is dedicated to finding texts according to a given criteria (Sebastiani, 2002) and it includes the classification of documents (research papers, technical reports, magazine articles, etc.). For topic classification (e.g., identification of documents about a given city or documents about bands and artists, etc.) documents are simply classified as being relevant to the topic or not; hence, classes are built as positive vs "everything else". Retrieval of relevant documents being the more important task, the focus in this case is on true positive classification. First comprehensive books on Machine Learning were published in late 1990's (Langley, 1996; Mitchell, 1997). As a discipline, ML borrowed measures from assortment of disciplines traditionally relied on empirical evidence, e.g., medical trials (Isselbacher & Braunwald, 1994), behavioural research (Cohen, 1988), information retrieval (IR) (van Rijsbergen, 1979; Salton & McGill, 1983). In some ways, text classification borrows from Information Extraction (IE) which preluded the use of Machine

[1] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.
[2] http://www.daviddlewis.com/resources/testcollections/reuters21578.

Table 1
Confusion matrix for binary classification and the corresponding array representation used in this paper.

Data class	Classified as pos	Classified as neg
pos	true positive (tp)	false negative (fn)
neg	false positive (fp)	true negative (tn)

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

Learning in automated text processing and understanding, e.g., the automated analysis and generation of synonymous texts (Boyer & Lapalme, 1985). IE and IR metrics in the evaluation of ML algorithms are an example of such borrowing. The evaluation metrics commonly used in Text Classification (*Precision*, *Recall*, *Fscore*) have their origin in IE. The formulas for these measures neglect the correct classification of negative examples, they instead reflect the importance of retrieval of positive examples in text/document classification:

Precision: the number of correctly classified positive examples divided by the number of examples labeled by the system as positive
Recall: the number of correctly classified positive examples divided by the number of positive examples in the data
Fscore: a combination of the above.

In recent years, the NLP and ML communities have turned their attention to the study of opinions, subjective statements, and sentiments. The corresponding empirical problems are represented by the classification of political debates, web postings or phone calls in which the main task is non-topic classification, e.g. vote classification, gender classification, sentiment classification, etc. Data for these studies are gathered from chart-boards, blogs, product and movie reviews, email, records of phone conversations and political debates, electronic negotiation transcripts, etc. Chart-boards, blogs and movie reviews are often used in sentiment analysis to find whether texts reflect a positive or negative opinion of the author on certain products or events. In this case, texts are classified according to opinion/sentiment labels (Pang, Lee, & Vaithyanathan, 2002). Email discussions, records of phone conversation and electronic negotiation transcripts are used in studies of individual behavior. The aim of such studies is to find what factors influence the behavior of a person in a specific situation. Classification of texts depends on the problem statement. Transcripts of the US Congress debates are used in the social network analysis, a new area of Artificial Intelligence research. Here a common task is to define important influence factors and predict the future behavior of members of a social group. In this case, records are classified according to the actions of speakers (Thomas, Pang, & Lee, 2006).

These sources represent records of *human communication* that convey meanings sent by a speaker and received by a hearer. These meanings can be complex and subtly expressed and constituted from both what is said and what is implied. So far, there is no common consensus on the choice of measures used to evaluate the performance of classifiers in opinion, subjectivity and sentiment analysis. Additional performance measures other than the above are *Accuracy* used in Pang et al. (2002) and Thomas et al. (2006), or the correspondence between *Precision* and *Recall* in Gamon, Aue, Corston-Oliver, and Ringger (2005). When going from document classification to the classification of human communication, it is important to know how different performance measures relate to each other in order to help resolve disagreements among performance evaluations. This phenomenon happens quite often in experimental studies.

3. Performance measures for classification

The correctness of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). These four counts constitute a confusion matrix shown in Table 1 for the case of the **binary classification**.

Table 2 presents the most often used measures for binary classification based on the values of the confusion matrix. *AUC* (*Area Under the Curve*),[3] captures a single point on the *Reception Operating Characteristic* curve; it can also be viewed as a linear transformation of *Youden Index* (Youden, 1950). We omit measures such as *BreakEvenPoint*, the point at which *Precision* = *Recall* (Goutte & Gaussier, 2005), and the combined $\frac{AUC}{Accuracy}$ (Huang & Ling, 2007) because their properties can be derived from the basic measures. However, we present *Fscore*'s properties because of its extensive use in text classification.

Table 3 presents the measures for **multi-class classification**. For an individual class C_i, the assessment is defined by tp_i, fn_i, tn_i, fp_i. $Accuracy_i, Precision_i, Recall_i$ are calculated from the counts for C_i. Quality of the overall classification is usually

[3] *AUC*, sometimes referred to as *Balanced Accuracy*.

Table 2
Measures for binary classification using the notation of Table 1.

Measure	Formula	Evaluation focus
Accuracy	$\frac{tp+tn}{tp+fn+fp+tn}$	Overall effectiveness of a classifier
Precision	$\frac{tp}{tp+fp}$	Class agreement of the data labels with the positive labels given by the classifier
Recall (*Sensitivity*)	$\frac{tp}{tp+fn}$	Effectiveness of a classifier to identify positive labels
Fscore	$\frac{(\beta^2+1)tp}{(\beta^2+1)tp+\beta^2 fn+fp}$	Relations between data's positive labels and those given by a classifier
Specificity	$\frac{tn}{fp+tn}$	How effectively a classifier identifies negative labels
AUC	$\frac{1}{2}\left(\frac{tp}{tp+fn}+\frac{tn}{tn+fp}\right)$	Classifier's ability to avoid false classification

Table 3
Measures for multi-class classification based on a generalization of the measures of Table 1 for many classes C_i: tp_i are true positive for C_i, and fp_i – false positive, fn_i – false negative, and tn_i – true negative counts respectively. μ and M indices represent micro- and macro-averaging.

Measure	Formula	Evaluation focus
Average Accuracy	$\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class effectiveness of a classifier
Error Rate	$\frac{\sum_{i=1}^{l}\frac{fp_i+fn_i}{tp_i+fn_i+fp_i+tn_i}}{l}$	The average per-class classification error
$Precision_\mu$	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$	Agreement of the data class labels with those of a classifiers if calculated from sums of per-text decisions
$Recall_\mu$	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions
$Fscore_\mu$	$\frac{(\beta^2+1)Precision_\mu Recall_\mu}{\beta^2 Precision_\mu+Recall_\mu}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions
$Precision_M$	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fp_i}}{l}$	An average per-class agreement of the data class labels with those of a classifiers
$Recall_M$	$\frac{\sum_{i=1}^{l}\frac{tp_i}{tp_i+fn_i}}{l}$	An average per-class effectiveness of a classifier to identify class labels
$Fscore_M$	$\frac{(\beta^2+1)Precision_M Recall_M}{\beta^2 Precision_M+Recall_M}$	Relations between data's positive labels and those given by a classifier based on a per-class average

assessed in two ways: a measure is the average of the same measures calculated for $C_1,\ldots,C_l$ (*macro-averaging* shown with an M index), or the sum of counts to obtain cumulative tp, fn, tn, fp and then calculating a performance measure (*micro-averaging* shown with μ indices). Macro-averaging treats all classes equally while micro-averaging favors bigger classes. As there is yet no well-developed multi-class *Reception Operating Characteristic* analysis (Lachiche & Flach, 2003), we do not include *AUC* in the list of multi-classification measures.

The quality of **multi-topic classification** (Table 4) is assessed through either partial or complete class label matching (Kazawa, Izumitani, Taira, & Maeda, 2005); the latter is often referred to as exact matching. We consider all classes and their labels as being equivalent. These measures thus count correct or incorrect label identification independently of their order or rank. We do not include such measures as *One-error* which counts how many times the top-ranked label was not a member of the predicted label set (Li et al., 2006). Some authors refer to it Exact Match Ratio as *Accuracy* (Zhu, Ji, Xu, & Gong, 2005). In Section 4, we show that these two measures are not interchangeable with respect to confusion matrix transformations; thus, they may not be equally applicable to similar settings.

For **hierarchical classification** measures (Table 5), we consider measures that incorporate the problem's hierarchy. These measures either evaluate descendant or ancestor performance (Kiritchenko, Matwin, Nock, & Famili, 2006). We omit

Table 4
Measures for multi-topic classification; I is the indicator function; $L_i = L_i[1],\ldots,L_i[l]$ denotes a set of class labels for x_i, $L_i[j]=1$ if C_j is present among the labels and 0, otherwise; L_i^c are labels given by a classifier, L_i^d are the data labels.

Measure	Formula	Evaluation focus
Exact Match Ratio	$\frac{\sum_{i=1}^{n} I(L_i^d=L_i^d)}{n}$	The average per-text exact classification
Labelling Fscore	$\frac{\sum_{i=1}^{n}\frac{2\sum_{j=1}^{l} L_i^c[j]L_i^d[j]}{\sum_{j=1}^{l}(L_i^c[j]+L_i^d[j])}}{n}$	The average per-text classification with partial matches
Retrieval Fscore	$\frac{\sum_{j=1}^{l}\frac{2\sum_{i=1}^{n} L_i^c[j]L_i^d[j]}{\sum_{i=1}^{n}(L_i^c[j]+L_i^d[j])}}{l}$	The average per-class classification with partial matches
Hamming Loss	$\frac{\sum_{i=1}^{n}\sum_{j=1}^{l} I(L_i^c[j]\neq L_i^d[j])}{nl}$	The average per-example per-class total error

Table 5
Measures for hierarchical classification: $C_{\downarrow}$ means subclasses of class C, $C^c_{\downarrow}$ denotes subclasses assigned by a classifier; $C^d_{\downarrow}$ – data class labels; similar notations apply to superclasses, which are denoted by $C_{\uparrow}$.

Measure	Formula	Evaluation focus
$Precision_{\downarrow}$	$\frac{\lvert C^c_{\downarrow} \cap C^d_{\downarrow}\rvert}{\lvert C^c_{\downarrow}\rvert}$	Positive agreement on subclass labels w.r.t. the subclass labels given by a classifier
$Recall_{\downarrow}$	$\frac{\lvert C^c_{\downarrow} \cap C^d_{\downarrow}\rvert}{\lvert C^d_{\downarrow}\rvert}$	Positive agreement on subclass labels w.r.t. the subclass labels given by data
$Fscore_{\downarrow}$	$\frac{(\beta^2+1)Precision_{\downarrow}Recall_{\downarrow}}{\beta^2 Precision_{\downarrow}+Recall_{\downarrow}}$	Relations between data's positive subclass labels and those given by a classifier
$Precision_{\uparrow}$	$\frac{\lvert C^c_{\uparrow} \cap C^d_{\uparrow}\rvert}{\lvert C^c_{\uparrow}\rvert}$	Positive agreement on superclass labels w.r.t. the superclass labels given by a classifier
$Recall_{\uparrow}$	$\frac{\lvert C^c_{\uparrow} \cap C^d_{\uparrow}\rvert}{\lvert C^d_{\uparrow}\rvert}$	Positive agreement on superclass labels w.r.t. the superclass labels given by data
$Fscore_{\uparrow}$	$\frac{(\beta^2+1)Precision_{\uparrow}Precision_{\uparrow}}{\beta^2 Precision_{\uparrow}+Precision_{\uparrow}}$	Relations between data's positive superclass labels and those given by a classifier

distance- and semantics-based measures suggested for hierarchical classification (Blockeel, Bruynooghe, Dzeroski, Ramon, & Struyf, 2002; Sun et al., 2003). These measures extend flat, non-hierarchical, measures by estimating differences and similarities between classes. However, in these measures, acceptable differences and similarities are often specified by users (Costa, Lorena, Carvalho, & Freitas, 2007). Thus, the obtained results may be subjective and user-specific. A similar restriction applies to depth-dependent measures, which relate classes by imposing vertical distances (Blockeel et al., 2002).

Data Mining has successfully exploited the invariant properties of interestingness measures for comparison of association and classification rules (Tan, Kumar, & Srivastava, 2004). Some invariant properties of binary classification measures have been discussed within broader studies of the classification of communication records (Sokolova & Lapalme, 2007). In the current study, we consider new invariant properties and expand discussed measures by including multi-class, multi-topic and hierarchical classification measures. Although the latter three types of classification are quite popular, their measures have not been studied to the same extent as for binary classification measures.

4. Invariance properties of measures

We focus on the ability of a measure to preserve its value under a change in the confusion matrix. A measure is *invariant* if its value does not change when a confusion matrix changes, i.e. invariance indicates that the measure does not detect the change in the confusion matrix. This inability can be beneficial or adverse, depending on the goals.

Let's consider a case when invariance to the change of *tn* is beneficial. Text classification extensively uses *Precision* and *Recall* (*Sensitivity*) which do not detect changes in *tn* when all other matrix entries remain the same. In document classification, a large number of unrelated documents constitute a negative class without having a single unifying characteristic. The criterion for the performance of a classifier is its performance on relevant documents, a well-defined unimodal positive class, independently of performance on the irrelevant documents. *Precision* and *Recall* do not depend on *tn*, but only on the correct labelling of positive examples (*tp*) and the incorrect labelling of examples (*fp* and *fn*). These measures provide the best perspective on a classifier's performance for document classification.

On contrast, the same invariance for the *tn* change can be an adversary. Consider the classification of human communication where negative classes are also important. In those problems, classes often have distinct features (male or female) for which both positive and negative classes are well-defined. The retrieval of a positive class, the discrimination between classes or the balance between retrieval from both classes are problem-dependent tasks. Thus, an appropriate evaluation measure should take into account the classification of negative examples and reflect the changes in *tn* when the other matrix elements stay the same.

We now examine eight invariance properties ($I_{1\leqslant k\leqslant 8}$) with respect to changes of elements in a confusion matrix. All the eight changes are results of elementary operations on matrices: addition, scalar multiplication, interchange of rows or columns. This set covers all relevant label distribution changes in a classification problem: interchange of positive and negative labels provided by data, interchange of those labels output by a classifier, change of a single segment (e.g., true positives), a uniform increase in the number of examples. Henceforth, $\overline{I_k}$ denotes the non-invariance of a transformation. We in detail discuss binary classification because other evaluation measures are derived from the binary confusion matrix and its performance measures. In several parts of the discussion, we refer to data quality. By this we understand how well examples represent the underlying notion (especially, ease of understanding and interpretability), how accurate is the data, including its labels, and the amount of noise (based on Wang & Strong (1996)). Thereinafter, $f([tp, fp, tn, fn])$ denotes a measure's value. Our claim is that the following invariance properties affect the applicability and trustworthiness of a measure.

4.1. (I_1) Exchange of positives and negatives

A measure $f([tp; fp; tn; fn])$ is invariant under exchange of positives and negatives if $f([tp; fp; tn; fn]) = f([tn; fp; tp; fn])$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} tn & fp \\ fn & tp \end{bmatrix}$$

This shows measure invariance with respect to the distribution of classification results because it does not distinguish *tp* from *tn* and *fn* from *fp* and may not recognize the 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 assessing the classification of consumer reviews than for document classification.

4.2. (I_2) Change of true negative counts

A measure $f([tp; fp; tn; fn])$ is invariant under the change of *tn* if all other matrix counts remain the same $f([tp; fp; tn; fn]) = f([tp; fp; tn'; fn])$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} tp & fn \\ fp & tn' \end{bmatrix}$$

This measure does not recognize the specifying ability of classifiers. Such evaluation may be more applicable to domains with a multi-modal negative class taken as "everything not positive". In the case of text classification, these invariant measures are suitable for the evaluation of document classification. If the measure is non-invariant, then it acknowledges the ability of classifiers to correctly identify negative examples. In this case, it may be reliable for comparison in domains with a well-defined, unimodal, negative class. Non-invariant measures are preferable for evaluating communications in which there are criteria for both positive and negative results.

4.3. (I_3) Change of true positive counts

A measure $f([tp; fp; tn; fn])$ is invariant under the change of *tp* if all other matrix counts remain the same $f([tp; fp; tn; fn]) = f([tp'; fp; tn; fn])$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} tp' & fn \\ fp & tn \end{bmatrix}$$

This measure does not recognize a classifier's sensitivity. Such evaluation can be complementary to other measures, but can hardly stay on its own. It may be reliable for comparison in domains with a well-defined, unimodal, negative class. As opposed to I_2, these invariant measures are not suitable for the evaluation of document classification. Non-invariant measures can be used as stand alone for evaluating classification with a strong positive class.

4.4. (I_4) Change of false negative counts

A measure $f([tp; fp; tn; fn])$ is invariant under the change of *fn* if all other matrix counts remain the same $f([tp; fp; tn; fn]) = f([tp; fp; tn; fn'])$

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} tp & fn' \\ fp & tn \end{bmatrix}$$

Invariance indicates measure stability under disagreement between the data and the negative labels assigned by a classifier. This is especially important for problems involving manual labelling. If a negative class has unreliable labels (Nigam & Hurst (2004) argue that humans can agree on only 74% of labels for negative opinion), an invariant measure may give misleading results. For non-invariant measures, its value's monotonicity is important. If the classifier evaluation improves when *fn* increases, the measure may favor a classifier prone to false negatives. The use of invariant and non-invariant measures should be decided based on problem and data characteristics.

4.5. (I_5) Change of false positive counts

A measure $f([tp; fp; tn; fn])$ is invariant under the change of *fp* if all other matrix counts remain the same $f([tp; fp; tn; fn]) = f([tp; fp'; tn; fn])$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} tp & fn \\ fp' & tn \end{bmatrix}$$

An invariant measure may provide reliable results when some of positive data labels are counter-intuitive. This can happen when the positive examples have outliers that cannot be explained by the mainstream data. We call such outliers *counterexamples*.

A non-invariant measure may not be suitable for data with many counterexamples. If the classifier evaluation improves when *fp* increases, the measure may favor a classifier prone to false positives. This is especially important for problems involving subjective labelling. Some data entries may not have consistent labels because of the difficulty of imposing

rigorous labelling rules. This can occur in the classification of records of long-term communications in which the data may contain a substantial number of counterexamples.

4.6. (I_6) Uniform change of positives and negatives

A measure $f([tp; fp; tn; fn])$ is invariant under a uniform change of positives and negatives if $f([tp; fp; tn; fn]) = f([k_1tp; k_1fp; k_1tn; k_1fn]); k_1 \neq 1$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} k_1tp & k_1fn \\ k_1fp & k_1tn \end{bmatrix}$$

An invariant measure is stable with respect to the uniform increase of data size, i.e., scalar multiplication of the confusion matrix. If we expect that for different data sizes the same proportion of examples will exhibit positive and negative characteristics, then the invariant measure may be a better choice for the evaluation of classifiers.

When a measure is non-invariant, then its applicability may depend on data sizes. The non-invariant measures may be more reliable if we do not know how representative the data sample is in terms of the proportion of positive/negative examples.

4.7. (I_7) Change of positive and negative columns

A measure $f([tp; fp; tn; fn])$ is invariant under columns' change if $f([tp; fp; tn; fn]) = f([k_1tp; k_1fp; k_2tn; k_2fn]); k_1 \neq k_2$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} k_1tp & k_2fn \\ k_1fp & k_2tn \end{bmatrix}$$

Suppose that different data sizes have the same proportion of positive and negative examples. This change in the confusion matrix is caused by changes in the proportion of positive and negative labels issued by an algorithm, i.e., the columns are multiplied by different scalars. This may happen when the quality of additional data substantially differs from the initial data sample (e.g., the information inflow can add more noise). However, an invariant measure does not show the performance change. Thus, it requires support of other measures to assess a classifier's performance on different classes.

A non-invariant measure reflects on the performance of a classifier on different classes. It is more appropriate if we can expect a change in the algorithm's performance across classes.

4.8. (I_8) Change of positive and negative rows

A measure $f([tp; fp; tn; fn])$ is invariant under rows' change if $f([tp; fp; tn; fn]) = f([k_1tp; k_2fp; k_2tn; k_1fn]); k_1 \neq k_2$.

$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix} \rightarrow \begin{bmatrix} k_1tp & k_1fn \\ k_2fp & k_2tn \end{bmatrix}$$

We again expect that different data sizes have the same proportion of positive and negative examples. Then the change in the confusion matrix corresponds to changes of an algorithm's performance within a positive (negative) class, i.e., the rows are multiplied by different scalars. For example, this may happen when a positive (negative) class is better represented in the new data. If we expect that different data sizes exhibit same quality of positive (negative) characteristics, then the invariant measure may be a better choice for the evaluation of classifiers.

When a measure is non-invariant, its applicability may depend on the quality of data classes. The non-invariant measures may be more reliable if we do not know how representative the data sample is in terms of the quality of positive and negative classes, which might be the case in web-posted consumer reviews.

For multi-class classification, we consider transformations of the confusion matrix for each class C_j. As expected, the measures retain their invariance properties regardless of *micro-* or *macro-averaging*.

For multi-topic classification, *Exact Match Ratio* and *Accuracy* have different invariant properties. Thus, referring to *Exact Match Ratio* as *Accuracy* may be misleading.

Measures used in hierarchical classification have a somewhat limited reliability because they evaluate the performance of a classifier either on subclasses or on superclasses, but not on both. Thus, invariance properties should be assessed with respect to the classification of subclasses – for $Precision_{\downarrow}$ and $Recall_{\downarrow}$, and superclasses – for $Precision_{\uparrow}$ and $Precision_{\uparrow}$.

Table 6 displays the invariance properties of the measures described in Tables 2–5. By assessing the invariant properties of commonly used measures, we show that $Precision$, $Precision_{\mu}$, $Precision_M$, $Precision_{\downarrow}$, $Precision_{\uparrow}$ exhibit same invariance characteristics. Thus, we group them as $Precision_G$ for general. Similarly, we group $Recall$, $Recall_{\mu}$, $Recall_M$, $Recall_{\downarrow}$, and $Recall_{\uparrow}$ as $Recall_G$, $Fscore$, $Fscore_{\mu}$, $Fscore_M$, $Fscore_{\downarrow}$, and $Fscore_{\uparrow}$ as $Fscore_G$, and, finally, *Accuracy*, *Average Accuracy*, and *Error Rate*, essentially 1-*Average Accuracy*, as $Accuracy_G$.

As a result, we further consider only those performance measures that vary in their invariance properties. Table 7 lists the measures and their properties. Our next step is to associate the invariant properties with particular settings.

Table 6
Invariance properties of performance measures (I_k) for different types of classification tasks. + denotes invariance and − non-invariance of the measure.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Binary classification							Table 2	
Accuracy	+	−	−	−	−	+	−	−
Precision	−	+	−	+	−	+	+	−
Recall (*Sensitivity*)	−	+	−	−	+	+	−	+
Fscore	−	+	−	−	−	+	−	−
Specificity	−	−	+	+	−	+	−	+
AUC	−	−	−	−	−	+	−	+
Multi-class classification							Table 3	
Average Accuracy	+	−	−	−	−	+	−	−
Error Rate	+	−	−	−	−	+	−	−
$Precision_{\mu}$	−	+	−	+	−	+	+	−
$Recall_{\mu}$	−	+	−	−	+	+	−	+
$Fscore_{\mu}$	−	+	−	−	−	+	−	−
$Precision_{M}$	−	+	−	+	−	+	+	−
$Recall_{M}$	−	+	−	−	+	+	−	+
$Fscore_{M}$	−	+	−	−	−	+	−	−
Multi-topic classification							Table 4	
Exact Match Ratio	−	−	−	+	+	−	−	−
Labelling Fscore	−	+	−	−	−	+	−	−
Retrieval Fscore	−	−	−	−	−	+	−	−
Hamming Loss	+	+	+	−	−	+	−	−
Hierarchical classification							Table 5	
$Precision_{\downarrow}$	−	+	−	+	−	+	+	−
$Recall_{\downarrow}$	−	+	−	−	+	+	−	+
$Fscore_{\downarrow}$	−	+	−	−	−	+	−	−
$Precision_{\uparrow}$	−	+	−	+	−	+	+	−
$Precision_{\uparrow}$	−	+	−	−	+	+	−	+
$Fscore_{\uparrow}$	−	+	−	−	−	+	−	−

Table 7
Performance measures that exhibit different invariance properties. + denotes invariance and − non-invariance of the measure.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
$Accuracy_G$	+	−	−	−	−	+	−	−
$Precision_G$	−	+	−	+	−	+	+	−
$Recall_G$(*Sensitivity*)	−	+	−	−	+	+	−	+
$Fscore_G$	−	+	−	−	−	+	−	−
Specificity	−	−	+	+	−	+	−	+
AUC	−	−	−	−	−	+	−	+
Exact Match Ratio	−	−	−	+	+	−	−	−
Labelling Fscore	−	+	−	−	−	+	−	−
Retrieval Fscore	−	−	−	−	−	+	−	−
Hamming Loss	+	+	+	−	−	+	−	−

5. Analysis of invariant properties

To identify similarities among the measures, we compare them according to their invariance and non-invariance properties shown in Table 7. First, we present measure outliers whose properties remarkably differ them from others. Two measures hold unique invariant properties: $Precision_G$ is the only measure invariant under vertical scaling (I_7) and *Exact Match Ratio* is the only measure non-invariant under uniform scaling ($\overline{I_6}$). Another exception is *Retrieval Fscore* which is sensitive to all the changes in the confusion matrix except for uniform scaling.

Next we generalize on the properties:

The invariance I_1 has been much discussed in the Machine Learning community, albeit from a negative point of view (Japkowicz, 2006). But we want to emphasize that this invariance makes $Accuracy_G$ and *Hamming Loss* robust measures for an algorithm's overall performance and insensitive to performance on a specific class. The corresponding non-invariance $\overline{I_1}$ means that the measures are sensitive to asymmetry of classification. This is a well-known characteristic for *Precision*, *Recall*, *Fscore* and *Specificity*, but not for *AUC*, which has been introduced only recently in text classification.

The invariance I_2 is a well-known property of *Precision*, *Recall*, and *Fscore* and less known for *Labelling Fscore* and *Hamming Loss*. Invariance under the change of *tn* has made them a tool of choice for the evaluation of document classification. The non-invariance $\overline{I_2}$ signifies that the use of non-invariant measures is more appropriate

on data with a unimodal negative class than with a multi-modal one. This implication is more important for *AUC* than for *Specificity*. The latter is usually used in combination with other measures, whereas the former might be applied separately.

The invariance I_3 so far eludes thorough studies. Measures are expected to be non-invariant under the change of *tp*. The non-invariant measures are used for evaluating classification with a strong positive class, such as for the evaluation of document classification. Only *Specificity* and *Hamming Loss* do not measure the *tp* change. *Specificity* was purposefully designed to avoid *tp*. The non-invariance of *Specificity* and *Hamming Loss* suggests they may be used in a combination with other measures. These two measures may be reliable for comparison in domains with a well-defined, unimodal, negative class.

The invariance I_4 under change in *fn* indicates that *Precision*, *Specificity*, and *Exact Match Ratio* may be more reliable when manual labelling follows rigorous rules for a negative class. In the absence of such rules, disagreement between the data labels and the negative labels assigned by a classifier can depend on subjective factors and fluctuate. Under such conditions, an invariant measure may give misleading results. All the $\overline{I_4}$ measures discussed above are monotone decreasing when *fn* increase hence, will not favor a classifier prone to false negatives.

The invariance I_5 under *fp* change indicates that *Recall* and *Exact Match Ratio* may provide reasonably conservative estimate when a positive class has counterexamples, i.e., outliers not explained by the mainstream positive examples. The other eight measures are non-invariant. However, they are monotone decreasing when *fp* increase, hence, they will not favor a classifier prone to false positives.

The invariance I_6 under uniform scaling holds for all the measures except *Exact Match Ratio*. The nine invariant measures adapt to different sizes of data. The non-invariance of *Exact Match Ratio* indicates that its results may not be comparable when obtained on data of widely different sizes.

The invariance I_7 under the scalar column change holds only for *Precision*. This supports a common practice of combining *Precision* with other measures when assessing classifier performance. The combination assures that the evaluation is less dependent on the data quality. All the other measures are non-invariant under the scalar column change. Thus, they are more reliable if an algorithm's performance is expected to change across classes with new data.

The invariance I_8 under the scalar row change indicates that *Recall*, *Specificity*, and *AUC* may be a better choice for the evaluation of classifiers if different data sizes exhibit same quality of positive (negative) characteristics. Examples are simulated or generated data under the same distribution. The other measures are non-invariant. They may be more reliable if the representative power of positive and negative classes is uncertain.

Invariance with respect to the matrix transformations is especially important because it connects evaluation measures to particular learning settings. We now summarize the applicability of these measures to two subfields of text classification: document classification and classification of human communications. One might be tempted to apply *Fscore* measures on any text classification evaluation. However, various classification problems exhibit different characteristics which may require different evaluation measures. Based on our analysis, we propose the following.

Since document classification data is often highly imbalanced, relevant documents constitute a small well-defined positive class, but the rest is a heterogeneous negative class built from non-relevant documents as "everything non-positive". Presence of a negative class that complements the positive class favors the use of the *Fscore* measures. In many such problems, examples of the positive class remain the same and the class keeps its modality, whereas examples of the negative class change. Since the *Fscore* measures' invariance under the change of correctly classified negative examples (I_2) prevents drastic changes, they will be less sensitive to changes in the negative class.

Classification of human communications is most often represented by sentiment classification applied to collections of free form texts containing product evaluations. The number and ratio of positive and negative examples depends on the popularity of a particular product. If reviewers have strong likes and dislikes, then both classes have well-defined characteristics. In this case, *Area Under the Curve* (*AUC*) may provide a more reliable classifier evaluation than *Precision* and *Recall*. Since *AUC* is non-invariant under the change of correctly classified negative examples ($\overline{I_2}$), it will detect possible changes in the negative class better than *Fscore* measures.

For other types of classification of communications in social activities, other measure combinations might also be suitable. Political debates and electronic negotiations are examples of such communications. Their data can exhibit a unimodal negative class and a large number of counterexamples. In political debates, counterexamples are records that praise the discussed matter, but vote against it at the end, either because of a hidden motive or randomness of behavior (Sokolova & Lapalme, 2007). In such cases, which are difficult even for human classification, *Accuracy*, with its invariance under the exchange of positives and negatives classification (I_1), and *Precision*, with its invariance under the change of false negative examples (I_5), may be used for a reliable evaluation of classifiers.

6. Conclusion and future work

In this study, we have analyzed twenty four performance measures used in the complete spectrum of Machine Learning classification tasks: binary, multi-class, multi-labelled, and hierarchical. Effects of changes in the confusion matrix on several

well-known measures have been studied. In all the cases, we have shown that the evaluation of classification results can depend on the invariance properties of the measures. A few cases required that we additionally considered monotonicity of the measure. These properties have allowed us to make fine distinctions in the relations between the measures. One way to insure a reliable evaluation is to employ a measure corresponding to the expected quality of the data, e.g., representativeness of class distribution, reliability of class labels, uni- and multi-modality of classes. To match measures with the data characteristics, we constructed the measure invariance taxonomy with respect to all relevant label distribution changes in a classification problem.

We supplemented the formal discussion by analyzing the applicability of performance measures on different subfields of text classification. We have shown that the classification of human communications differs from document classification, and thus that these two types of text classification may require different performance measures.

Our study has dealt with measures used in text classification but it could be extended to other language applications of Machine Learning. The next step would be to study measures used in Machine Translation. This will considerably expand the measure list. Applicability of the measures to traditional Natural Language Processing tasks, e.g., word sense disambiguation, parsing, is another topic of considerable interest. It would also be useful to analyze in more details a measure's monotonicity, especially its behavior with respect to extreme classification results, such as when the labels provided by the data and a classifier are independent. Person authentication problems, in which the appropriate measures are a *false acceptance rate* and a *false rejection rate* (Bengio et al., 2005), is another example of possible applications.

Acknowledgments

This work has been funded by the Natural Sciences and Engineering Research Council of Canada and the Ontario Centres of Excellence. We thank Elliott Macklovitch for fruitful suggestions on an early draft. We thank anonymous reviewers for helpful comments.

References

Asuncion, A., & Newman, D. (2007). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science. <http://www.ics.uci.edu/mlearn/MLRepository.html>.

Bengio, S., Mariéthoz, J., & Keller, M. (2005). The expected performance curve. In *Proceedings of the ICML'05 workshop on ROC analysis in machine learning* (pp. 43–50).

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 440–447). Association for Computational Linguistics.

Blockeel, H., Bruynooghe, M., Dzeroski, S., Ramon, J., & Struyf, J. (2002). Hierarchical multi-classification. In *KDD-2002: Workshop on multi-relational data mining* (pp. 21–35).

Bobicev, V., & Sokolova, M. (2008). An effective and robust method for short text classification. In *Proceedings of the association for the advancement of artificial intelligence (AAAI-2008)* (pp. 1444–1445). AAAI Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Costa, E., Lorena, A., Carvalho, A., & Freitas, A. (2007). A review of performance evaluation measures for hierarchical classifiers. In *Proceedings of the AAAI 2007 workshop "Evaluation methods for machine learning"* (pp. 1–6).

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*, 1895–1923.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. John Wiley & Sons.

Eisner, R., Poulin, B., Szafron, D., Lu, P., & Greiner, R. (2005). Improving protein function prediction using the hierarchical structure of the gene ontology. In *Proceedings of IEEE symposium on computational intelligence in bioinformatics and computational biology* (pp. 1–10).

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Proceedings of the 6th international symposium on intelligent data analysis (IDA 2005)* (pp. 121–132).

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Proceedings of 27th European conference on IR research (ECIR 2005)* (pp. 345–359).

Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1997). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR-97)* (pp. 192–201).

Huang, J., & Ling, C. (2007). Constructing new and better evaluation measures for machine learning. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI'2007)* (pp. 859–864).

Isselbacher, K., & Braunwald, E. (1994). *Harrison's principles of internal medicine*. McGraw-Hill.

Japkowicz, N. (2006). Why question machine learning evaluation methods? In *Proceedings of the AAAI'06 workshop on evaluation methods for machine learning* (pp. 6–11).

Kazawa, H., Izumitani, T., Taira, H., & Maeda, E. (2005). Maximal margin labeling for multi-topic text categorization. In *Advances in neural information processing systems (NIPS'04)*, (Vol. 17, pp. 649–656).

Kiritchenko, S., Matwin, S., Nock, R., & Famili, A. F. (2006). Learning and evaluation in the presence of class hierarchies: Application to text categorization. In *Proceedings of the 19th Canadian conference on AI (AI'2006)* (pp. 395–406).

Lachiche, N., & Flach, P. A. (2003). Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In *Proceedings of ICML'2003* (pp. 416–423).

Langley, P. (1996). *Elements of machine learning*. San Francisco, Calif: Morgan Kaufmann.

Li, T., Zhang, C., & Zhu, S. (2006). Empirical studies on multi-label classification. In *Proceedings of the 18th IEEE international conference on tools with artificial intelligence* (pp. 86–92).

Marchand, M., & Shawe-Taylor, J. (2002). The set covering machine. *Journal of Machine Learning Research, 3*, 723–746.

Mewes, H.-W., Albermann, K., Heumann, K., Lieb, S., & Pfeiffer, F. (1997). MIPS: A database for protein sequences, homology data and yeast genome information. *Nucleic Acids Research, 25*(1), 28–30.

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Nigam, K., & Hurst, M. (2004). Towards a robust metric of opinion. In *Proceedings of the AAAI spring symposium on exploring attitude and affect in text* (pp. 98–105). AAAI Press.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of empirical methods of natural language processing (EMNLP'02)* (pp. 79–86).

Rijsbergen, C. (1979). *Information retrieval* (2nd ed.). London: Butterworths.
Salton, G., & McGill, M. (1993). *Introduction to modern information retrieval*. New York: McGraw-Hill.
Salzberg, S. L. (1999). On comparing classifiers: A critique of current research and methods. *Data Mining and Knowledge Discovery, 1*, 1–12.
Sasaki, Y., Rea, B., & Ananiadou, S. (2007). Multi-topic aspects in clinical text classification. In *Proceedings of the 2007 IEEE international conference on bioinformatics and biomedicine* (pp. 62–70). IEEE Computer Society.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.
Shawe-Taylor, J., & Christianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.
Snyder, B., & Barzilay, R. (2007). Database-text alignment via structured multilabel classification. In *Proceedings of the international joint conference on artificial intelligence (IJCAI-2007)* (pp. 1713–1718).
Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *Proceedings of the ACS Australian joint conference on artificial intelligence* (pp. 1015–1021).
Sokolova, M., & Lapalme, G. (2007). Performance measures in classification of human communication. In *Proceedings of the 20th Canadian conference on artificial intelligence (AI'2007)* (pp. 159–170).
Sun, A., Lim, E.-P., & Ng, W.-K. (2003). Performance measurement framework for hierarchical text classification. *Journal of the Americal Society for Information Science and Technology, 54*(11), 1014–1028.
Tan, P., Kumar, V., & Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems, 29*(4), 293–313.
Tikk, D., & Biró, G. (2003). Experiments with multi-label text classifier on the Reuters collection. In *Proceedings of the international conference on computational cybernetics (ICCC 03)* (pp. 33–38).
Thomas, M., Pang, B., & Lee, L. (2006). 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* (pp. 327–335).
Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–34.
Wilson, T., Wiebe, J., & Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence, 22*(2), 73–99.
Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*, 69–90.
Youden, W. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.
Zhu, S., Ji, X., Xu, W., & Gong, Y. (2005). Multi-labelled classification using maximum entropy method. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 274–281).