Information Retrieval Journal (2019) 22:581–619
https://doi.org/10.1007/s10791-019-09355-y

A comparison of filtering evaluation metrics based on formal constraints

Enrique Amigó[1] · Julio Gonzalo[1] · Felisa Verdejo[1] · Damiano Spina[2]

Received: 14 April 2016 / Accepted: 20 March 2019 / Published online: 1 April 2019

Abstract
Although document filtering is simple to define, there is a wide range of different evaluation measures that have been proposed in the literature, all of which have been subject to criticism. Our goal is to compare metrics from a formal point of view, in order to understand whether each metric is appropriate, why and when, in order to achieve a better understanding of the similarities and differences between metrics. Our formal study leads to a typology of measures for document filtering which is based on (1) a formal constraint that must be satisfied by any suitable evaluation measure, and (2) a set of three (mutually exclusive) formal properties which help to understand the fundamental differences between measures and determining which ones are more appropriate depending on the application scenario. As far as we know, this is the first in-depth study on how filtering metrics can be categorized according to their appropriateness for different scenarios. Two main findings derive from our study. First, not every measure satisfies the basic constraint; but problematic measures can be adapted using smoothing techniques that and makes them compliant with the basic constraint while preserving their original properties. Our second finding is that all metrics (except one) can be grouped in three families, each satisfying one out of three formal properties which are mutually exclusive. In cases where the application scenario is clearly defined, this classification of metrics should help choosing an adequate evaluation measure. The exception is the Reliability/Sensitivity metric pair, which does not fit into any of the three families, but has two valuable empirical properties: it is strict (i.e. a good result according to reliability/sensitivity ensures a good result according to all other metrics) and has more robustness that all other measures considered in our study.

Keywords Document filtering · Evaluation metrics · Evaluation methodologies

✉ Julio Gonzalo
julio@lsi.uned.es
http://nlp.uned.es

[1] NLP & IR Research Group, UNED, calle Juan del Rosal, 16, 28040 Madrid, Spain

[2] Present Address: Computer Science and Information Technologies, RMIT, Melbourne, Australia

1 Introduction

Document Filtering is a generic problem involved in a wide set of tasks such as spam detection (Cormack and Lynam 2005), Information Retrieval over user profiles (Hoashi et al. 2000), post retrieval selection for on-line reputation management (Amigó et al. 2012), etc. In essence, document filtering is a binary classification task with priority (one is the class of interest, the other is meant to be discarded). It consists of discerning relevant from irrelevant documents from an input document stream. In spam filtering, for instance, the system must keep relevant e-mails and discard unwanted mails.

Although document filtering is simple to define, there is a wide range of different evaluation measures that have been proposed in the literature, all of which have been subject to criticism. Just as an illustration, TREC (the Text Retrieval Evaluation Conference) has organized at least three filtering tasks, all of them using different evaluation metrics: the Filtering track used utility (Hull 1998), the Spam track chose Lam% (Cormack and Lynam 2005), and the legal track employed a variation of F (Hedin et al. 2009). In fact, the choice of an appropriate, flawless evaluation measure seems to be still controversial in many filtering scenarios.

Our goal is to provide a systematic, formal comparison of existing evaluation metrics that helps us determine when they are appropriate and why. Previous comparisons between metrics have focused on issues such as stability of measures across datasets, ability to discriminate systems with statistical significance, or sensitivity to small changes in the input. We take a different approach: we focus on establishing a set of formal constraints (Amigó et al. 2009; Fang et al. 2004) that define properties of filtering metrics. Some formal constraints must be satisfied by any suitable metric, and other constraints help understanding and comparing metrics according to their properties. A key novelty of our analysis is that it is grounded on a probabilistic interpretation of measures that facilitates formal reasoning.

First, we assume one basic constraint that should be satisfied by any evaluation metric, for any filtering scenario. Our analysis shows that many criticisms to existing metrics can be explained in terms of the (lack of) satisfaction of this constraint. In particular, some of the most popular measures (such as the F measure of Precision and Recall for the positive class and Lam%) fail to satisfy them. However, we also show that redefining measures in probabilistic terms and applying smoothing techniques lead to alternative definitions of Lam% and F measure that—when the smoothing technique is chosen with care—have a similar behavior but comply with our basic constraints.

Even measures that satisfy the basic constraint, however, can say different things about the comparative performance of systems. Our starting point to understand their differences is a key empirical observation: in a filtering dataset (Amigó et al. 2010), measures differ substantially, and *the sharpest differences reside in how they evaluate non-informative outputs* (systems whose output is independent from the input, such as a system that always returns the same label for every item). This serves as inspiration to define three mutually-exclusive properties that depend on how measures handle non-informative outputs. The three properties, then, define three families of metrics, and provide a clear-cut criterion to choose the most adequate measure (or family of measures) for a given application scenario.

In addition to our formal analysis—and its practical outcomes—we also report empirical results on (1) the practical effects of our proposed smoothed measures, (2) the relative strictness of metrics, and (3) metric robustness with respect to variations in the set of test

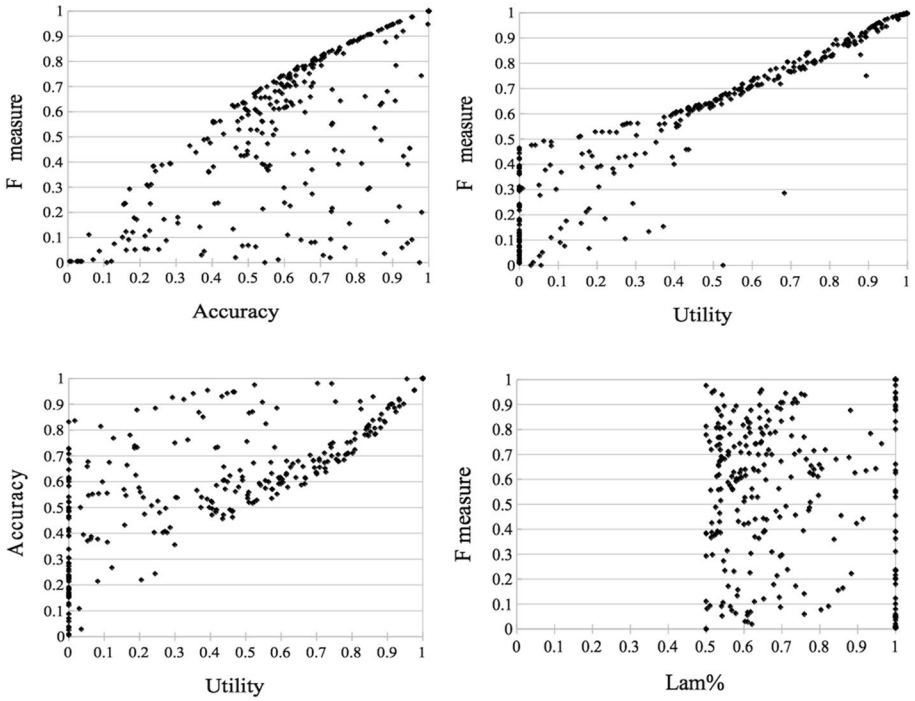

Fig. 1 Correspondence between popular measures in the WEPS-3 evaluation campaign. Each dot corresponds to a system output for one test case

cases. This empirical analysis complements our formal study and provides deeper insights into the differences of behavior between metrics.

This paper is structured as follows: in Sect. 2, we begin with a preliminary experiment on how measures disagree and why. Then in Sect. 3, we present our formal analysis of measures. In Sect. 4 we introduce the smoothed versions of measures that do not comply with our basic constraints. Finally, Sect. 5 presents our empirical analysis, Sect. 6 discusses related work and Sect. 7 summarizes our conclusions.

2 A preliminary experiment on how measures disagree and why

In this section we perform a preliminary experiment on how measures disagree that motivates our study, showing how different the veredict of measures can be on the same dataset. And, more importantly, it partly suggest how to differentiate measures and classify them in families, because it shows that measure disagreement concentrates mainly on non-informative outputs (those that do not depend on the input).

2.1 Measure disagreement

Our initial experiment consists of comparing the most popular measures used in the TREC filtering evaluation campaigns over the WePS-3 dataset (see Sect. 4 for a

description of each measure, and Sect. 6.1 for a detailed description of the WePS-3 task and the dataset). Figure 1 shows the correspondence between measures for systems participating in the WePS-3 evaluation campaign. Each graph compares two standard measures, and each dot corresponds with a system output for one test case in the collection. In order to illustrate the behavior of measures under different system outputs for the same topic, the grey squares represent the outputs for one (randomly selected) topic.

The graphs clearly illustrate with four examples that, in general, the correlation between measures is lower than expected:

- *F measure versus Accuracy*. The F measure (harmonic mean of Precision and Recall for the positive class) seems to be a lower bound on the value of Accuracy, which can take values arbitrarily larger than the F measure, but not smaller.
- *F measure versus Utility*. This is the only metric pair that has a high correlation in our experiments, and only for values above 0.5. Below 0.5, both metrics can say radically different things about the quality of a system.
- *Utility versus accuracy*. Utility also seems to be a lower bound on the value of accuracy, but beyond that there is little correlation between both metrics.
- *F measure versus Lam%*. The patterns in the graph F-measure versus Lam% are very particular. As we explain in the next section, the reason is that Lam% is a measure based on information gain, and it considers the probabilistic dependence between the system output and the gold standard signals. As a consequence, Lam% assigns a score of 0.5 to any random output. This explains the vertical line at Lam% 0.5. Also, as we explain in Sect. 4, it is possible to achieve a maximal Lam% score without predicting the correct class in most cases, which explains the vertical line at Lam%=1. Overall, the plot shows little correlation between both measures.

This lack of correlation implies that system ranking can be severely affected by the metric choice. Also, it means that a system development cycle—where the system is repeatedly tested and improved with respect to a certain evaluation measure—can be easily biased by the measure selected. Therefore, it is crucial to understand how and why measures differ, in order to prevent the use of inadequate measures for a given task and application scenario.

2.2 The role of non-informative outputs

We now select, from the WePS-3 dataset, all the non-informative outputs. We use *non-informative output* to refer to those cases where the automatic classification is statistically independent of the real document classification. In other words, when there is no correlation between the system output and the gold standard. For instance, returning the same label for all documents (the system accepts everything or discards everything) would be a non-informative output. Also, a system that always picks up a random selection of the input documents as relevant is also non-informative. The set of WePS-3 systems includes the baseline systems provided by the organization.

Figure 2 illustrates how the correspondence between measures looks like when we only display results for non-informative outputs. Comparing both Figs. 1 and 2 we see that the scores of non-informative outputs tend to draw the limits of the dotted areas in Fig. 1. In general, this means that the non-informative outputs include most of the extreme cases of measure disagreement. Our conclusion is that a key factor of disagreement between metrics

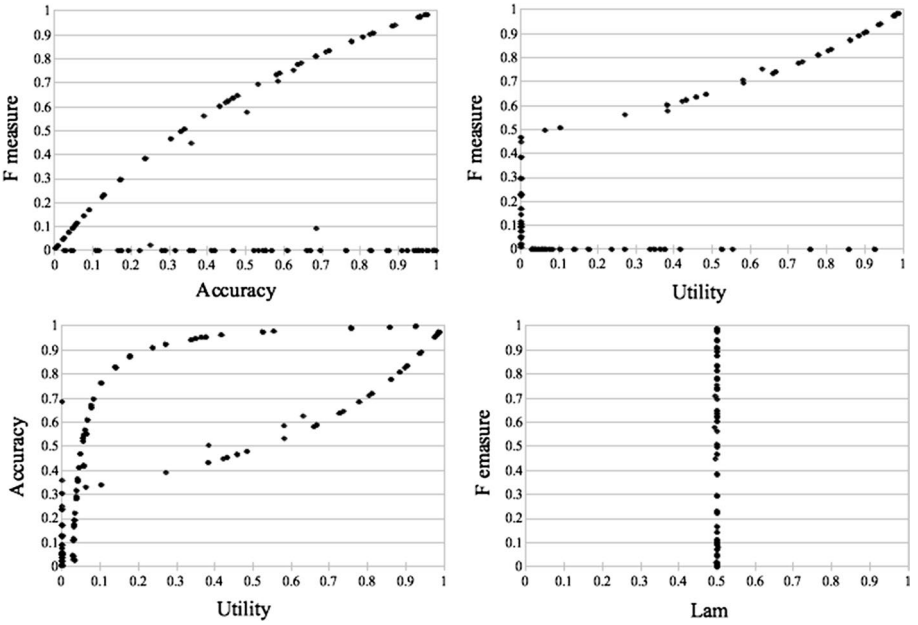

Fig. 2 Correspondence between measure scores for non-informative systems in the WEPS-3 evaluation campaign. Each dot corresponds to a system output for one test case

is how non-informative outputs are scored by measures. In other words, the treatment of non-informative outputs is a strong defining characteristic of a filtering evaluation measure.

Note that purely non-informative systems are artificial (i.e. simply used as baselines for comparison purposes). One could argue that it is not crucial how measures evaluate artificial systems, but only how they evaluate real systems. And this would be a reasonable objection. However, many real systems may have a near non-informative behavior. In fact, in the WePS-3 dataset there are many system outputs which have low informativeness.

The empirical observation that metrics differ most in how they evaluate non-informative outputs has served as inspiration for our formal analysis, and has led to the definition of three mutually exclusive properties of metrics which produce a classification of measures in three families. Given an application scenario, choosing which property is adequate leads to one of the metric families, thus guiding the metric selection process.

3 Theoretical framework

In this section, we formulate the basic constraints (which must be satisfied by any metric) and properties (which further characterize metrics) that help assessing filtering evaluation measures. In order to enable our formal analysis, we first introduce a probabilistic notation to describe measures and measure properties.

Note that our study focuses only on measures that assess the overall quality of systems, rather than measures that cover partial quality aspects (such as False Positive Rate, False Negative Rate, Recall, True Negative Rate, Precision, Negative Predictive Value, Prediction-conditioned Fallout, Prediction-conditioned Miss, Rate of Positive Predictions or Rate of Negative Predictions). For instance, Precision and Recall are partial and complementary quality aspects, and they can be used to assess the overall quality of a system if they are combined. In this case, the most popular way of combining them is via a weighted harmonic mean (the F measure).

We also restrict our study to measures that work on binary decisions (relevant vs. irrelevant), rather than on a ranked list of documents. Typically, a filtering system—as any binary classifier—outputs a probability of relevance for every item,[1] and the final classification implies choosing a threshold for this probability. Then, items above/below the threshold are classified as relevant/irrelevant. One way of evaluating document filtering is by inspecting the rank of documents (ordered by decreasing probability of relevance) and then measuring precision and recall at certain points in the rank. The advantage of this type of evaluation is that the classification algorithm can be evaluated independently from how the threshold is finally set. Some examples of this type of evaluation are ROC (Receiver Operating Characteristic) (Provost and Fawcett 1997) and AUC (Area Under the Curve) (Ling et al. 2003), which compare the classification performance across decreasing classification threshold values. For document filtering tasks, some researchers also evaluate Precision at a certain number of retrieved documents (Robertson and Hull 2001; Callan 1996) or average across recall levels (Persin 1994). Other related measures are Mean Cross-entropy (Good 1952), Root-mean-squared error, Calibration Error (Fawcett and Niculescu-Mizil 2007), SAR and Expected Cost (all of them available and described, for instance, in the R package called ROCR.[2]) These metrics, however, do not consider the ability of systems to predict the ratio of relevant documents in the input document stream, which is, in practice, a crucial aspect of system quality (Schapire et al. 1998); a good ranking may turn into a poor classification if the cutoff point is not chosen adequately. In order to consider this aspect, metrics have to work on the binary output of the filtering system, rather than on the intermediate internal rank. Therefore, we restrict ourselves to these kind of metrics.

3.1 The filtering task: a probabilistic notation

We understand the filtering task as follows. A filtering input consists of a document[3] set $\mathcal{T}$ which contains relevant (subset $\mathcal{G}$) and irrelevant documents (subset $\neg\mathcal{G}$).[4] $\mathcal{G}$ is the subset of documents manually assessed as relevant, and its complementary $\neg\mathcal{G}$ is the subset of documents manually assessed as irrelevant.

The system output is represented with a subset $\mathcal{S}$ containing all documents labeled as positive by the system. Its complementary set $\neg\mathcal{S}$represents documents labeled as negative by the system. Given an input test set $\mathcal{T} = \mathcal{G} \cup \neg\mathcal{G}$, a metric returns a certain quality score $Q(\mathcal{S})$ for the system with output set $\mathcal{S}$.

[1] Or, more precisely, a quantity which can be mapped into a probability of relevance using some growing monotonic function.

[2] http://cran.r-project.org/web/packages/ROCR/ROCR.pdf.

[3] For the sake of readability, we will speak of *documents*. However, our conclusions can be applied to any kind of items.

[4] Letter G is chosen for *Gold standard*.

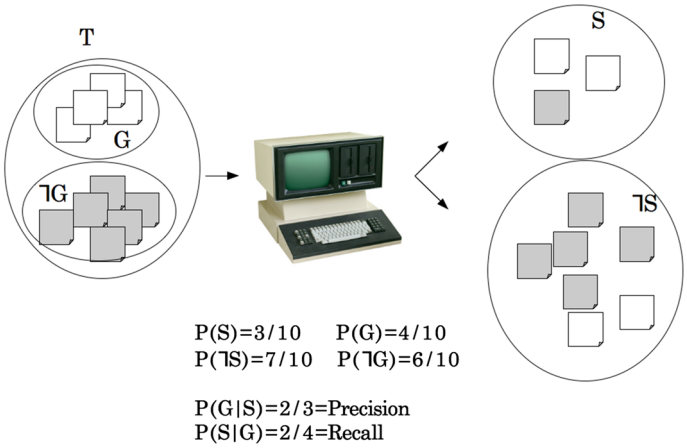

Fig. 3 Interpretation and notation for the filtering task

Table 1 Relationship between the traditional contingency matrix and our probabilistic notation

$TP = \|\mathcal{S} \cap \mathcal{G}\| \sim P(\mathcal{G}\|\mathcal{S})P(\mathcal{S})$	$FP = \|\mathcal{S} \cap \neg\mathcal{G}\| \sim P(\neg\mathcal{G}\|\mathcal{S})P(\mathcal{S})$
$FN = \|\neg\mathcal{S} \cap \mathcal{G}\| \sim P(\mathcal{G}\|\neg\mathcal{S})P(\neg\mathcal{S})$	$TN = \|\neg\mathcal{S} \cap \neg\mathcal{G}\| \sim P(\neg\mathcal{G}\|\neg\mathcal{S})P(\neg\mathcal{S})$

We will use the simplified notation $P(\mathcal{G})$ to denote the probability $P(e \in \mathcal{G})$ measured over the space of samples $\mathcal{T}$. We use the same notation for any subset of $\mathcal{T}$. Using this notation we can express quality metrics; for instance, Precision (fraction of relevant documents in the subset labeled as positive by the system) is $P(\mathcal{G}|\mathcal{S})$, and can be computed as $\frac{|\mathcal{S}\cap\mathcal{G}|}{|\mathcal{S}|}$. Figure 3 illustrates an example where $\mathcal{G}$ contains four documents and $\mathcal{S}$ contains three documents with the following values for Precision and Recall.

The traditional representation is the *contingency matrix*, which uses four subsets: true positives (TP) are items labeled as positive and relevant; false positives (FN) are items labeled as negative but relevant; true negatives (TN) are items labeled as negative and irrelevant; and false negatives (FN) are labeled as negative and irrelevant. Table 1 illustrates the correspondence between our notation and the contingency matrix.

Our notation is not standard, and it does not seem simpler at first sight; but it is crucial for us to provide formal proofs in the remainder of the paper, and to propose smoothing mechanisms for the metrics that require it.

Note that neither the contingency matrix nor our probabilistic notation allow to consider a notion of document redundancy: for instance, the penalty for discarding a redundant relevant document is the same as the penalty for discarding a unique relevant document. In this work (as in many other approaches to the subject) we assume that filtering is a process preliminary to redundancy removal and is, therefore, evaluated independently.

The evaluation process requires to estimate some probabilities over observable data. Some of these probabilities are:

- the ratio of relevant documents in the input stream: $(P(\mathcal{G}) \sim \frac{|\mathcal{G}|}{|\mathcal{T}|})$
- the system output size $(P(\mathcal{S}) \sim \frac{|\mathcal{S}|}{|\mathcal{T}|})$

Table 2 Summary of our notation

$\mathcal{T}$	Set of documents
$\mathcal{G}$	Documents manually assessed as relevant (positive)
$\neg\mathcal{G}$	Documents manually assessed as irrelevant (negative)
$\mathcal{S}$	Documents labeled relevant by the system
$\neg\mathcal{S}$	Documents labeled as irrelevant by the system
$P(\mathcal{G})$	Probability for a document $e \in \mathcal{T}$ of belonging to the relevant set $\mathcal{G}$
$P(\mathcal{G}\|\mathcal{S})$	Precision ($\frac{\|\mathcal{S}\cap\mathcal{G}\|}{\|\mathcal{S}\|}$)
$P(\mathcal{S}\|\mathcal{G})$	Recall ($\frac{\|\mathcal{S}\cap\mathcal{G}\|}{\|\mathcal{G}\|}$)
$\mathcal{S}_{\neg i}$	Non-informative systems: $P(\mathcal{S}_{\neg i} \cap \mathcal{G}) = P(\mathcal{S}_{\neg i})P(\mathcal{G})$
$\mathcal{S}_T = T$	Placebo baseline system (everything is labeled positive)
$\mathcal{S}_\emptyset = \emptyset$	Zero baseline system (everything is labeled negative)

- several conditional probabilities such as Precision ($P(\mathcal{G}|\mathcal{S}) \sim \frac{|\mathcal{S}\cap\mathcal{G}|}{|\mathcal{S}|}$) or Recall ($P(\mathcal{S}|\mathcal{G}) \sim \frac{|\mathcal{S}\cap\mathcal{G}|}{|\mathcal{G}|}$).

The probabilistic representation allows to define non-informative outputs $\mathcal{S}_{\neg i}$ as those whose output set $\mathcal{S}$ is chosen independently from their relevance $\mathcal{G}$:

$$P(\mathcal{S}_{\neg i} \cap \mathcal{G}) = P(\mathcal{S}_{\neg i})P(\mathcal{G})$$

The non-informativeness property can be also expressed as:

$$P(\mathcal{S}_{\neg i}|\mathcal{G}) = P(\mathcal{S}_{\neg i}) \quad \vee \quad P(\mathcal{G}|\mathcal{S}_{\neg i}) = P(\mathcal{G})$$

For the discussions to follow, it is interesting to think of two particular non-informative outputs. The first one classifies all documents as relevant, returning the original document set without modifications. We will refer to this system as the *Placebo* baseline ($\mathcal{S}_T = T$).[5] The second one is the *Zero* system, which returns an empty output: $\mathcal{S}_\emptyset = \emptyset$.

Table 2 summarizes some useful expressions in our notation.

3.2 The strict monotonicity axiom

Sebastiani (2015) proposed a basic axiom as a formal constraint that must be satisfied by any classification evaluation measure. It states that any relabeling of an item into its correct category must produce an increase in any appropriate measure score. Using our notation:

$$\mathcal{S} = \mathcal{S}' \cup \{e\} \wedge e \in \mathcal{G} \Rightarrow Q(\mathcal{S}) > Q(\mathcal{S}')$$
$$\neg\mathcal{S} = \neg\mathcal{S}' \cup \{e\} \wedge e \in \neg\mathcal{G} \Rightarrow Q(\mathcal{S}) > Q(\mathcal{S}')$$

This axiom also implies that the maximum score is achieved only when every item is correctly classified.

Assuming that $\mathcal{S}$ and $\mathcal{G}$ are two sets, this constraint is closely related to Tversky's *Monotonicity Axiom* for similarity between sets (Tversky 1977). As Sebastiani proved, although

[5] What our definition of the placebo baseline implies is that document filtering is an asymmetric process in terms of the positive/negative labels. This is implicit in most literature on the subject: for instance, precision and recall are assumed by default to be computed on the relevant class.

this intuitive axiom seems obvious, it is not satisfied by some popular measures, as we discuss in Sect. 4.

3.3 Measure properties and use cases

The previous axiom is a constraint that must be satisfied by any suitable metric. We now focus on *properties*, which are structurally similar but are intented to characterize how some metrics work (rather than prescribing how they should work), and help distinguishing measures. Note that the properties we are about to introduce are mutually exclusive: any metric can satisfy at most one of them.

In the previous section we have showed how the evaluation of non-informative outputs is what makes measures different. Therefore, we focus on establishing properties that describe different ways of handling non-informative outputs. We exemplify how these properties are useful with a single task in the context of Online Reputation Monitoring: *name ambiguity resolution*. Given all the online posts that contain the name of an entity (e.g. a company) to be monitored, we want to select the texts that do refer to the company, and discard the texts that refer to something else. For instance, if the entity is the telecom company *Orange*, we want to filter out appearances of orange that refer to the colour, the fruit, etc.

Scenario 1: Absolute gain for single documents classifications. Sometimes, the quality of a non-informative output $S_{\neg i}$ depends on the absolute gain/loss associated with correctly/incorrectly classified elements.
In the name ambiguity task, the system has to assign a positive label to the items that refer to, for instance, the *Orange* company. Let us suppose that the output of the system is used by a competitor that wants to advertise the advantages of its services to people that are talking about Orange. For every False Positive, there is a quantifiable loss (in time or money), as well as for every True Positive there is a quantifiable average gain, etc. Depending on the relative cost/profit of every incorrect/correct label, a system that assigns positive labels to all items (what we call a *placebo* output S_T) can be better or worse than assuming that nothing is about the company (zero output ($S_\emptyset$).
According to this, we define the *Absolute Weighting* property as the ability of measures to assign an absolute weight to relevant (versus non relevant) documents in the output regardless of the output size. A measure satisfies the absolute weighting property if it has a parameter that determines the relative profit of selecting a relevant document with respect to the cost of selecting an irrelevant document. Depending on the value of the parameter, adding together a relevant and an irrelevant document to the positive class may have an overall positive effect on the measure (the profit is higher than the cost), or a negative overall effect (the cost of selecting the irrelevant document is higher than the benefit of adding the relevant one). We formalize this property by saying that there exists a threshold value for the parameter which determines if adding one relevant and one irrelevant documents to the positive output set $\mathcal{S}$ improves the system output or not. Formally, being[6]:

$$\mathcal{S}' \equiv \mathcal{S} \cup \{e_{\mathcal{G}} \in \mathcal{G}\} \cup \{e_{\neg\mathcal{G}} \in \neg\mathcal{G}\}$$

then there exists a parameter c and a threshold θ such that:

[6] The formula assumes that both $e_{\mathcal{G}}$ and $e_{\neg\mathcal{G}}$ did not already belong to $\mathcal{S}$.

$$c > \theta \Leftrightarrow Q_c(S) > Q_c(S')$$

When increasing the size of a non-informative output ($|S_{\neg i}|$), both the amount of relevant and irrelevant documents labeled as positive grow. Given that the system is non-informative, the relative growth of relevant versus irrelevant documents in the output is fixed, and only depends on the ratio of relevant documents in the input stream. Therefore, we can express this property in terms of non-informative output scores: If $S_{\neg i}$ and $S'_{\neg i}$ are two different non-informative outputs, the property requests that there exists a certain parameter that determines if one non-informative output is better (when the parameter is above a certain threshold) or worse (when it is below the threshold) than other:

$$c > \theta \Leftrightarrow Q_c(S_{\neg i}) > Q_c(S'_{\neg i})$$

Scenario 2: Any non-informative output is equally useless. Let us now consider a use case in which the system output is used to estimate how frequently online texts that contain the word *Orange* refer to the telecom company and, subsequently, to estimate the online presence of the company.

In this case, any non-informative output is equally useless, because it will predict a ratio of relevant documents which is independent from the actual data. This leads to a *Non-Informativeness Fixed Quality* property that we formulate as follows: for any non-informative output $S_{\neg i}$ its quality is constant $Q(S_{\neg i})$.[7] That is:

$$Q(S_{\neg i}) = k$$

We will refer to measures satisfying this property as *Informativeness-based measures*.

Scenario 3: Doing nothing is better than doing random. Finally, let us consider a third scenario in which items labeled as positive are examined by experts in Public Relations in order to identify and handle potential reputation alerts. In this scenario, recall is crucial, because the risk of failing to detect a reputation alert is much worse than having to examine an irrelevant post.

In these conditions, discarding all documents is catastrophical: the reputation experts simply cannot do their job. Returning all documents (placebo baseline), on the other hand, implies a lot of extra work, but it is not nearly as harmful than the zero baseline. In general, the more the system removes texts randomly, the more harmful the classifier is.

For these cases we establish a *Non-Informativeness Growing Quality* property: The quality of a non-informative output grows with the size of its positive class:

$$Q(S_{\neg i}) \sim |S_{\neg i}|$$

Obviously, this property is not compatible with the previous ones.

Note that our three usage scenarios exemplify that a task specification (name ambiguity resolution in our case) is not enough to select appropriate evaluation measures; we also need to specify how the output of the system is going to be used to determine how it should be evaluated.

[7] Note that if the measure also satisfies the monotonicity axiom, this constant will be low.

Table 3 Basic constraints, properties and measures

	Axiom	Properties		
	Strict monotonocity	Absolute weighting	Non-inf. fixed quality	Non-inf growing quality
Utility measures				
Acc (weighted)	✓	✓	✗	✗
Utility	✓	✓	✗	✗
Informativeness measures				
Lam%	✗	✗	✓	✗
Odds	✗	✗	✓	✗
Phi, MAAC, KapS, Chi, MI	✓	✗	✓	✗
Class-oriented measures				
F measure	✗	✗	✗	✓
Reliability and sensitivity				
F(R,S)	✓	✗	✗	✗

We now turn to the formal analysis of the most popular filtering metrics in terms of the monotonicity constraint and the three mutually exclusive properties.

4 Formal analysis of measures

In this section we present an analytical study of several Filtering evaluation measures, in terms of how they satisfy the monotonicity axiom and the three mutually-exclusive properties. The outcome of the formal analysis is summarized in Table 3, and the main points are:

- All metrics in the study, except one (Reliability/Sensitivity), belong to one of the three families defined by our proposed properties. Therefore, in order to select an appropriate measure for a given scenario, a crucial step is to decide how non-informative outputs should be assessed. More specifically, what is the quality of the zero output (discarding all) with respect to the placebo output (accepting all). If they are equivalent, we must employ an informativeness-based measure. If accepting everything is better than discarding randomly, the best option is employing Precision/Recall on the positive class. If the answer depends on the relative profit/cost of each combination in the contingency matriz, then we should use a Utility-based measure.
- Some popular measures like Precision and Recall or Lam% fail to satisfy the basic *Strict Monotonocity* axiom. We will propose smoothing techniques to fix these problems in Sect. 5.

We now discuss the formal properties of each of the metrics analyzed, grouped according to the properties defined earlier.

4.1 Utility-based measures

Utility-based measures are those that can be expressed as a linear combination of the four components in the contingency matrix (Hull 1998):

$$Utility_{\alpha_1,\alpha_2,\alpha_3,\alpha_4} \equiv \alpha_1 TP + \alpha_2 TN - \alpha_3 FP - \alpha_4 FN$$

In other words, there is an absolute reward for each type of correct labeling, and an absolute penalty for each type of error. The resulting score can be scaled according to the size of the positive and negative classes in the input stream ($P(\mathcal{G})$ and $P(\neg\mathcal{G})$). In our notation, scaled true positives correspond to $P(\mathcal{S}|\mathcal{G})P(\mathcal{G})$, true negatives to $P(\neg\mathcal{S}|\neg\mathcal{G})P(\neg\mathcal{G})$, false positives to $P(\mathcal{S}|\neg\mathcal{G})P(\neg\mathcal{G})$ and false negatives to $P(\neg\mathcal{S}|\mathcal{G})P(\mathcal{G})$.

The **Accuracy** measure (proportion of correctly classified documents) and the **Error Rate** (1-Accuracy) are two particular cases of Utility measures which reward equally true positives and true negatives. The result is scaled over the input stream size. Implicitly, accuracy penalizes also the false positive and false negatives. The Accuracy measure can be expressed in terms of conditional probabilities:

$$Acc(\mathcal{S}) = \frac{TP + TN}{T} \simeq P(\mathcal{S}|\mathcal{G})P(\mathcal{G}) + P(\neg\mathcal{S}|\neg\mathcal{G})P(\neg\mathcal{G})$$

In Androutsopoulos et al. (2000), a weighted version of Accuracy is proposed:

$$WAcc(\mathcal{S}) = \frac{\lambda TP + TN}{\lambda(TP + FN) + TN + FP} \simeq \frac{\lambda P(\mathcal{S}|\mathcal{G})P(\mathcal{G}) + P(\neg\mathcal{S}|\neg\mathcal{G})P(\neg\mathcal{G})}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})}$$

Basically, Weighted Accuracy is a Utility measure which assigns a relative weight to true positives and normalizes the score according to the ratio of relevant documents in the input stream.

The most common **Utility** version assigns a relative α weight between true positives and false positives:

$$Utility(\mathcal{S}) = \alpha TP - FP \simeq \alpha P(\mathcal{S}|\mathcal{G})P(\mathcal{G}) - P(\mathcal{S}|\neg\mathcal{G})P(\neg\mathcal{G})$$

A drawback of Utility is that the range of possible scores varies depending on the size of the dataset. Several normalization methods have been proposed (Hull 1998; Hoashi et al. 2000). In general, they consider the maximum score that can be achieved in each input stream. We do not tackle this issue here.

4.1.1 Axioms and properties

With respect to the *Strict Monotonocity Axiom*, adding an irrelevant document to the output $\mathcal{S}$ reduces true negatives, and adding a relevant document increases true positives. Therefore, accuracy satisfies this constraint. A similar reasoning can be applied to the traditional Utility measure.

The characteristic of Utility-based metrics is that it is possible to assign an absolute weight to relevant (versus non relevant) documents in the output regardless of the output size. Thus, they satisfy the *Absolute Weighting* property (see proof in the "Appendix"). But note that, although Accuracy can be considered a Utility-based measure, it does not directly satisfy the *Absolute Weighting* property, given that its definition does not include any parameter. However, the weighted accuracy proposed in Androutsopoulos et al. (2000) does satisfy this property, and it is a generalization of Accuracy (see proof in the

"Appendix"). In summary, utility-based measures assign growing or decreasing scores to non-informative outputs depending on the measure parameterization and the ratio of relevant documents in input stream (see details in the *Proofs* appendix).

4.2 Informativeness-based measures

This family of measures satisfies the *Non-Informativeness Fixed Quality* property. That is, they score equally any non-informative solution. We first focus on Lam%, which is possibly the most popular metric in this family in document filtering scenarios. Then we also analyze other metrics in this family.

Lam% **(Logistic average misclassification rate)** was defined for the problem of spam detection as the geometric mean of the odds ratio of misclassified ham, ($P(\neg S|\mathcal{G})$) and ratio of misclassified spam ($P(S|\neg\mathcal{G})$). Maximum Lam% represents minimum quality:

$$lam\% = logit^{-1}\left(\frac{logit(P(\neg S|\mathcal{G})) + logit(P(S|\neg\mathcal{G}))}{2}\right)$$

$$logit(x) = log\left(\frac{x}{1-x}\right) \quad logit^{-1}(x) = \frac{e^x}{1+e^x}$$

With respect to the *Strict Monotonocity Axiom*, a well-known problem of Lam% is that when either $P(\neg S|\mathcal{G})$ or $P(S|\neg\mathcal{G})$ are zero, *lam%* is minimal (i.e. maximal quality) regardless of the other measure component (Qi et al. 2010). This behavior implies that the Strict Monotonocity Axiom is not satisfied by Lam%.

This is a problem that could be fixed with smoothing methods. Consider, for example, an output with ten positive labels ($|S| = 10$) that all correspond to true relevant documents. This can easily be reached in practice by establishing a very high classification threshold; then the system will retrieve very few documents, but most likely they will all be relevant, and therefore spam misclassification ($P(S|\neg\mathcal{G})$) will be zero. But having zero misclassified spam documents in our data does not imply that the true probability of misclassification is zero when we provide a different set of documents to the classifier; it only means that it is very low. Therefore, a possible solution is to apply some smoothing mechanism for the estimation of $P(S|\neg\mathcal{G})$. We will tackle this issue in Sect. 5.

From the point of view of measure properties, Lam% assigns a fixed score to every non-informative system output ($Lam\%(S_{\neg i}) = 0.5$),and therefore satisfies the *Non-Informativeness Fixed Quality* property. The Lam% score for non-informative outputs is always 0.5 (see Lam% results for non-informative outputs in Fig. 2)—see proof in the "Appendix".

We have used Lam%, as a representative masure of its family, in the empirical study reported in Sect. 6. But let us review the formal properties of other common measures that also assign a fixed value to all non-informative system outputs.

The Phi correlation coefficient is expressed in terms of false and true positives and negatives (TP,FP,TN and FN)[8]:

$$Phi = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}$$

[8] For the sake of readability, we use here the traditional notation for the contingency matrix components.

Phi is always zero if $S_{\neg_i}$ is non informative (see proof in the "Appendix").

Something similar happens with the odds ratio (Karon and Alexander 1958):

$$Odds(S_{\neg_i}) = \frac{TP \times TN}{FN \times FP}$$

which is 1 for every non-informative system output (see proof in the "Appendix"). As well as the original Lam% measure, the Odds ratio does not satisfy the *Growing Quality* constraint: if $TN = 0$, then the measure is not sensitive to TP.

The Macro Average Accuracy (Mitchell 1997) is a modified Accuracy measure (MAAc) that also gives the same results for any non-informativeness measure. It is defined as the arithmetic average of the partial accuracies of each class:

$$MAAc(S_{\neg_i}) = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} = \frac{P(S|\mathcal{G}) + P(\neg S|\neg\mathcal{G})}{2}$$

Its value is always $\frac{1}{2}$ if the output is non-informative (see proof in the "Appendix"). Note that, in spite of its name, MAAc is not a utility measure, as it is not a linear combination of the components of the contingency matrix.

The Kappa statistic (Cohen 1960) is another example of informativeness-based measure. Kappa is defined as:

$$KapS(S) = \frac{\text{Accuracy} - \text{Random Accuracy}}{1 - \text{Random Accuracy}}$$

where the Random Accuracy represents the Accuracy obtained randomly by an output with size $|S|$. This measure returns zero for any non-informative output (see proof in the "Appendix").

Another metric in this family is the Chi square test statistic:

$$\begin{aligned} Chi(S) =& \frac{(|S \cap \mathcal{G}| \times |\neg S \cap \neg\mathcal{G}| - |S \cap \neg\mathcal{G}| \times |\neg S \cap \mathcal{G}|) + |T|}{|S| + |\mathcal{G}| + |\neg S| + |\neg\mathcal{G}|} \\ =& \frac{(P(S|\mathcal{G}) \times P(\neg S|\neg\mathcal{G}) - P(S|\neg\mathcal{G}) \times P(\neg S|\mathcal{G})) + 1}{2} \end{aligned}$$

which returns $\frac{1}{2}$ for any non-informative output (see proof in the "Appendix").

Finally, Mutual Information (MI):

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

can be applied to the evaluation of binary classifiers as follows:

$$\begin{aligned} MI(S) =& P(S \wedge \mathcal{G}) log \frac{P(S \wedge \mathcal{G})}{P(S)P(\mathcal{G})} + P(\neg S \wedge \mathcal{G}) log \frac{P(\neg S \wedge \mathcal{G})}{P(\neg S)P(\mathcal{G})} \\ &+ P(S \wedge \neg\mathcal{G}) log \frac{P(S \wedge \neg\mathcal{G})}{P(S)P(\neg\mathcal{G})} + P(\neg S \wedge \neg\mathcal{G}) log \frac{P(\neg S \wedge \neg\mathcal{G})}{P(\neg S)P(\neg\mathcal{G})} \end{aligned}$$

If an output $S_{\neg_i}$ is non-informative, every component in the sum is zero (e.g. $P(S_{\neg_i})P(\mathcal{G}) = P(S_{\neg_i} \wedge \mathcal{G})$, and then the log of the fraction is zero)). Therefore, $MI(S_{\neg_i})$ is zero.

4.3 Class-oriented measures: precision and recall

The third measure family includes those that assume some asymmetry between classes. These measures are suitable for applications where one class is of more interest than the other, as is the case of Information Retrieval (Sokolova et al. 2006) and Information Filtering tasks.

The most representative measure in this family is the combination of Precision and Recall for the relevant class. We will focus here on their combination via F measure (Van Rijsbergen 1974), which is a weighted harmonic mean of Precision and Recall, although the same conclusions are valid for the product (Hull 1997). The F measure is computed as:

$$F_\alpha(S) = \frac{1}{\frac{\alpha}{P(\mathcal{G}|S)} + \frac{1-\alpha}{P(S|\mathcal{G})}}$$

where $P(\mathcal{G}|S)$ and $P(S|\mathcal{G})$ are Precision and Recall respectively, in our probabilistic notation. α is a parameter that sets their relative weight, with $\alpha = 0.5$ giving the same weight to both.

4.3.1 Axioms and properties

As Sebastiani proved (Sebastiani 2015), the F measure does not satisfy the *Strict Monotonocity Axiom*. The reason is that Precision is not able to distinguish between outputs that contain only irrelevant documents: it is zero for any output without relevant documents. If we take an output without true positives and we correctly move an item from the set of false positives to the set of true negatives, F does not improve, because Precision remains zero. In Sect. 5 we discuss and propose smoothing methods to address this problem.

The F score is the only measure in our analysis that satisfies the *Non-Informativeness Growing Quality*: for any non-informative output, its F score is higher if it returns a larger set of positive labels (see proof in the "Appendix").

As we showed in Sect. 3, this property is not compatible with the other two properties (*Non Informativeness Fixed Quality* and *Absolute Weighting*).

4.4 Reliability and sensitivity

Reliability and *Sensitivity* (R and S) (Amigó et al. 2013) is a precision/recall measure pair which can be used for filtering, ranking and clustering, and also for general document organization problems that combine these three tasks. It is a generalization of the BCubed Precision and Recall metrics (used to evaluate Clustering systems) (Amigó et al. 2009). *Reliability* measures to what extent, for a given item, its relationships with other items predicted by the system do exist in the gold standard. Reversely, *Sensitivity* computes to what extent, for a given item, its true relationships with other items are predicted by the system output. An average over all items d in the dataset gives BCubed Precision and Recall overall scores.

For every document d, R and S are computed as:

$$\text{Reliability}\ (d) \equiv P_{d'}(r_g(d,d')|r_s(d,d'))$$
$$\text{Sensitivity}\ (d) \equiv P_{d'}(r_s(d,d')|r_g(d,d'))$$

where $r_g(d, d')$ and $r_s(d, d')$ are relationships between d and d' in the gold-standard and in the system output, respectively. P_d stands for the Probability measured on the sample space of all possible documents d.

Two types of binary relationships are considered in the original formulation: priority (item 1 is more relevant than item 2) and relatedness (item 1 and item 2 are related). The projection of Reliability and Sensitivity to Clustering uses only relatedness relationships, and is equivalent to BCubed Precision and Recall. The projections to Ranking and Filtering tasks use only priority relationships: Ranking obtains priority relationships from (graded) relevance assessments, and Filtering from the binary classes: items in the positive class are more relevant than items in the negative class.

In the case of filtering tasks, any positive document has more priority than any negative document. Therefore Reliability is computed as the probability of true positives ($P(\mathcal{G} \wedge \mathcal{S})$) multiplied by their ratio of correct relationships (i.e. the probability of irrelevant documents within the discarded set, $P(\neg\mathcal{G}|\neg\mathcal{S})$) plus the probability of true negatives ($P(\neg\mathcal{G} \wedge \neg\mathcal{S})$) multiplied by their correct relationships (the probability of relevant documents within the accepted documents, $P(\mathcal{G}|\mathcal{S})$):

$$\begin{aligned} \mathit{Reliability}(\mathcal{S}) &= P(\mathcal{G} \wedge \mathcal{S})P(\neg\mathcal{G}|\neg\mathcal{S}) + P(\neg\mathcal{S} \wedge \neg\mathcal{G})P(\mathcal{G}|\mathcal{S}) \\ &= P(\mathcal{G}|\mathcal{S})P(\mathcal{S})P(\neg\mathcal{G}|\neg\mathcal{S}) + P(\neg\mathcal{G}|\neg\mathcal{S})P(\neg\mathcal{S})P(\mathcal{G}|\mathcal{S}) \\ &= (P(\mathcal{S}) + P(\neg\mathcal{S}))P(\mathcal{G}|\mathcal{S})P(\neg\mathcal{G}|\neg\mathcal{S}) = P(\mathcal{G}|\mathcal{S})P(\neg\mathcal{G}|\neg\mathcal{S}) \end{aligned}$$

which is the product of precisions over both the positive and the negative classes:

$$\mathit{Reliability}(\mathcal{S}) = \mathit{Precision}_{\mathcal{S}} \times \mathit{Precision}_{\neg\mathcal{S}}$$

Replacing $\mathcal{S}$ with $\mathcal{G}$, we obtain an analogous result for Sensitivity, which corresponds to the product of both recalls:

$$\mathit{Sensitivity}(\mathcal{S}) = P(\mathcal{S}|\mathcal{G})P(\neg\mathcal{S}|\neg\mathcal{G}) = \mathit{Recall}_{\mathcal{G}} \times \mathit{Recall}_{\neg\mathcal{G}}$$

In the literature, Reliability and Sensitivity are usually combined via the F measure or weighted harmonic mean *F(R,S)*. *F(R,S)* has been used in the context of Online Reputation Management evaluation campaigns (Amigó et al. 2012) to evaluate filtering tasks where texts containing the (ambiguous) name of an entity of interest have to be classified as referring to the entity or not. Remarkably, it is the only metric in our study that do not belong to any of the three metric families induced by our proposed mutually-exclusive properties. In Sect. 6 we will see that, on the other hand, F(R,S) has empirical advantages over the rest of metrics in our study.

4.4.1 Axioms and properties

The *Strict Monotonicity Axiom* is not satisfied by F(R,S), because Reliability is zero in all cases without true positives. If we move a document from the false positives to the true negatives, the axiom requires that F(R,S) should increase. In fact, Precision on the negative class (which is a component of Reliability) increases; but precision on the positive class remains zero and dominates the product (Reliability is zero) and the harmonic mean with Sensitivity is also zero, as is also the case of F(P,R).

Reliability and Sensitivity (combined via F measure) are the only metric pair that does not satisfy any of our mutually-exclusive properties, and therefore does not fit into any of our metric families:

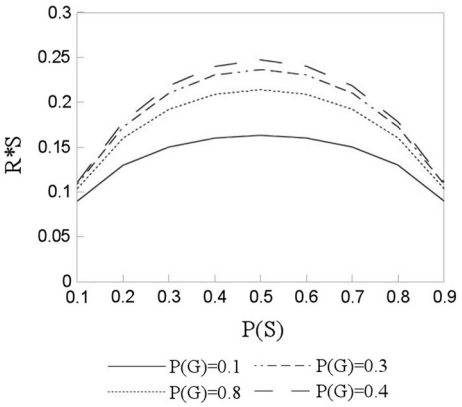

Fig. 4 F(R,S) scores for non-informative outputs. The horizontal axis represents the amount of randomly selected documents returned by the system output (i.e. labeled as positive). The vertical axis represents the F(R,S) score. Each curve represents a certain ratio of relevant documents in the input stream

- Both the zero (everything negative) and placebo (everything positive) outputs, which are non-informative, receive the minimal score, because they are not able to identify any priority relationship. Therefore, F(R,S) does not satisfy the *Non-Informativeness Growing Quality* of class-oriented metrics.
- On the other hand, not every non-informative output receives the same score. Any random distribution of documents into S and $\neg S$ produces some correct relationships by chance. Therefore, F(R,S) does not belong to the class of informativeness-based metrics.
- There is no parameter to define the relative weight of classification decisions, and therefore the metric does not belong to the utility-based family of measures.

Analytically, we can nevertheless say a couple of things about how F(R,S) handles non-informative outputs:

- Assigning all input documents to the same class (what we call *zero* and *placebo* non-informative baselines) produces a minimum score, given that at least one of the precision or recall for one of the classes is zero.
- Every non-informative output receives a score below 0.25 (see proof in the "Appendix").

Figure 4 illustrates the behavior of F(R,S) for non-informative outputs. The horizontal axis represents the amount of randomly selected documents returned (labeled positive) by the system output. The vertical axis represent the F(R,S) score. Each curve corresponds to a system output which returns a given ratio of positive labels. As the figure shows, the highest possible value of F(R,S) is 0.25, when the random assignment gives half of the items to the positive class.

5 Smoothing measures

As we have seen in the previous section, F-measure and Lam% fail to satisfy the *Strict Monotonicity* axiom. According to our probabilistic interpretation, the reason why F-measure and Lam% fail to satisfy the basic constraints is related to how conditional probabilities are estimated over just a few samples. For instance, if a system output S contains 10 positive

Table 4 Laplace's correction applied to the contingency matrix

	Relevant docs.	Irrelevant docs.
Returned docs	$\|\mathcal{S} \cap \mathcal{G}\| + 1$	$\|\mathcal{S} \cap \neg\mathcal{G}\| + 1$
Returned docs	$\|\neg\mathcal{S} \cap \mathcal{G}\| + 1$	$\|\neg\mathcal{S} \cap \neg\mathcal{G}\| + 1$

documents ($|\mathcal{S}| = 10$) and they are all irrelevant ($|\mathcal{S} \cap \mathcal{G}| = 0$), then the true probability of finding a relevant document in the output ($P(\mathcal{G}|\mathcal{S})$) should be some unknown value, lower than $\frac{1}{10}$, but not necessarily zero. Actually, zero is the less reliable estimation, because, over a large enough dataset, we will likely find, purely by chance, at least one relevant document. This reasoning can be applied to other conditional probabilities implicit in measures such as recall $P(\mathcal{S}|\mathcal{G})$ or precision $P(\mathcal{G}|\mathcal{S})$.

In general, we assume that the implicit estimation of conditional probabilities in all measures should be revised when the ratio of relevant documents ($\frac{|\mathcal{G}|}{|T|}$) or positive system outputs ($\frac{|\mathcal{S}|}{|T|}$) is extremely low or high.

We now turn to discuss which is the best way of smoothing F-measure and Lam% to make them compliant with our formal constraints.

5.1 Laplace's correction

A popular smoothing mechanism to is *Laplace's correction*. Assuming that all the components in the contingency matrix are equally likely (prior knowledge), this method simply adds one unit per component. Table 4 shows how Laplace's correction is applied to the contingency matrix.

This correction ensures that all matrix components are always larger than zero. The resulting estimation for Precision ($P(\mathcal{G}|\mathcal{S})$) is:

$$P(\mathcal{G}|\mathcal{S}) = \frac{|\mathcal{S} \cap \mathcal{G}| + 1}{|\mathcal{S}| + 2}$$

This smoothed F-measure satisfies the *Strict Monotonicity Axiom*. Unlike the original, non-smoothed version, now Precision is never zero, and therefore F can always decrease when adding irrelevant documents to the output.

All metrics considered in our analysis can be smoothed in a similar way. The smoothed Lam% measure also satisfies the axiom, given that the misclassified relevant $P(\neg\mathcal{S}|\mathcal{G})$ and the misclassified irrelevant documents $P(\mathcal{S}|\neg\mathcal{G})$ are never zero. Therefore, it is necessary to reduce both in order to optimize the Lam% score. Returning a reduced set of relevant documents is no longer enough to maximize the score.

A problem of Laplace's correction is that assuming that all the components in the contingency matrix are equiprobable may not be a good prior. For instance, suppose that relevant documents are extremely unfrequent (e.g. $|\mathcal{G}| = 1$ and $|T| = 100{,}000$). Then Laplace's correction assumes that the system is able to capture the unique relevant document by adding one unit in $|\mathcal{S} \cap \mathcal{G}| \equiv TP$. This assumption leads to an artificial effect of system informativeness: a smoothed non-informative output becomes an informative output. Therefore, the properties which are grounded on the behavior of measures over non-informative outputs are not preserved.

Let us consider the zero output $\mathcal{S}_\emptyset$, which is non-informative, to further illustrate this problem. Its smoothed version $P(\mathcal{G}|\mathcal{S}_\emptyset)$ does not preserve non-informativeness:

$$P(\mathcal{G}|S_\emptyset) = \frac{|S_\emptyset \cap \mathcal{G}| + 1}{|S_\emptyset| + 2} = \frac{1}{2} \neq P(\mathcal{G})$$

Recall that non-informative outputs are those for which $P(\mathcal{G}|S_{\neg i}) = P(\mathcal{G})$.

Let us analyze the consequences of applying this smoothing method to the F measure. In principle, a class-oriented measure such as F(P,R) prefers non-informative outputs that discard less documents, and therefore the F score for S_T (Placebo) is always higher than the score for the Zero system $S_\emptyset$. But the smoothed Precision and Recall for the non-informative outputs S_T (Placebo) and Zero system are:

$$\begin{aligned}
Recall_{Smooth}(S_\emptyset) &= \frac{TP+1}{TP+1+FN+1} = \frac{|S_\emptyset \cap \mathcal{G}| + 1}{|S_\emptyset \cap \mathcal{G}| + 1 + |\neg S_\emptyset \cap \mathcal{G}| + 1} \\
&= \frac{|\emptyset \cap \mathcal{G}| + 1}{|\emptyset \cap \mathcal{G}| + 1 + |\mathcal{T} \cap \mathcal{G}| + 1} = \frac{1}{|\mathcal{G}| + 2} \\
Recall_{Smooth}(S_\mathcal{T}) &= \frac{TP+1}{TP+1+FN+1} = \frac{|\mathcal{T} \cap \mathcal{G}| + 1}{|\mathcal{T} \cap \mathcal{G}| + 1 + |\emptyset \cap \mathcal{G}| + 1} = \frac{|\mathcal{G}| + 1}{|\mathcal{G}| + 2} \\
Precision_{Smooth}(S_\emptyset) &= \frac{TP+1}{TP+1+FP+1} = \frac{|S_\emptyset \cap \mathcal{G}| + 1}{|S_\emptyset \cap \mathcal{G}| + 1 + |S_\emptyset \cap \neg\mathcal{G}| + 1} = \frac{1}{2} \\
Precision_{Smooth}(S_\mathcal{T}) &= \frac{TP+1}{TP+1+FP+1} = \frac{|\mathcal{T} \cap \mathcal{G}| + 1}{|\mathcal{T} \cap \mathcal{G}| + 1 + |\mathcal{T} \cap \neg\mathcal{G}| + 1} = \frac{|\mathcal{G}| + 1}{|T| + 2}
\end{aligned}$$

Therefore, the smoothed Recall is still higher in the Placebo output ($S_\mathcal{T}$) than in the Zero output:

$$\frac{|\mathcal{G}| + 1}{|\mathcal{G}| + 2} > \frac{1}{|\mathcal{G}| + 2}$$

However, if $|\mathcal{G}| < \frac{|T|}{2}$ then the precision for the Zero output is higher:

$$|\mathcal{G}| < \frac{|T|}{2} \Longrightarrow |\mathcal{G}| < \frac{|T|+2}{2} - 1 \Longrightarrow \frac{|\mathcal{G}| + 1}{|T| + 2} < \frac{1}{2}$$

Therefore, depending on the relative weight of Precision in F (α value), the zero system $S_\emptyset$ can outperform the Placebo system S_T and the *Non-Informativeness Growing Quality* property is not preserved.

A similar problem occurs when we apply Laplace's correction over the Lam% measure. Adding one element to each component in the contingency matrix may transform a non-informative output into an informative output, achieving a Lam% score different than the constant value that any non-informative output should achieve (0.5). Therefore, the *non-informativeness fixed quality* property that characterizes Lam% is not preserved.

5.2 Non-informative smoothing

In order to comply with the strict monotonicity axiom while preserving the other properties of metrics, we propose to *assume non-informativeness as prior knowledge*. We will use it here to

modify Laplace's correction, but the same reasoning can be applied to other smoothing techniques (Agresti and Hitchcock 2005).

Non-informativeness implies that the discrete variables (or sets) $\mathcal{S}$ and $\mathcal{G}$ are independent of each other. Therefore we can add the joint probability to each matrix component in this way:

	Relevant docs.	Irrelevant docs.
Returned docs	$\lvert\mathcal{S}\cap\mathcal{G}\rvert + P(\mathcal{G})P(\mathcal{S})$	$\lvert\mathcal{S}\cap\neg\mathcal{G}\rvert + P(\neg\mathcal{G})P(\mathcal{S})$
Returned docs	$\lvert\neg\mathcal{S}\cap\mathcal{G}\rvert + P(\mathcal{G})P(\neg\mathcal{S})$	$\lvert\neg\mathcal{S}\cap\neg\mathcal{G}\rvert + P(\neg\mathcal{G})P(\neg\mathcal{S})$

The resulting computation for Precision $P(\mathcal{G}|\mathcal{S})$ is:

$$P(\mathcal{G}|\mathcal{S}) = \frac{|\mathcal{S}\cap\mathcal{G}| + P(\mathcal{G})P(\mathcal{S})}{|\mathcal{S}| + P(\mathcal{S})}$$

$P(\mathcal{G})$ represents the prior knowledge about $P(\mathcal{G}|\mathcal{S})$. That is, a priori, the system is non-informative and the ratio of relevant documents in the output corresponds to the ratio of relevant documents in the input stream. On the other hand, $P(\mathcal{S})$ represents the weight assigned to the prior knowledge. Assigning the same weight to the prior knowledge as Laplace's correction, we obtain:

$$P(\mathcal{S}|\mathcal{G}) = \frac{|\mathcal{S}\cap\mathcal{G}| + 2P(\mathcal{G})}{|\mathcal{S}| + 2}$$

Note that this value is equivalent to Laplace's correction when $P(\mathcal{G}) = \frac{1}{2}$, slightly larger if $P(\mathcal{G})$ grows, an slightly lower if $P(\mathcal{G})$ is lower than $\frac{1}{2}$.

Now, given a non-informative system $\mathcal{S}_{\neg i}$, we have the following smoothed conditional probability estimation, which preserves the nature of a non-informative output:

$$\begin{aligned} P(\mathcal{G}|\mathcal{S}_{\neg i}) &= \frac{|\mathcal{S}_{\neg i}\cap\mathcal{G}| + 2P(\mathcal{G})}{|\mathcal{S}_{\neg i}| + 2} = \frac{|T|P(\mathcal{S}_{\neg i})P(\mathcal{G}) + 2P(\mathcal{G})}{|\mathcal{S}_{\neg i}| + 2} \\ &= \frac{P(\mathcal{G})(|T|P(\mathcal{S}_{\neg i}) + 2)}{|\mathcal{S}_{\neg i}| + 2} = \frac{P(\mathcal{G})(|\mathcal{S}_{\neg i}| + 2)}{|\mathcal{S}_{\neg i}| + 2} = P(\mathcal{G}) \end{aligned}$$

Therefore, the condition $P(\mathcal{G}|\mathcal{S}_{\neg i}) = P(\mathcal{G})$ is preserved and the measure properties are not affected.

We apply the same procedure to all the conditional probabilities:

$$\begin{aligned} P(\mathcal{S}|\mathcal{G}) &= \frac{|\mathcal{S}\cap\mathcal{G}| + 2P(\mathcal{S})}{|\mathcal{G}| + 2} = 1 - P(\neg\mathcal{S}|\mathcal{G}) \\ P(\mathcal{S}|\neg\mathcal{G}) &= \frac{|\mathcal{S}\cap\neg\mathcal{G}| + 2P(\mathcal{S})}{|\neg\mathcal{G}| + 2} = 1 - P(\neg\mathcal{S}|\neg\mathcal{G}) \\ P(\mathcal{G}|\mathcal{S}) &= \frac{|\mathcal{S}\cap\mathcal{G}| + 2P(\mathcal{G})}{|\mathcal{S}| + 2} = 1 - P(\neg\mathcal{G}|\mathcal{S}) \\ P(\mathcal{G}|\neg\mathcal{S}) &= \frac{|\neg\mathcal{S}\cap\mathcal{G}| + 2P(\mathcal{G})}{|\neg\mathcal{S}| + 2} = 1 - P(\neg\mathcal{G}|\neg\mathcal{S}) \end{aligned}$$

The non-informative smoothed versions (i.e. $F_{sm_{\neg i}}$ or $Lam\%_{sm_{\neg i}}$) are computed in the same way as the original measures, but using the informativeness-based smoothing procedure when estimating the previous conditional probabilities.

Table 5 Basic constraints, properties and measures

	Axiom	Properties		
	Strict monotonicity	Absolute weighting	Non-inf. fixed quality	Non-inf. growing quality
Utility measures				
Weighted Accuracy	✓	✓	✗	✗
Utility	✓	✓	✗	✗
Informativeness measures				
Lam%	✗	✗	✓	✗
Odds	✗	✗	✓	✗
$\text{Lam\%}_{smL}$	✓	✗	✗	✗
Odds_{smL}	✓	✗	✗	✗
$\text{Lam\%}_{smN}$	✓	✗	✓	✗
Odds_{smN}	✓	✗	✓	✗
Phi, MAAC, KapS, Chi, MI	✓	✗	✓	✗
Class-oriented measures				
F measure	✗	✗	✗	✓
F_{smL}	✓	✗	✗	✗
F_{smN}	✓	✗	✗	✓

Table 5 shows the properties and constraints satisfied by measures and their smoothed versions. The Laplace smoothed version is represented by the subindex X_{smL}. The subindex X_{smN} represents the informativeness-based smoothing. As the table shows, Laplace's smoothing fixes compliance with the basic axiom, but at the cost of breaking the natural properties of measures. Informativeness-based smoothing also makes measures compliant with monotonicity, but also preserves the way they handle non-informative outputs.

In conclusion, if we expect very large or very small sets of positive documents in the system output, or a very large fraction of relevant documents in the dataset, we need to apply a smoothing method to preserve strict monotonicity, and we can apply our informativeness-based smoothing in order to preserve the original properties of metrics.

6 Experiments

In addition to the formal analysis, which is the primary contribution of our work, we want to further characterize and compare the behavior of metrics empirically. First of all, we want to study the empirical behavior of the smoothed versions of the F-measure and Lam% that we have proposed purely on formal arguments. Then we will compare measures in terms of their strictness, their robustness across data sets, and in terms of how they rank systems.

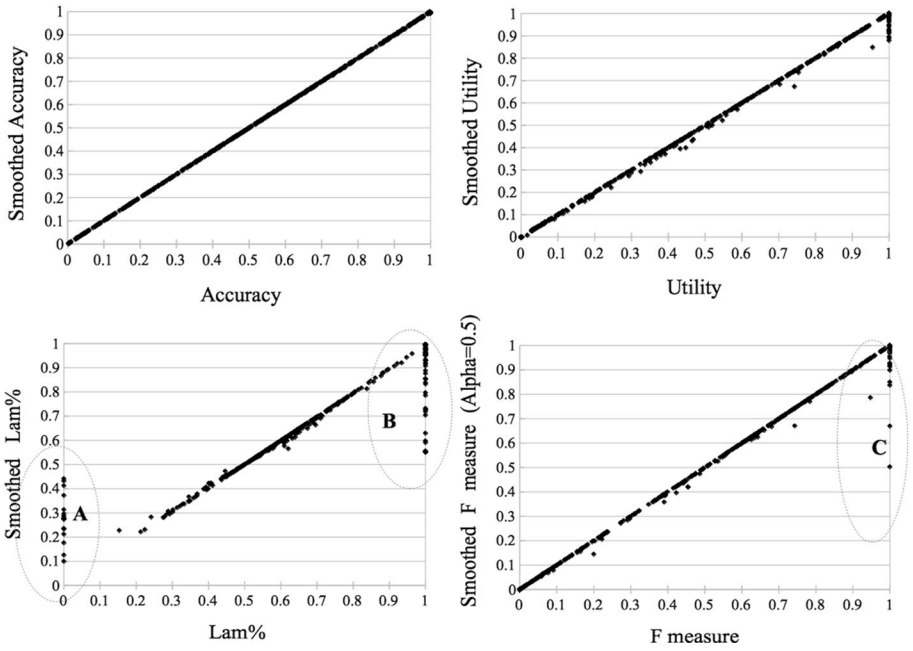

Fig. 5 The effect of smoothing Accuracy, Utility, F measure and Lam% using the informativeness-based correction. Each dot corresponds to a system output for one test case (one company name)

6.1 Experimental setting

For our experiments, we have employed the evaluation corpus and system results from the second task in the WePS3 competition, *Online Reputation Management* (Amigó et al. 2010). Given a company name and a stream of tweets containing the name, the task consisted of classifying Twitter entries (Krishnamurthy et al. 2008) as relevant (related) when they refer to a certain company and irrelevant (unrelated) otherwise.

The test set includes tweets for 47 companies and the training set comprises 52 company names. For each company, around 400 tweets were retrieved using the company name as query. The training and test corpora were crowdsourced using Mechanical Turk (Le et al. 2010) using five annotations per tweet with reasonable inter-annotator agreement rates. The ratio of related tweets per company name varies widely across companies, which suits our purposes well. The statistics are described in Amigó et al. (2010). We will refer to each test case (tweets for a company) as an *input stream* or *topic*. Five research teams participated in the competition, and sixteen runs were evaluated. The organizers included two naive baseline systems: the placebo system (all tweets are about the company) and its opposite (no tweet is about the company).

6.2 The effect of smoothing

We want to investigate empirically what is the effect of applying non-informative smoothing to the evaluation measures, and how it compares to the standard Laplace's correction.

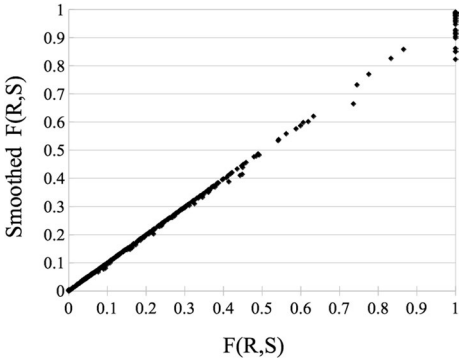

Fig. 6 The effect of smoothing Reliability and Sensitivity (F(R,S)) using the informativeness based correction. Each dot corresponds to a system output for one test case (one company name)

Figure 5 shows the relationship between smoothed and non-smoothed measures. We have inverted the Lam% values (i.e. we use 1-Lam%) for an easier interpretation of the graph. In the case of Utility, we have used its normalized version (Hull 1998; Hoashi et al. 2000) in order to have comparable results across test cases. Each dot represents a single system output for a single test case (a company name). The horizontal axis represents the original measures and the vertical axis represent the smoothed versions using the Laplace's technique. Note that we can apply the smoothing procedure to any measure which is computed from the contingency matrix, including Accuracy or Utility.

in the case of Utility-based metrics (Accuracy and Utility in the figure), the smoothing has little practical impact. Accuracy, in fact, does not change for any system output. In the case of Utility, there are only a few cases where smoothed utility gives a slightly lower score. The reason is that the probabilities are not computed over single classes, and therefore the imbalances in the data do not have a significant effect on the probability computation.

With respect to the F measure, the overall effect is similar to Utility: in just a few cases, the smoothed version gives a slightly lower score. The only exception is the dot marked as C, where the smoothed version is around 10% lower than the original F score.

The sharpest difference occurs in the case of Lam%. In general, the correlation is almost perfect; but in the extreme values of Lam% the situation changes drastically. When Lam%=1 (region B in the figure), its smoothed version can be anywhere from 1 to near 0.5 (which is the score for non-informative outputs). Recall that Lam% overscores system outputs without misclassified irrelevant documents $P(\mathcal{S}|\neg\mathcal{G}) = 0$ even if not all relevant documents appear in the output. The smoothed version solves this, and therefore some outputs that receive a high Lam% score are penalized by the smoothed version. Reversely, when Lam%=0 (region A in the figure) the smoothed version can be anywhere from 0.1 to almost 0.5.

Figure 6 shows that Reliability and Sensitivity also modify their behavior when smoothed. Although the correlation is in general almost perfect, outputs with a perfect F(R,S) score can now receive values from 1 to almost 0.3.

Figure 7 compares the use of Laplace's correction with our informativeness-based correction. In the case of F measure, the overall correlation between both methods is high, but some outputs are penalized by the informativeness-based correction (for instance dots A and B in the figure). In these cases, the system returns only a few documents ($|\mathcal{S}| << |T|$), and there are only a few relevant documents in the dataset ($|\mathcal{G}| << |T|$). Therefore, Laplace's correction overscores the output by adding one element in the true positive

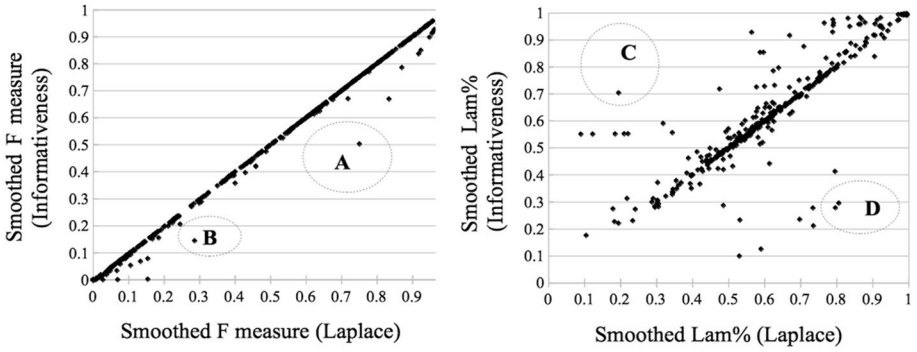

Fig. 7 Laplace's correction versus informativeness-based correction for F measure and Lam%

component of the contingency matrix. The informativeness-based correction takes into account the ratio of relevant documents in the input stream, and prevents such overscoring.

The sharpest differences between both corrections appear in the case of Lam%. In some cases, the informativeness-based correction rewards systems: for instance, if there is a large amount of relevant documents in the input stream ($|\mathcal{G}| \quad T$) and the output size is low ($|S| << |T|$), the informativeness-based correction assumes less misclassifications than Laplace's correction (e.g. dot C in the figure). On the other hand, if there are no relevant documents in the small positive output and the ratio of relevant documents is low, then the informativeness-based correction assumes more misclassifications than the original Laplace correction (dots D in the figure). In this cases, the informativeness-based correction penalizes more than the Laplace correction.

Overall, our recommendation is to use non-informative smoothing in cases where the classes in the test cases are highly imbalanced, to prevent the few cases where metrics can overestimate or understimate errors.

6.3 Strictness

In this section we follow the definition and estimation of strictness given by Amigó et al. (2013). Given a set of measures, one of them is stricter if it is a lower bound on the quality assessments of the other measures; in other words, if it penalizes systems for all flaws detected by the other measures. Consider, for instance, Accuracy and Lam%. Depending on the dataset, sometimes a high Accuracy score can be achieved just by assigning every sample to the most frequent class. Informativeness-oriented measures such as Lam%, on the other hand, penalize such strategy. Reversely, high Lam% scores can be achieved by minimizing the false negative or the false positive sets; for instance, returning only a few high-confidence samples as positive (see Sect. 4.2). With respect to this strategy, Accuracy would be stricter, as it penalizes such behavior. A measure is stricter than Accuracy and Lam% if it penalizes both types of wrong system behavior.

Within our set of measures, we say that a measure is strict if it penalizes anything that at least other (reasonable) metric penalizes. The effect is that a high score with a strict measure implies a high score according to the rest of measures is achieved.

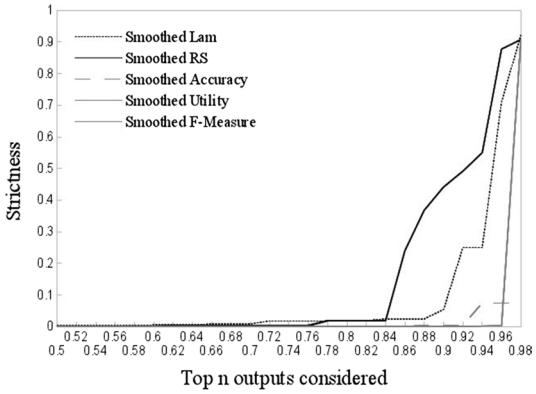

Fig. 8 Strictness of measures computed using the top $n\%$ system outputs in the ranking produced by each measure

Strictness means that a highly ranked output according to the metric is highly ranked according to any other measure. Following (Amigó et al. 2013), in order to compute the strictness of the metric m with respect to another metric m', we (1) rank all outputs from all topics according to m and m'. (2) Then, we select the top outputs according to the measure m. (3) We consider the lowest ranking position according to the other metric m' within these outputs. (4) The global strictness of each measure is the minimum strictness with respect to all the other of metrics. Formally, being $\mathcal{O}$ the set of outputs in all topics and being $rank(m, o)$ the ranking position of the output o regarding the measure m:

$$rank(m, o) = P_{o' \in \mathcal{O}}(m(o) \geq m(o'))$$

That is, the top, middle and bottom ranks are 1, 0.5 and 0 respectively. The set of top ranked outputs according to m is:

$$top(m, th) = \{o \in O | rank(m, o) \geq th\}$$

The strictness of m with respect to other metric m' is:

$$Strictness_{th}(m, m') = min_{o \in top(m,th)}(rank(m', o)))$$

and the overall strictness of m given a metric set $\mathcal{M}$ is:

$$Strictness(m) = min_{m' \in \mathcal{M}}(Strictness(m, m'))$$

Figure 8 shows the results. Each curve represents the strictness of a metric computed using the top $n\%$ values of each system output. We have considered informativeness-based smoothing variants. As the figure shows, R and S are substantially stricter than other metrics above 80% of the top ranked outputs. This means that the minimum ranking position for these outputs according to other metrics is higher than in the case of the other metrics. The second strictest metric is Lam%, which belongs to the informativeness-based measure family. Note that when the input stream is not well balanced, then the F measure and utility based metrics overscore non-informative outputs, which makes them less strict than R, S and Lam%.

In order to better understand the strictness of R, S with respect to the other metric, in Fig. 9 we have compared Reliability and Sensitivity values (combined with the F measure)

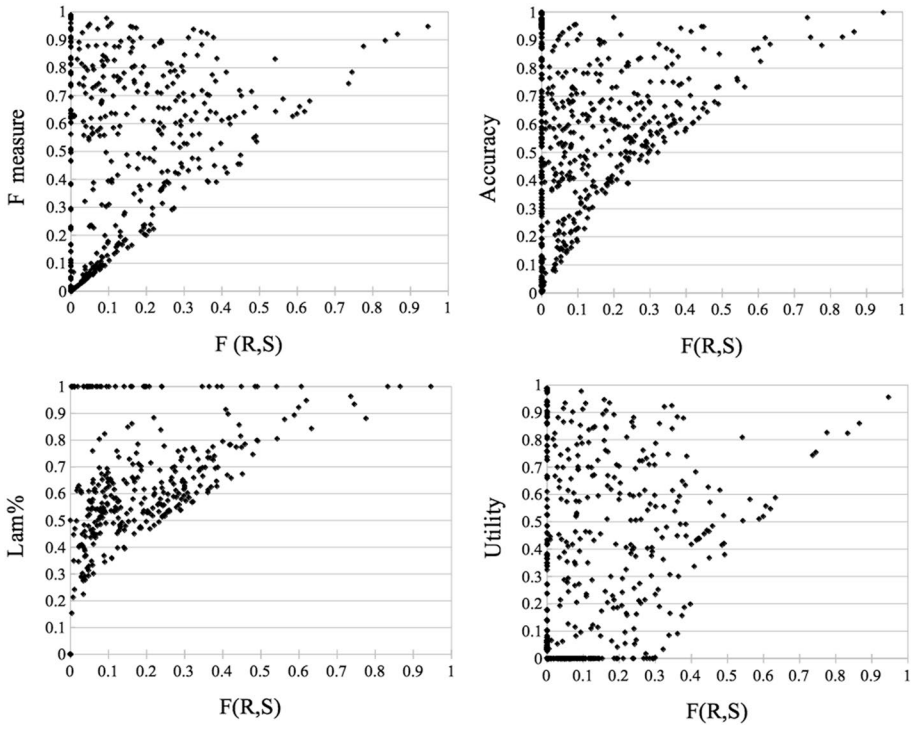

Fig. 9 F(R,S) values for every (topic,system) pair compared to other metrics

with the values of other salient measures. Each dot represents a system output, and all outputs from all systems in the dataset have been considered.[9] Note that F(R,S) is strictly lower than smoothed Lam%, accuracy and F(P,R) for virtually every system and every test case in the dataset. Only in the case of utility there is an exception in the area of low F(R,S) values (0-0.4), where there seems to be little correlation between both metrics.

We can provide an analytical explanation for the behavior of $F(R, S)$ with respect to the other metrics. First, a low precision or recall in the positive class directly implies a low R and S:

$$P(S|\mathcal{G}) << 1 \Longrightarrow P(S|\mathcal{G})P(\neg S|\neg\mathcal{G}) << 1 \Longrightarrow F(R,S) << 1$$

If the output is non-informative (and then Lam% is low, as well as any other measure in its family), then we cannot have a high precision and recall of discriminative relationships. For instance, Lam% is grounded on the ratio of misclassified documents $P(S|\neg\mathcal{G})$ and $P(\neg S|\mathcal{G})$. Then:

$$P(\neg S|\mathcal{G}) >> 0 \Longrightarrow P(S|\mathcal{G}) << 1 \Longrightarrow F(R,S) << 1$$

And finally, if most documents are false positives or false negatives (which implies a low Accuracy or Utility) then the ratio of correct relationships from all documents necessarily decreases:

[9] For easier comparison, the lam% scale has been reversed from 0 to 1.

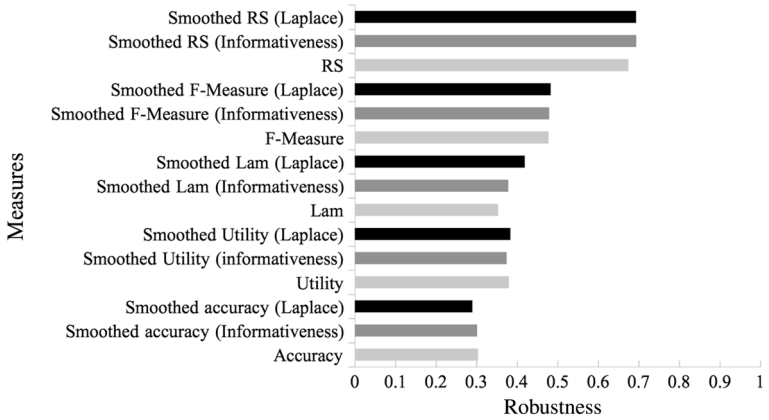

Fig. 10 Robustness of measures

$$
\begin{aligned}
TP + TN << T &\Longrightarrow P(\mathcal{S}|\mathcal{G})P(\mathcal{G}) + P(\neg\mathcal{S}|\neg\mathcal{G})P(\neg\mathcal{G}) << 1 \\
&\Longrightarrow (P(\mathcal{G}) << 1 \wedge P(\neg\mathcal{S}|\neg\mathcal{G}) << 1) \vee (P(\neg\mathcal{G}) << 1 \wedge P(\mathcal{S}|\mathcal{G}) << 1) \\
&\Longrightarrow P(\mathcal{S}|\mathcal{G}) << 1 \vee P(\neg\mathcal{S}|\neg\mathcal{G}) << 1 \Longrightarrow F(R,S) << 1
\end{aligned}
$$

Given the low correlation between filtering measures, strictness can sometimes be a highly desirable property. If a task / test collection does not prescribe how the output of the filtering system is going to be used, obtaining a high value with a strict metric guarantees that the system can be used in different usage scenarios.

6.4 Robustness across data sets

The robustness of a metric is its ability to return consistent results across different data sets. In combination with other metric properties, it can be a valuable property, because it contributes to the predictive power of an experimental outcome.

Our last experiment is an empirical assessment of the robustness of measures across datasets/test cases. As we discuss in Sect. 7, there are many ways to meta-evaluate measures according to its robustness: for instance, robustness to noise, analysis of variance (ANOVA), *consistency* or *discriminacy*. Here we follow the meta-evaluation criterion of Amigó et al. (2013), which consists of measuring the correlation of measure system rankings across topics.

For this, we compute the Spearman correlation[10] between system rankings obtained over 1000 pairs of randomly selected topics. Being $Q_m(o,t)$ the score according to the measure m for the output o in the topic t:

$$Robustness(m) = corr_{o\in\mathcal{O}}(Q_m(o,t), Q_m(o,t'))$$

The results are shown in Fig. 10. The most remarkable result is the large difference between the robustness of F(R,S) (Reliability/Sensitivity) with respect to the rest of measures.

[10] Initially we applied the Pearson coefficient. However, the results were not consistent, due to scaling issues (non-linear correlations).

Utility-based measures have low robustness, because they are very sensitive to the characteristics of the dataset (in particular, to the ratio of relevant documents in the input). And class-oriented measures (F measure) tend to be more robust than the informativeness-based measure Lam%.

With respect to the effect of the smoothing techniques, there seems to be no consistent improvement with respect to the original measures.

6.5 Ranking systems

Finally, we compare the system scores for each of the metrics,in order to illustrate how a set of systems can be ranked in an evaluation campaign. Table 6 shows all the runs in the WEPS-3 evaluation campaign ranked by Reliability and Sensitivity. Measures from other families are also included in the table. We have included a *Random baseline* system which assigns randomly half of the documents to the positive class.

All measures agree on which is the best system (LSIR.EPFL 1). Beyond that, the correspondence between rankings is lower than would be expected for metrics which have the same purpose. For instance, the Pearson correlation between the F measure and the Accuracy rankings is 0.5. The reason is that, by definition, class-oriented metrics penalize systems that are close to the zero output. For instance, SINAI 1 achieves a high accuracy but a low F measure.

An interesting question is which systems are better than a non-informative output. For each metric (column), figures in boldface are the scores that improve all non-informative baselines for that metric. According to the F measure over Precision and Recall, there is only one system that improves the upper bound of non-informative outputs. In other words, if we consider that the quality of a non-informative output is correlated with its size (*Non-informativeness growing* property) then most systems do not improve the Placebo baseline. According to Lam%, most systems improve the fixed score for non-informative outputs (0.5). Therefore, if we consider that any non-informative output is equally useless, then all systems represent an improvement over non-informative baselines. According to Accuracy, some systems improve the non-informative outputs and some systems do not. Therefore, if we consider that any correct classification decision is equally important (Accuracy) some approaches are better than non-informative outputs. Our conclusion is that measures are complementary, and that understanding the assumptions of each measure is crucial to interpret their results.

6.6 Wrap up

The outcome of our experiments provides two practical consequences on the use of filtering evaluation measures: (1) although smoothed versions are highly correlated with the original measures, using them avoids potential over and underestimations of the quality of systems in cases where the classes are highly imbalanced; and (2) Reliability and Sensitivity is the metric pair with the highest strictness and robustness of all measures considered. Therefore, if a use case does not clearly point to one of the three measure families (or if the output of the filtering system is going to be used in multiple scenarios), Reliability/Sensitivity should be the preferred metric. In any case, studying the results of different metrics provides additional insights into the behavior of systems.

Table 6 Systems in WEPS-3 evaluation campaign ranked by smoothed F(R,S)

System	Accuracy	Utility	Lam%	smoothed lam%	F	Smoothed F	Smoothed F(R,S)
LSIR.EPFL 1	**0.83**	**0.64**	**0.71**	**0.72**	**0.63**	**0.63**	**0.25**
ITC-UT 1	**0.75**	**0.52**	**0.63**	**0.64**	0.49	0.49	**0.2**
ITC-UT 3	**0.67**	**0.41**	**0.6**	**0.61**	0.41	0.4	**0.18**
UVA 1	0.56	**0.22**	**0.54**	**0.53**	0.36	0.36	**0.17**
Random	0.5	0.21	0.46	0.49	0.38	0.38	0.15
KALMAR 4	0.46	**0.34**	**0.57**	**0.56**	0.46	0.46	0.15
ITC-UT 2	**0.73**	**0.53**	**0.64**	**0.63**	0.51	0.51	0.15
ITC-UT 4	**0.64**	**0.42**	**0.61**	**0.6**	0.43	0.42	0.14
KALMAR 5	0.44	**0.35**	**0.58**	**0.56**	0.47	0.47	0.13
KALMAR 2	0.44	**0.29**	**0.55**	**0.54**	0.43	0.43	0.13
KALMAR 3	0.4	**0.26**	**0.56**	**0.55**	0.39	0.39	0.12
SINAI 1	**0.63**	**0.37**	**0.64**	**0.64**	0.29	0.29	0.11
KALMAR 1	0.48	**0.31**	**0.56**	**0.52**	0.42	0.42	0.1
SINAI 3	0.46	**0.31**	0.5	0.5	0.36	0.36	0
SINAI 5	0.51	**0.32**	0.5	0.5	0.28	0.28	0
SINAI 4	0.61	**0.3**	0.5	0.5	0.17	0.17	0
SINAI 2	0.56	0.19	0.5	0.5	0	0	0
Zero output	0.57	0.19	0.5	0.5	0	0	0
Placebo	0.43	0.4	0.5	0.5	0.53	0.53	0

For each column, figures in boldface are results that improve all non-informative baselines (random, placebo, zero output)

7 Related work

In this section, we first review related work on meta-evaluating classification measures, and then we also briefly review related work on meta-evaluation based on formal constraints for other tasks.

7.1 Measure analysis for binary classification problems

7.1.1 Sebastiani's axioms

Possibly, Sebastiani's work (2015) is the closest in spirit to our analysis. The author proposed a set of basic axioms to be satisfied by measures. The first one is the *Strict Monotonicity* axiom, which is considered in our work as the main basic constraint for measures. Sebastiani proved that the traditional F measure (on Precision and Recall) does not satisfy this. In this case, our contribution builds on this analysis, and proposes a technique that leads to a smoothed version of the F measure that satisfies the monotonicity axiom and, at the same time, preserves its other analytical properties.

His second axiom, (*Continuous Differentiability*), states that the evaluation measure must be continuous and differentiable over the true positive and true negative. We did not consider this aspect in our study. However, according to the author, measures fail to satisfy it for the case of zero values in the contingency matrix. Something similar happens with

the third and fourth axioms *Strong Definiteness* and *Weak Definiteness*, which state that the measures must be defined for any gold standard or system output. The four axioms are satisfied by the F measure when interpreting it in probabilistic terms and applying smoothing techniques. Apart from this, we can consider that the third and four constraints are inferred from the *Strict Monotonicity* axiom. Notice that the axiom states a comparison of scores which must be definable for every system output and gold standard.

The fifth axiom sets a restriction about the measure value range. We did not cover this aspect of measures in our work, given that we focus on the intrinsic measure properties rather than on scale aspects. Interestingly, the sixth and seventh axioms proposed by Sebastiani are equivalent to our *Non-Informativeness Fixed Quality* property. That is, a random or trivial classifier must achieve the same score regardless the gold standard. As we have argumented, we do not think this is a basic axiom (something that all filtering metrics should hold), but a property that helps characterizing a family of metrics. For instance, in the case where the positive class is going to be inspected by online reputation experts, the (non-informative) option of labeling everything as positive is much harmless than the (equally non-informative) option of labeling everything as negative: in the first case, the result is a substantial increase in the manpower needed to examine the positive class; but, in the second case, performing the reputation analysis simply becomes impossible.

7.1.2 Sensitivity and robustness

One criterion to compare evaluation measures is sensitivity in Analysis of Variance (Bradley 1997). Along this line, Ling presented a rigorous definition of *consistency* and *discriminacy* (Ling et al. 2003). These meta-evaluation criteria focus on the ability to capture slight differences between classifiers. In Ferri et al. (2009), measures are meta-evaluated in terms of robustness with respect to noise in system outputs (which is introduced artificially in their experimentation). In general, all these meta-evaluation criteria are oriented to the statistical consistence of evaluation measures. In contrast, our approach focuses on clarifying their analytical behavior and their underlying assumptions.

7.1.3 Grouping measures by correlation to each other

Other studies categorize measures empirically by computing their mutual correlation (Ferri et al. 2009; Caruana and Niculescu-Mizil 2005). An interesting result is that, in general, measures tend to be less correlated to each other in imbalanced data sets. This observation highlights the importance of selecting an appropriate measure when the ratio of relevant documents ($P(\mathcal{G})$ in our notation) varies across test cases.

7.1.4 Ferri's measure categorization

Ferri et al. (2009) grouped classification evaluation measures in three categories. First, some measures are based on how well the system ranks the samples (e.g ROC or AUC). We have excluded them from our study, as we focused on the evaluation of binary, discrete classification outputs, where the system must predict the optimal classification threshold. Ferri et al. distinguish between *probabilistic* measures and measures based on a *qualitative* understanding of errors. The probabilistic measures consider the deviation from the true probability of errors. These measures are closely related to our family of informativeness-based measures. The qualitative measures include both Utility based

metrics and class-oriented measures. Unlike in our study, Ferri et al. do not provide a formal distinction between measure families.

7.1.5 Caruana's measure categorization

In Caruana and Niculescu-Mizil (2005) another measure categorization is proposed. One family is "threshold measures", which groups all our three measure families, and the other two sets compare the system versus the reference ranking and therefore, they are excluded from our study. Therefore, our work can be seen as a formal investigation of the subfamilies in the first group proposed by Caruana and Niculescu-Mizil (2005).

7.1.6 Solokova's invariance properties

Solokova proposed a formal categorization of threshold measures (Sokolova 2006). She focused on the invariance of measures under a change in the contingency matrix (true positive, false positive, etc.). These properties are:

- Invariance under the exchange of true positive (*TP*) with true negative (*TN*) and false negative (*FN*) with false positive (*FP*). Absolute weighting based measures are invariant under certain weighting schemes. Measures from the other two families are in general non invariant.
- Invariance under the change in *TN* when all other matrix entries remain the same. According to the authors, all the precision/recall based measures are invariant under the change of *TN*. This property characterizes the class-oriented measure family. This is closely related to our *Non-Informativeness Growing Quality* property. Intuitively, changing the size of a non-informative output $|S_{\neg i}|$ produces a trade-off between components in the contingency matrix. If the measure is not sensitive to one of the components, then increasing the non-informative output size can be always beneficial.
- Invariance under the change in *FP* when all other matrix entries remain the same. The non-invariance is necessary if the measure satisfies the *Strict Monotonicity* axiom.
- Invariance under the classification scaling:

$$TP \Rightarrow k_1 TP \quad TN \Rightarrow k_2 TN$$
$$FP \Rightarrow k_1 FP \quad FN \Rightarrow k_2 FN$$

  where $k_1, k_2 > 0$. This invariance does not hold for any of our measure families. In fact, according to the author, this invariance is only satisfied by Precision ($P(\mathcal{G}|S)$), which is a partial measure that does not satisfy the *Strict Monotonicity* axiom.

7.1.7 Wrap up

In short, there exists in the state of the art a clear distinction between threshold measures for discrete binary outputs versus ranking evaluation measures. The Utility and Accuracy measures have been distinguished from other binary measures, but not formally. There

exists an informal category based on informativeness (probabilistic measures). And there exists also an indirect property that discriminate class-oriented measures (invariance over changes in *TN*).

The main contribution of our analysis with respect to the state of the art is to establish a framework, based on the concept of informativeness, which formally distinguishes three families of measures and clarifies the basic assumptions that define each measure set. A strength of our categorization scheme is that how measures evaluate non informative outputs determines measure disagreement as well as measure families.

7.2 Formal constraints for information access problems

Formal constraints as a tool to analyze and categorize evaluation metrics have previously been used in other information access problems: (Amigó et al. 2009) proposed four constraints for extrinsic clustering evaluation measures, which are only satisfied by the Bcubed precision & recall metric pair. Amigó et al. (2013) postulates five constraints for document retrieval evaluation measures which no metric in the state of the art satisfies, and propose a new metric pair, *Reliability* and *Sensitivity*, which comply with all constraints and can also be applied to tasks that mix retrieval, clustering and filtering aspects. Busin and Mizzaro (2013) also introduces a wide range of constraints that cover many aspects of the document retrieval problem, in an attempt to characterize document retrieval evaluation measures. Amigó et al. (2018) proposes a measure to evaluate search results diversification (Rank-Biased Utility), designed to comply with a set of formal constraints for the problem of search with diversity. The metric takes into account redudancy and user effort associated to the inspection of documents in the ranking.

Besides the analysis of evaluation measures, formal constraints have also been used to analyze and improve document retrieval models.[11] For instance, Fang et al. (2004) is a seminal work that postulates a number of constraints on tf*idf weights, which lead to a reformulation of some popular weighting schemes—such as okapi weighting—that result in better document retrieval effectiveness (Lv and Zhai 2011); proposes two constraints to lower-bound term frequency normalization (Fang and Hui 2006; Fang 2008); introduce formal constraints to model semantic term matching and query expansion (Clinchant and Gaussier 2011); propose a constraint on document frequency for pseudo-relevance feedback models; and (Karimzadehgan and Zhai 2012) introduces formal constraints to model translation estimations for document retrieval based on statistical translation models. Recently, a SIGIR workshop on the topic (*Axiomatic Thinking for Information Retrieval and Related Tasks*) (Amigo et al. 2017) has contributed to highlight the relevance of axiomatic thinking in several areas of Information Retrieval.

In general, formal constraints have proved to be a powerful analysis tool in several aspects of Information Access problems, which starts from foundational aspects rather than circumstantial empirical observations, and ultimately provide qualitative and quantitative improvements on the systems.

[11] See Fang and Zhai (2014) for an extensive discussion on the topic.

7.3 Wrap up

In summary, the main contributions of this paper with respect to the state of the art are: (1) a formal analysis and categorization of measures into families that starts from a probabilistic interpretation, which relates them with their suitability for particular user scenarios; (2) a proposal of smoothing techniques in order to keep the basic properties of metrics; (3) an empirical study of metrics based on their strictness (a good result with a strict measure ensures a good results with respect to other measures) and robustness; and (4) based on our formal and empirical results, a set of best-practice recommendations to select the most appropriate measure in a given application scenario.

8 Conclusions

The current variety of approaches to document filtering evaluation may be not only the consequence of the different nature of the various filtering tasks, but also a reflection of the lack of a systematic, analytical comparison of the properties of evaluation metrics. Our work attempts to fill this gap by presenting a comparison of measures based on formal constraints and properties.

We have relied on only one basic constraint (an axiom to be satisfied by any valid evaluation measure) that was first proposed by Sebastiani, the strict monotonicity constraint; and we have proved that not all popular measures satisfy it. We have also shown that non-compliant measures (such as Precision/Recall and Lam%) can be modified, under a probabilistic interpretation, to comply with the monotonicity constraint while preserving their properties. Our smoothing technique replaces the equiprobability assumption of Laplace's correction with a probability based on the input distribution.

Our analysis also shows that the main difference between metrics can be explained in terms of how non-informative outputs are evaluated. As a result, many evaluation measures for document filtering can be grouped in three families, each satisfying one out of three formal properties which are mutually exclusive. Utility-based measures reward good decisions in the classification process, stating an absolute weight for relevant versus irrelevant documents. Informativeness-based measures penalize good decisions which are taken by chance, considering that any non-informative output is equally useless. Finally, Class-oriented measures penalize reduced outputs (low recall), considering that the quality of non-informative outputs correlates with its size (in other words, doing nothing is better than randomly discarding information).

Finally, we have also studied the Reliability/Sensitivity metric pair, which does not fit into any of the three families, and has two distinctive empirical properties: (1) it is stricter than all other metrics in our study: a high Reliability/Sensitivity score ensures high scores with all other measures; and (2) it is more robust to changes in the set of test items than all other metrics in our study.

Our results do not prescribe any particular measure as the best option for every conceivable document filtering scenario. But, from the results of our formal analysis and our experimentation, a reasonable methodology to select and adequate measure for a particular document filtering scenario would be the following:

1. Decide how non-informative outputs should be evaluated, and select a measure in the appropriate family accordingly.

2. If such decision cannot be made (because the scenario is too general, for instance) compare results of measures from each of the families, and use the Reliability/Sensitivity metric pair as a stricter evaluation criterion.
3. If a highly unbalanced input is expected, compute measures in probabilistic terms with the non-informative smoothing mechanisms proposed in this paper, in order to avoid a biased analysis.

Acknowledgements Funding was provided by Secretaría de Estado de Investigación, Desarrollo e Innovación, Ministerio de Economía, Industria y Competitividad, Gobierno de España (Grant No. TIN2015-71785-R, project Vemodalen).

Appendix: Formal proofs

Proof Utility satisfies the *Absolute Weighting* property

The characteristic of Utility-based metrics in general, and accuracy in particular, is that they assign an absolute weight to relevant (versus non relevant) documents in the output regardless of the output size. For instance, in the case of the Utility measure U_α, being $S_{\neg i}$ and $S'_{\neg i}$ two non-informative outputs:

$$\begin{aligned} U_\alpha(S_{\neg i}) &= \alpha P(S_{\neg i}|\mathcal{G})P(\mathcal{G}) - P(S_{\neg i}|\neg\mathcal{G})P(\neg\mathcal{G}) \\ &= \alpha P(\mathcal{G}|S_{\neg i})P(S_{\neg i}) - P(\neg\mathcal{G}|S_{\neg i})P(S_{\neg i}) = P(S_{\neg i})(\alpha P(\mathcal{G}) - P(\neg\mathcal{G})) \end{aligned}$$

Therefore, if $\alpha = \frac{P(\neg\mathcal{G})}{P(\mathcal{G})}$ then the score of non-informative outputs is fixed. If $\alpha > \frac{P(\neg\mathcal{G})}{P(\mathcal{G})}$, the score of non-informative outputs grows with its size, and reversely if $\alpha < \frac{P(\neg\mathcal{G})}{P(\mathcal{G})}$. In summary, the value of the α parameter determines the relative score of two non-informative outputs. □

Proof Weighted Accuracy satisfies *Absolute Weighting*

Note that, although Accuracy can be considered a Utility-based measure, it does not directly satisfy the *Absolute Weighting* property, given that its definition does not include any parameter. However, the weighted accuracy proposed in Androutsopoulos et al. (2000) does satisfy this property, and it is a generalization of Accuracy (see proof in this section).

$$\begin{aligned} \text{Weighted Accuracy}(S_{\neg i}) &= \frac{\lambda P(S_{\neg i}|\mathcal{G})P(\mathcal{G}) + P(\neg S_{\neg i}|\neg\mathcal{G})P(\neg\mathcal{G})}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})} \\ &= \frac{\lambda P(S_{\neg i}|\mathcal{G})P(\mathcal{G}) + P(\neg S_{\neg i})P(\neg\mathcal{G})}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})} \\ &= \frac{\lambda P(S_{\neg i})P(\mathcal{G}) + (1 - P(S_{\neg i}))P(\neg\mathcal{G})}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})} \\ &= \frac{\lambda P(S_{\neg i})P(\mathcal{G}) + P(\neg\mathcal{G}) - P(S_{\neg i})P(\neg\mathcal{G})}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})} \end{aligned}$$

If we derive over $P(S_{\neg i})$ we obtain:

$$\frac{\lambda P(\mathcal{G}) - P(\neg\mathcal{G})}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})} = \frac{\lambda 2P(\mathcal{G}) - 1}{\lambda P(\mathcal{G}) + P(\neg\mathcal{G})}$$

Therefore, the score of a non-informative output grows or decreases with its size depending on whether λ is larger or smaller than $\frac{1}{2P(\mathcal{G})}$. □

Proof Lam% satisfies *Non-Informativeness Fixed Quality*

Given a non informative output $\mathcal{S}_{\neg i}$, then:

$$lam\%(\mathcal{S}_{\neg i}) = logit^{-1}\left(\frac{logit(P(\mathcal{S}_{\neg i})) + logit(P(\neg\mathcal{S}_{\neg i}))}{2}\right)$$

But given that:

$$\begin{aligned} logit(P(\mathcal{S}_{\neg i})) =& log\left(\frac{P(\mathcal{S}_{\neg i})}{1 - P(\mathcal{S}_{\neg i})}\right) = log\left(\frac{1 - P(\neg\mathcal{S}_{\neg i})}{P(\neg\mathcal{S}_{\neg i})}\right) \\ =& -log\left(\frac{P(\neg\mathcal{S}_{\neg i})}{1 - P(\neg\mathcal{S}_{\neg i})}\right) = -logit(P(\neg\mathcal{S}_{\neg i})) \end{aligned}$$

The two components in the numerator cancel out each other:

$$logit(P(\mathcal{S}_{\neg i})) + logit(P(\neg\mathcal{S}_{\neg i})) = -logit(P(\neg\mathcal{S}_{\neg i})) + logit(\neg P(\mathcal{S}_{\neg i})) = 0$$

Therefore, given any non-informative output $\mathcal{S}'$, the fixed resulting score is 0.5. □

Proof Phi satisfies *Non-Informativeness Fixed Quality*

$$Phi = \frac{TP.TN - FP.FN}{\sqrt{(TP + FN).(TN + FP).(TP + FP).(TN + FN)}}$$

Phi is always zero if $\mathcal{S}_{\neg i}$ is non informative (see proof in this section), because then the two numerator components cancel each other:

$$\begin{aligned} TP.TN =& P(\mathcal{S}_{\neg i}|\mathcal{G})P(\mathcal{G})P(\neg\mathcal{S}_{\neg i}|\neg\mathcal{G})P(\neg\mathcal{G}) = P(\mathcal{S}_{\neg i})P(\mathcal{G})P(\neg\mathcal{S}_{\neg i})P(\neg\mathcal{G}) \\ =& P(\mathcal{S}_{\neg i}|\neg\mathcal{G})P(\mathcal{G})P(\neg\mathcal{S}_{\neg i}|\mathcal{G})P(\neg\mathcal{G}) \\ =& P(\mathcal{S}_{\neg i}|\neg\mathcal{G})P(\neg\mathcal{G})P(\neg\mathcal{S}_{\neg i}|\mathcal{G})P(\mathcal{G}) = FP.FN \end{aligned}$$

And therefore *Phi* is zero. □

Proof Odds Ratio satisfies *Non-Informativeness Fixed Quality*

If $\mathcal{S}_{\neg i}$ is non informative:

$$\begin{aligned} Odds(\mathcal{S}_{\neg i}) =& \frac{TP.TN}{FN.FP} = \frac{P(\mathcal{S}_{\neg i}|\mathcal{G})P(\mathcal{G})P(\neg\mathcal{S}_{\neg i}|\neg\mathcal{G})P(\neg\mathcal{G})}{P(\mathcal{S}_{\neg i}|\neg\mathcal{G})P(\neg\mathcal{G}).P(\neg\mathcal{S}_{\neg i}|\mathcal{G})P(\mathcal{G})} \\ =& \frac{P(\mathcal{S}_{\neg i})P(\mathcal{G})P(\neg\mathcal{S}_{\neg i})P(\neg\mathcal{G})}{P(\mathcal{S}_{\neg i})P(\neg\mathcal{G})P(\neg\mathcal{S}_{\neg i})P(\mathcal{G})} = 1 \end{aligned}$$

□

Proof Macro Average Accuracy satisfies *Non-Informativeness Fixed Quality*

$$MAAc(S_{\neg i}) = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} = \frac{P(S|\mathcal{G}) + P(\neg S|\neg\mathcal{G})}{2}$$

If $S_{\neg i}$ is non-informative then:

$$\begin{aligned} MAAc(S_{\neg i}) =& \frac{P(S_{\neg i}|\mathcal{G}) + P(\neg S_{\neg i}|\neg\mathcal{G})}{2} = \frac{P(S_{\neg i}) + P(\neg S_{\neg i})}{2} \\ =& \frac{P(S_{\neg i}) + 1 - P(S_{\neg i})}{2} = \frac{1}{2} \end{aligned}$$

□

Proof Kappa statistic satisfies *Non-Informativeness Fixed Quality*

The Kappa statistic is defined as:

$$KapS(S) = \frac{\text{Accuracy} - \text{Random Accuracy}}{1 - \text{Random Accuracy}}$$

where *Random Accuracy* represents the Accuracy obtained randomly by an output with size $|S|$. In our probabilistic notation, Kappa can be expressed as:

$$KapS(S) = \frac{(P(S|\mathcal{G})P(\mathcal{G}) + P(\neg S|\neg\mathcal{G})P(\neg\mathcal{G})) - (P(S)P(\mathcal{G}) + P(\neg S)P(\neg\mathcal{G}))}{1 - (P(S)P(\mathcal{G}) + P(\neg S)P(\neg\mathcal{G}))}$$

If $S_{\neg i}$ is non informative then $P(S_{\neg i}|\mathcal{G}) = P(S_{\neg i})$, and the formula returns zero. □

Proof Chi-square satisfies *Non-Informativeness Fixed Quality*

$$\begin{aligned} Chi(S) =& \frac{(|S \cap \mathcal{G}|.|\neg S \cap \neg\mathcal{G}| - |S \cap \neg\mathcal{G}|.|\neg S \cap \mathcal{G}|) + |T|}{|S| + |\mathcal{G}| + |\neg S| + |\neg\mathcal{G}|} \\ =& \frac{(P(S|\mathcal{G}).P(\neg S|\neg\mathcal{G}) - P(S|\neg\mathcal{G}).P(\neg S|\mathcal{G})) + 1}{2} \end{aligned}$$

If an output $S_{\neg i}$ is non informative then:

$$Chi(S_{\neg i}) = \frac{(P(S_{\neg i})P(\neg S_{\neg i}) - P(S_{\neg i})P(\neg S_{\neg i}) + 1}{2} = \frac{1}{2}$$

□

Proof The F measure of Precision and Recall for the positive class satisfies *non-informativeness growing quality*

The F measure for a non-informative output grows with its size (i.e. with the ratio of items labeled as positive by the system), because

$$F_\alpha(S_{\neg i}) = F_\alpha(P(\mathcal{G}|S_{\neg i}), P(S_{\neg i}|\mathcal{G})) = F_\alpha(P(\mathcal{G}), P(S_{\neg i}))$$

The F measure *Independence* property (Van Rijsbergen 1974) states that, if the first parameter is fixed (in our case, $P(\mathcal{G})$), F grows with the second parameter (in our case, $P(\mathcal{S}_{\neg i})$, which is the probability that an item receives a posivitive label). Therefore,

$$F_\alpha(P(\mathcal{G}), P(\mathcal{S}_{\neg i})) \sim P(\mathcal{S}_{\neg i})$$

which satisfies the non-informativeness growing quality. □

Proof Every non informative output receives an F(R,S) score lower than 0.25.

Given a non informative input $\mathcal{S}_{\neg i}$, $\mathrm{F}_\alpha(\mathrm{R}(\mathcal{S}_{\neg i}), \mathrm{S}(\mathcal{S}_{\neg i}))$ can be expressed as:

$$\begin{aligned}
\mathrm{F}_\alpha(\mathrm{R}(\mathcal{S}_{\neg i}),\ \mathrm{S}(\mathcal{S}_{\neg i})) &= \left(\frac{\alpha}{P(\mathcal{G}|\mathcal{S}_{\neg i})P(\neg\mathcal{G}|\neg\mathcal{S}_{\neg i})} + \frac{1-\alpha}{P(\mathcal{S}_{\neg i}|\mathcal{G})P(\neg\mathcal{S}_{\neg i}|\neg\mathcal{G})}\right)^{-1} \\
&= \left(\frac{\alpha}{P(\mathcal{G})P(\neg\mathcal{G})} + \frac{1-\alpha}{P(\mathcal{S}_{\neg i})P(\neg\mathcal{S}_{\neg i})}\right)^{-1} \\
&= \left(\frac{\alpha}{P(\mathcal{G})(1-P(\mathcal{G}))} + \frac{1-\alpha}{P(\mathcal{S}_{\neg i})(1-P(\mathcal{S}_{\neg i}))}\right)^{-1}
\end{aligned}$$

We can prove easily that if $0 \leq x \leq 1$, then the function $f = x(1-x)$ is upper bounded by 0.25.[12] Therefore, according to the harmonic mean properties, the maximal value of F(R,S) is:

$$\mathrm{F}_\alpha(\mathrm{R}(\mathcal{S}_{\neg i}),\ \mathrm{S}(\mathcal{S}_{\neg i})) \leq \left(\frac{\alpha}{0.25} + \frac{1-\alpha}{0.25}\right)^{-1} = \left(\frac{\alpha+1-\alpha}{0.25}\right)^{-1} = 0.25$$

□

References

Agresti, A., & Hitchcock, D. B. (2005). *Bayesian inference for categorical data analysis: A survey*. Technical report.

Amigó, E., Artiles, J., Gonzalo, J., Spina, D., Liu, B., & Corujo, A. (2010). WePS3 evaluation campaign: Overview of the on-line reputation management task. In *2nd Web people search evaluation workshop (WePS 2010), CLEF 2010 conference, Padova Italy*.

Amigó, E., Corujo, A., Gonzalo, J., Meij, E., & de Rijke, M. (2012). Overview of RepLab 2012: Evaluating online reputation management systems. In *CLEF (online working notes/labs/workshop)*.

Amigo, E., Fang, H., Mizzaro, S., & Zhai, C. (2017). Axiomatic thinking for information retrieval and related tasks. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, SIGIR '17, pp. 1419–1420, New York, 2017. ACM.

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*(4), 461–486.

Amigó, E., Gonzalo, J, & Verdejo, F. (2013). A generic measure for document organization tasks. In *Proceedings of ACM SIGIR*, pp. 643–652. ACM Press.

[12] We omit the proof; it is enough to solve the equation f'(x)=0.

Amigó, E., Spina, D., & Carrillo-de-Albornoz, J. (2018). An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *CoRR*, abs/1805.02334.

Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive bayesian anti-spam filtering. In *CoRR*, cs.CL/0006013.

Bradley, Andrew P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Busin, L., & Mizzaro, S. (2013). Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 conference on the theory of information retrieval*, ICTIR '13, pp. 8:22–8:29, New York, NY, 2013. ACM.

Callan, J. (1996). Document filtering with inference networks. In *Proceedings of the nineteenth annual international ACM SIGIR conference on research and development in information retrieval*, pp. 262–269.

Caruana, R., & Niculescu-Mizil, A. (2005). An empirical comparison of supervised learning algorithms using different performance metrics. In *Proceedings of 23rd international conference machine learning (ICML06)*, pp. 161–168.

Clinchant, S., & Gaussier, E. (2011). Is document frequency important for PRF? In *Advances in information retrieval theory*, pp. 89–100. Springer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37.

Cormack, G., & Lynam, T. (2005). TREC 2005 spam track overview. In *Proceedings of the fourteenth text retrieval conference 8TREC 2005)*.

Fang, H. (2008). A re-examination of query expansion using lexical resources. In *ACL*, vol. 2008, pp. 139–147. Citeseer.

Fang, H., Tao, T., & Zhai, C. X. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49–56. ACM.

Fang, H., & Zhai, C. X. (2006). Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 115–122. ACM.

Fang, H, & Zhai, C. X. (2014). Axiomatic analysis and optimization of information retrieval models. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval*, SIGIR '14, pp. 1288–1288, New York, NY, 2014. ACM.

Fawcett, T., & Niculescu-Mizil, A. (2007). PAV and the ROC convex hull. *Machine Learning*, *68*, 97–106.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society Series B (Methodological)*, *14*, 107–114.

Hedin, B., Tomlinson, S., Baron, J. R., & Oard, D. W. (2009). *Overview of the TREC 2009 legal track*.

Hoashi, K., Matsumoto, K., Inoue, N., & Hashimoto, K. (2000). Document filtering method using nonrelevant information profile. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, SIGIR '00, pp. 176–183, New York, NY, 2000. ACM.

Hull, David A. (1997). The TREC-6 filtering track: Description and analysis. *Proceedings of the TREC*, *6*, 33–56.

Hull, D. A. (1998). The TREC-7 filtering track: Description and analysis. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of TREC-7, 7th text retrieval conference*, pp. 33–56, Gaithersburg, US, 1998. National Institute of Standards and Technology, Gaithersburg, US.

Karimzadehgan, M., & Zhai, C. X. (2012). Axiomatic analysis of translation language model for information retrieval. In *Advances in information retrieval*, pp. 268–280. Springer, Berlin.

Karon, B. P., & Alexander, I. E. (1958). Association and estimation in contingency tables. *Journal of the American Statistical Association*, *23*(2), 1–28.

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on online social networks*, pp. 19–24, New York, NY, 2008. ACM.

Le, A., Ajot, J., Przybocki, M., & Strassel, S. (2010). Document image collection using Amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*, pp. 45–52, Los Angeles, June 2010. Association for Computational Linguistics.

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A statistically consistent and more discriminating measure than accuracy. In *IJCAI*, pp. 519–526.

Lv, Y., & Zhai, C. X. (2011). Lower-bounding term frequency normalization. In *Proceedings of the 20th ACM international conference on information and knowledge management*, CIKM '11, pp. 7–16, New York, NY, 2011. ACM.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.

Persin, Michael. (1994). Document filtering for fast ranking. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pp. 339–348, New York, NY, 1994. Springer, New York.

Provost, F. J., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Knowledge discovery and data mining*, pp. 43–48.

Qi, Haoliang, Yang, Muyun, He, Xiaoning, & Li, Sheng. (2010). Re-examination on lam% in spam filtering. In *Proceedings of the SIGIR 2010 conference, Geneva, Switzerland*.

Robertson, S., & Hull, D. A. (2001). The TREC-9 filtering track final report. In *Proceedings of TREC-9*, pp. 25–40.

Schapire, R. E., Singer, Y., & Singhal, A. (1998). Boosting and Rocchio applied to text filtering. In *Proceedings of ACM SIGIR*, pp. 215–223. ACM Press.

Sebastiani, F. (2015). An axiomatically derived measure for the evaluation of classification algorithms. In *ICTIR*, pp. 11–20.

Sokolova, M. (2006). Assessing invariance properties of evaluation measures. In *Proceedings of NIPS'06 workshop on testing deployable learning and decision systems*.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. *AI 2006: Advances in artificial intelligence*, pp. 1015–1021.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Van Rijsbergen, C. (1974). Foundation of evaluation. *Journal of Documentation*, *30*(4), 365–373.