A Proposal for New Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks*

Francisco José Valverde-Albacete[1],
Jorge Carrillo-de-Albornoz[1], and Carmen Peláez-Moreno[2]

[1] Departamento de Lenguajes y Sistemas Informáticos
Univ. Nacional de Educación a Distancia, c/ Juan del Rosal, 16. 28040 Madrid, Spain
{fva,jcalbornoz}@lsi.uned.es
[2] Departamento de Teoría de la Señal y de las Comunicaciones
Universidad Carlos III de Madrid, 28911 Leganés, Spain
carmen@tsc.uc3m.es

Abstract. In this paper we propound the use of a number of entropy-based metrics and a visualization tool for the intrinsic evaluation of Sentiment and Reputation Analysis tasks. We provide a theoretical justification for their use and discuss how they complement other accuracy-based metrics. We apply the proposed techniques to the analysis of TASS-SEPLN and RepLab 2012 results and show how the metric is effective for system comparison purposes, for system development and postmortem evaluation.

1 Introduction

The appropriate evaluation of multi-class classification is a founding stone of Machine Learning. For Sentiment and Reputation Analysis (SA and RA), where different polarities—for instance *positive, neutral, negative*—and several degrees of such polarities may be of interest, it is a crucial tool.

However, accuracy-based methods in predictive analytics suffer from the well-known accuracy paradox, viz. a high level of accuracy is not a necessarily an indicator of high classifier performance [1, 2, 3]. In other words, a high accuracy figure does not necessarily imply that the classifier has been able to model the underlying phenomena.

Since accuracy-improving methods try to improve the *heuristic rule* of minimizing the number of errors, we have to question whether rather than a shortcoming of accuracy, this paradox might be *a shortcoming of the heuristic.*

An alternative heuristic is to maximize the information transferred from input to output through the classification process, as described by the contingency matrix. In [4] an information-theoretic visualization scheme was proposed,

* FJVA and JCdA are supported by EU FP7 project LiMoSINe (contract 288024). CPM has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2011-26807 for this paper.

the *entropy triangle*, where the *mutual information (MI)* of the contingency matrix is related to the distance of the input and output distributions from uniformity and to the *variation of information* [5], another distance measuring how much information from input was not learnt and how much information at the output is not predicted by the classifier.

Unfortunately, MI is expressed in bits, not in efficiency, and this detracts from its intended reading as a metric. Furthermore, it is actually one aspect of a tripolar manifestation [4], hence not adequate as a *binary* indicator of goodness. Also, it measures how well has the classifier learnt the input distribution, but not what its expected accuracy is.

On the other hand, the Normalized Information Transfer (NIT) factor [6] is a measure that relates to MI in the same way that the reduction in perplexity of a language model relates to the entropy of a source: it quantifies how well the classifier has done its job of reducing the uncertainty in the input distribution. This reading allows us to justify an Entropy-Modulated Accuracy that can be used as a complement to more standard, error-based metrics, like precision, recall or F-score.

In the following we introduce more formally these two tools (Section 2) and apply them to the systems that took part in the last TASS-SEPLN and RepLab 2012 campaigns (Section 3). We conclude with some suggestions for their use.

2 The Entropy Triangle and the Normalized Information Transfer

2.1 The Entropy Triangle: A Visualization Tool

The entropy triangle is a contingency matrix visualization tool based on an often overlooked decomposition of the joint entropy of two random variables[4]. Figure 1 shows such a decomposition showing the three crucial regions:

- The *mutual information*,

$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}}$$

- The *variation of information*, the addition of the conditional perplexities on input and output [5],

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} \tag{1}$$

- And the *entropy decrement* between the uniform distributions with the same cardinality of events as P_X and P_Y and the entropy of the joint distribution where both are independent,

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y} \quad . \tag{2}$$

Note that all of these quantities are positive. In fact from the previous decomposition the following *balance equation* is evident,

$$\begin{gathered} H_{U_X \cdot U_Y} = \Delta H_{P_X \cdot P_Y} + 2 * MI_{P_{XY}} + VI_{P_{XY}} \qquad (3) \\ 0 \leq \Delta H_{P_X \cdot P_Y}, MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_X \cdot U_Y} \end{gathered}$$

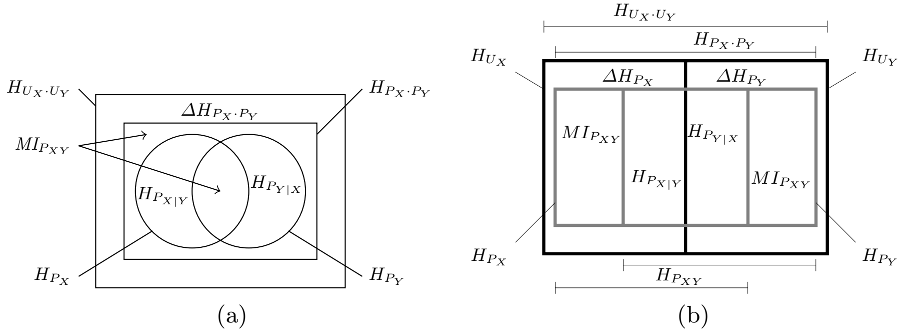


Fig. 1. Extended entropy diagrams related to a bivariate distribution, from [4]. The bounding rectangle is the joint entropy of two uniform (hence independent) distributions U_X and U_Y of the same cardinality as input probability distribution P_X and output P_Y, resp. The expected mutual information $MI_{P_{XY}}$ appears *twice* in (a) and this makes the diagram split for each variable symmetrically in (b).

where the bounds are easily obtained from distributional considerations. If we normalize (3) by the overall entropy $H_{U_X \cdot U_Y}$ we obtain the equation of the 2-simplex in entropic space,

$$\begin{aligned} 1 &= \Delta' H_{P_X \cdot P_Y} + 2 * MI'_{P_{XY}} + VI'_{P_{XY}} \\ 0 &\leq \Delta' H_{P_X \cdot P_Y}, MI'_{P_{XY}}, VI'_{P_{XY}} \leq 1 \end{aligned} \tag{4}$$

representable by a De Finetti or ternary entropy diagram or simply *entropy triangle (ET).*

The evaluation of classifiers is fairly simple using the schematic in Fig. 2.

1. Classifiers on the bottom side of the triangle *transmit no mutual information* from input to output: they have not profited by being exposed to the data.
2. Classifiers on the right hand side have diagonal confusion matrices, hence *perfect (standard) accuracy.*
3. Classifiers on the left hand side operate on perfectly balanced data distributions, hence they are *solving the most difficult multiclass problem* (from the point of view of an uninformed decision).

Of course, combinations of these conditions provide specific kinds of classifiers. Those at the apex or close to it are obtaining the highest accuracy possible on very balanced datasets and transmitting a lot of mutual information hence they are the *best classifiers* possible. Those at or close to the left vertex are essentially not doing any job on very difficult data: they are *the worst classifiers.* Those at or close to the right vertex are not doing any job on very easy data for which they claim to have very high accuracy: they are *specialized (majority) classifier*s and our intuition is that they are the kind of classifiers that generate the accuracy paradox [1].

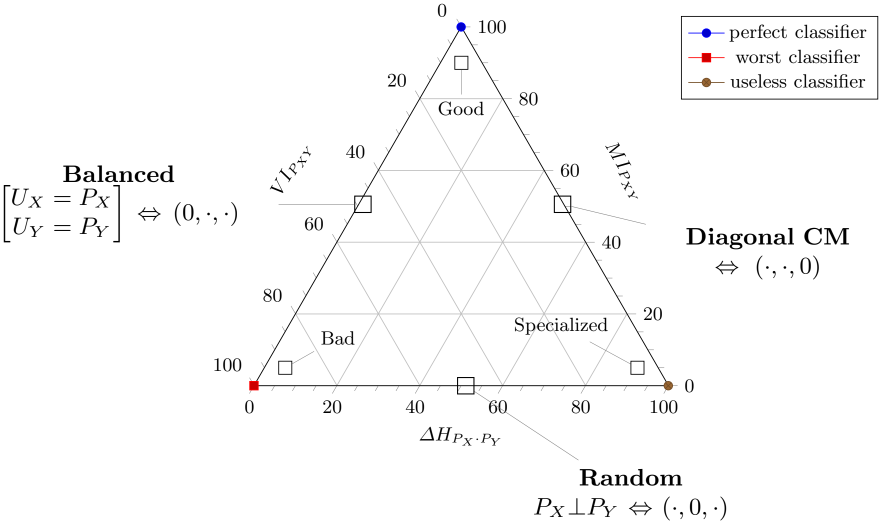

perfect classifier
worst classifier
useless classifier
Good
Bad
Specialized
$VI_{P_{XY}}$
$MI_{P_{XY}}$
$\Delta H_{P_X \cdot P_Y}$
0
20
40
60
80
100
Balanced
$\begin{bmatrix} U_X = P_X \\ U_Y = P_Y \end{bmatrix} \Leftrightarrow (0,\cdot,\cdot)$
Diagonal CM
$\Leftrightarrow (\cdot,\cdot,0)$
Random
$P_X \perp P_Y \Leftrightarrow (\cdot,0,\cdot)$

Fig. 2. Schematic Entropy Triangle showing interpretable zones and extreme cases of classifiers. The annotations on the center of each side are meant to hold for that whole side.

In just this guise, the ET has already been successfully used in the evaluation of Speech Recognition systems [4, 7]. But a simple extension of the ET is to endow it with a graduated axis or colormap that also allows us to visualize the correlation of such information-theoretic measures with other measures like accuracy, greatly enhancing its usefulness. Examples of its use can be seen in Figs. 3 and 4, and this is the main tool we propose to complement other Sentiment Analysis metrics.

2.2 The Normalized Information Transfer (NIT) Factor and the Entropy-Modified Accuracy (EMA)

The problem with the ET is that in spite of being helpful as a visualization and exploration tool, it does not allow for system ranking at the heart of modern competitive, task-based evaluation. For such purposes we use a corrected version of the accuracy and a measure derived from mutual information.

A measure of the *effectiveness of the learning process* is the *information transfer factor* $\mu_{XY} = 2^{MI_{P_{XY}}}$ but we prefer to report it as a fraction of the number of classes, the *Normalized Information Transfer factor (NIT)*,

$$q(P_{XY}) = \frac{\mu_{XY}}{k} = 2^{MI_{P_{XY}} - H_{U_X}} \tag{5}$$

The NIT is explained in the context of the perplexity of the classifier [6]. The quantity $\mu_X = 2^{MI_{XY}}$ is interpreted there as the reduction in the number of classes afforded by a classifier on average, as seen from the point of view of an uninformed decision: the higher this reduction, the better. In the worst case—random decision—, this reduction is $MI_{P_{XY}} = 0, 2^{MI_{P_{XY}}} = 1$ whence the NIT is $1/k$. In the best possible case (perfect classifier, balanced class distribution) this reduction is $MI_{P_{XY}} = log_2 k, 2^{MI_{P_{XY}}} = k$, whence the normalized rate is 1 so that the range of the NIT factor is $1/k \leq q((P_{XY}) \leq 1$ matching well the intuition that a random decision on a balanced data set can only guess right $1/k$ of the times on average but the best informed decision guesses right always.

Considering the two paragraphs above, $k_{X|Y} = 2^{H_{P_{X|Y}}}$ can be interpreted as the *remanent number of equiprobable classes* seen by the classifier (after learning the task). But $k_{X|Y}$ is precisely the number of equiprobable classes the classifier sees after subtracting the NIT, whence the *entropy-modulated accuracy (EMA)* of the classifier would be

$$a'(P_{XY}) = 1/k_{X|Y} = 2^{-H_{P_{X|Y}}}$$

We can see that the EMA is corrected by the input distribution and the learning process, i.e. the more efficient the learning process, the higher the NIT and the higher the EMA but, the more imbalanced the input class distribution, the lower k_X and the higher the EMA.

Note that this last commentary makes the EMA a suspicious metric: classifiers should only be compared when the effective perplexities of the tasks they are applied to are comparable, that is, with similar k_X . For classifiers across tasks, then, the NIT is a better measure of success, although when measuring performance *on the same task*, modified accuracy is a good metric. In the following, we will report both.

3 Experiments and Evaluation

3.1 Sentiment Analysis in TASS-SEPLN

The aim of the TASS-SEPLN competition was to classify tweets into different degrees of *Sentiment* polarity. The data consists of tweets, written in Spanish by nearly 200 well-known personalities and celebrities of the world [8]. Each tweet is tagged with its global polarity, indicating whether the text expresses a positive, negative or neutral sentiment, or no sentiment at all. Five levels have been defined: *strong positive* (P+), *positive* (P), *neutral* (NEU), *negative* (N), *strong negative* (N+) and one additional *no sentiment* tag (NONE). Table 1 shows the distribution of these classes in the training and test sets, and their effective perplexities: the training sets are much more balanced.

In TASS-SEPLN, polarity classification is evaluated as two different tasks. The goal of TASS5 is to automatically classify each of the tweets into one of the 5 polarity levels mentioned above. However, prior to this classification, the task requires to filter out those tweets not expressing any sentiment (i.e., those tagged as NONE), so the number of classes is $k = 6$. TASS3 consists in classifying each tweet in 3 polarity classes (*positive*, *neutral* and *negative*). To this end, tweets

Table 1. Distribution of tweets per polarity class in the TASS corpus

TASS5	P+	P	NEU	N	N+	NONE	TOTAL	k_X
training	1 764	1 019	610	1 221	903	1 702	7 219	5.6
testing	20 745	1 488	1 305	11 287	4 557	21 416	60 798	4.1
TASS3								
training		2 783	610	2 124		1 702	7 219	3.6
testing		22 233	1 305	15 844		21 416	60 798	3.2

tagged as positive and strong positive are merged into a single category (*positive*), and tweets tagged as negative and strong negative into another (*negative*). This task is called TASS3 but has $k = 4$.

Table 2 shows the numeric results of the different metrics on the (a) TASS3 and (b) TASS5 tasks. These data reveal that the EMA is much lower than normal accuracy and that there would be some reordering of the ranking if EMA was the ranking criterion. In particular, some sets of submissions are systematically pushed downwards in the table according to EMA. These phenomenon warrants some postmortem analysis of the results of such systems.

Furthermore, some systems, specifically those with $\mu_{XY} \approx 1.000$, essentially took random decisions but their accuracies were well above random. This is a strong result that shows the inadequacy of accuracy for such evaluations.

Figure 3 presents the ET visualization of the performance of the different systems at either task, revealing some interesting results. First, in both tasks four systems are closer to the upper vertex of the triangle implying a better behaviour than the others. However, their distance to the apex of the ET indicates that even these systems are still far from solving the task, that is, being able to model the different polarities captured in the data, even though the best accuracy is 72.3% in TASS3, 67.8% in TASS5. This is another strong hint that *high accuracy does not correlate with high performance in the task.* Furthermore, the triangles show that two systems (correlative submissions in either tasks) are placed very close to the base of triangle, which suggests both random decision and specialization as majority classifiers, despite their achieving an accuracy of around 35% in both tasks. These are the very same systems with $\mu_{XY} \approx 1.000$.

Second, while the accuracy of the systems is better in TASS3 than in TASS5 (as expected, since the complexity of the problem increases with the number of classes), the evaluation according to the ET shows that the behaviour of the systems is, in practice, the same in both tasks. In our opinion, the explanation can be found in the evaluation methodology and distribution of classes in the dataset: for TASS3, *positive* and *strong positive* tweets are merged in a single category, and *negative* and *strong negative* tweets are merged in another category. But since the number of tweets in the *positive* and *strong negative* categories is very low in comparison with the number of tweets in the remaining categories, the effect of misclassifying tweets of these two categories in TASS5 is not that marked, in terms of accuracy.

Table 2. Perplexities, accuracy (a), EMA (a'_X) and NIT factor (q_X) for the TASS test runs. . The ranking by accuracy (official) and by EMA have some inversions (red=should sink, green=should rise).

(a) TASS3: $k = 4, k_X = 3.2$

TASS3 run	$k_{X\mid Y}$	μ_{XY}	a	a'_X	q_X
daedalus-1	2.090	1.539	0.723	0.478	0.385
elhuyar-1	2.265	1.420	0.711	0.441	0.355
l2f-1	2.258	1.424	0.691	0.443	0.356
l2f-3	2.256	1.426	0.690	0.443	0.356
l2f-2	2.312	1.391	0.676	0.432	0.348
atrilla-1	2.541	1.266	0.620	0.394	0.316
sinai-4	2.706	1.189	0.606	0.370	0.297
uned1-1	2.735	1.176	0.590	0.366	0.294
uned1-2	2.766	1.163	0.588	0.362	0.291
uned2-1	2.819	1.141	0.501	0.355	0.285
imdea-1	2.953	1.089	0.459	0.339	0.272
uned2-2	3.033	1.061	0.436	0.330	0.265
uned2-4	2.900	1.109	0.412	0.345	0.277
uned2-3	3.070	1.048	0.404	0.326	0.262
uma-1	2.649	1.214	0.376	0.377	0.304
sinai-2	3.212	1.001	0.358	0.311	0.250
sinai-1	3.213	1.001	0.356	0.311	0.250
sinai-3	3.216	1.000	0.351	0.311	0.250

(b) TASS5: $k = 6, k_X = 3.2$

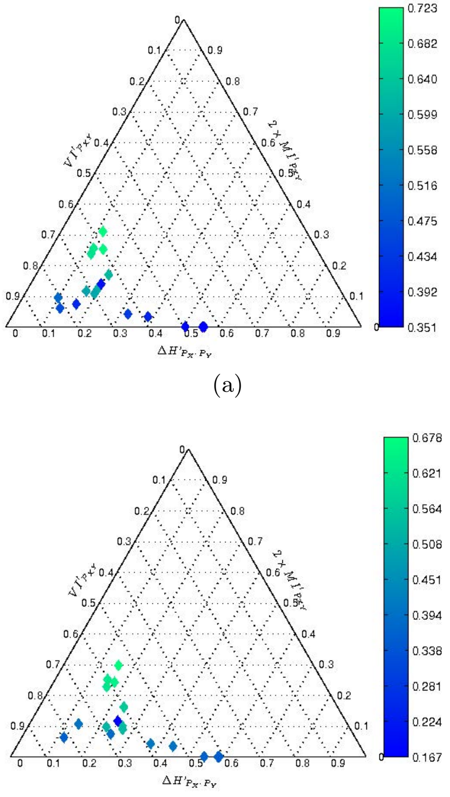
TASS5 run	$k_{X\mid Y}$	μ_{XY}	a	a'_X	q_X
daedalus-1	2.413	1.705	0.678	0.414	0.284
elhuyar-1	2.664	1.545	0.653	0.375	0.257
l2f-1	2.625	1.567	0.634	0.381	0.261
l2f-3	2.620	1.570	0.633	0.382	0.262
l2f-2	2.734	1.505	0.622	0.366	0.251
atrilla-1	3.077	1.337	0.570	0.325	0.223
sinai-4	3.432	1.199	0.547	0.291	0.200
uned1-2	3.505	1.174	0.538	0.285	0.196
uned1-1	3.454	1.191	0.525	0.290	0.199
uned2-2	3.809	1.080	0.404	0.263	0.180
uned2-1	3.395	1.212	0.400	0.295	0.202
uned2-3	3.865	1.064	0.395	0.259	0.177
uned2-4	3.600	1.143	0.386	0.278	0.190
imdea-1	3.674	1.121	0.360	0.272	0.187
sinai-2	4.107	1.002	0.356	0.243	0.167
sinai-1	4.110	1.001	0.353	0.243	0.167
sinai-3	4.113	1.000	0.350	0.243	0.167
uma-1	3.338	1.232	0.167	0.300	0.205

3.2 Reputation Analysis in RepLab 2012

RepLab 2012 was an evaluation campaign aimed at comparing classification systems trained to determine whether a tweet content has positive, negative or neutral implications for corporate reputation [9]. This task is related to sentiment analysis and opinion mining, but differs in some important points: not only opinions or subjective content are being analysed, but also polar facts, i.e. objective information that might have negative or positive implications for a company's reputation. For instance, "Barclays plans additional job cuts in the next two years" is a fact with negative implications for reputation. Since more than 1 out of 3 tweets are polar facts affecting reputation without containing sentiments or emotions, the number of cases that cannot be correctly captured using sentiment analysis techniques alone is very significant.

Moreover, the focus is set on the decisive role that the point of view or perspective can play since, for example, the same information may be negative from the point of view of the clients and positive from the point of view of investors. For instance, "R.I.P. Michael Jackson. We'll miss you" has a negative associated sentiment for fans, but a positive implication for the reputation of Michael Jackson.

The data set was manually labelled by experts for 6 companies (for training) and 31 companies (for testing) both in English and Spanish. The distribution of tweets among classes is summarized in Table 3.



(a)

(b)

Fig. 3. Entropy triangles for the TASS Sentiment Analysis tasks for 3 (a) and 5 (b) polarity degrees. Colormap correlates with accuracy.

Figure 4 shows the performance of the different systems submitted to the RepLab 2012 evaluation on the Entropy Triangle, whose analysis seems to indicate that classifying reputation polarity is a more complex task than classifying sentiment polarity, since the results in the RepLab 2012 show that most systems present a nearly random behaviour (obtaining very bad performances in the more balanced test distribution). This is further supported on lower accuracies *and* EMAs.

Only one system (the one above the others) presents results that suggest that, even reporting a low performance, is differentiating correctly between classes. Notoriously, this system is knowledge-supervised, while most of the rest approaches are based in machine learning statistical supervised approaches.

In contrast, the system to the middle of the bottom side of the triangle is specialized returning to every input the label of the majority class. This deduction from the theoretical side was corroborated by its authors declaring that this last system classifies all instances as positive [10], the majority class *in training*. This

Table 3. Distribution of tweets per polarity class in the RepLab 2012 corpus. Effective perplexities are very different for training and testing.

Dataset	P	NEU	N	TOTAL	k_X
training	885	550	81	1 516	2.32
testing	1 625	1 488	1 241	4 354	2.98

was a profitable strategy in terms of accuracy according to the training set (see Table 3) but certainly not in the test set where the classes are not that skewed (hence accuracies in the 30%). This extreme behaviour is perfectly identified in the ET and with the NIT factor and it would have been detected irrespective of the test set distribution. In fact, this system is the last in the ranking according to both EMA and NIT whilst holding the 24th position out of 35, according to accuracy. Since many of the systems of the competition were based on statistical modelling, similar behaviours can be observed due to the marked imbalance of the training set classes.

An example of this is the system presented to both evaluations (RepLab 2012 [11] and TASS-SEPLN [12]). This system, based on sentiment analysis techniques [13], achieved a reasonably good performance in TASS3, but was considerably worse in the RepLab 2012. This behaviour seems to corroborate our hypothesis that polarity for reputation and sentiment analysis are substantially different tasks. Finally, it is also worth mentioning that both tasks should take into consideration the presence of irony. Few works have dealt with the effect of irony when analyzing polarity [14, 15], but its correct analysis should increase the performance of SA and RA approaches. Our intuition is that this phenomenon

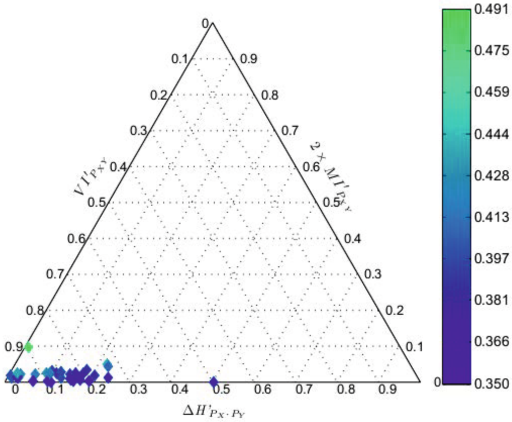


Fig. 4. Entropy triangles for the whole population of systems presented to the RepLab2012 Reputation Analysis. The colormap encodes accuracy. The task is not solved, even as a collective effort, taking the NIT as the criterion.

Table 4. Relevant perplexities, accuracy $a(P_{XY})$, EMA $a'(P_{XY})$ and NIT factor $q_X(P_{XY})$ for RepLab 2012 confusion matrices. k_X is not homogeneous due to the possibility of submitting only part of the results.

RepLab 2012	k_X	$k_{X\|Y}$	μ_{XY}	a	a'_X	q_X
polarity-Daedalus-1	2.982	2.678	1.113	0.491	0.373	0.371
polarity-HJHL-4	2.775	2.629	1.056	0.439	0.380	0.352
profiling-uned-5	2.982	2.897	1.029	0.436	0.345	0.343
profiling-BMedia-4	2.982	2.899	1.029	0.427	0.345	0.343
profiling-BMedia-5	2.982	2.911	1.024	0.420	0.343	0.341
profiling-uned-2	2.982	2.902	1.027	0.418	0.345	0.342
profiling-uned-4	2.982	2.902	1.027	0.418	0.345	0.342
profiling-BMedia-2	2.982	2.911	1.024	0.415	0.344	0.341
profiling-OPTAH-2.tx	2.981	2.841	1.049	0.408	0.352	0.350
profiling-BMedia-3	2.982	2.924	1.020	0.398	0.342	0.340
profiling-BMedia-1	2.982	2.941	1.014	0.398	0.340	0.338
profiling-OXY-2	2.982	2.938	1.015	0.396	0.340	0.338
profiling-uned-1	2.982	2.892	1.031	0.396	0.346	0.344
profiling-uned-3	2.982	2.892	1.031	0.396	0.346	0.344
profiling-OXY-1	2.982	2.939	1.015	0.394	0.340	0.338
polarity-HJHL-1	2.775	2.685	1.034	0.391	0.372	0.345
profiling-ilps-4	2.982	2.962	1.007	0.391	0.338	0.336
profiling-ilps-3	2.982	2.914	1.023	0.385	0.343	0.341
profiling-ilps-1	2.982	2.962	1.007	0.384	0.338	0.336
profiling-kthgavagai	2.982	2.922	1.020	0.383	0.342	0.340
profiling-ilps-5	2.982	2.876	1.037	0.382	0.348	0.346
profiling-OPTAH-1.tx	2.981	2.904	1.026	0.380	0.344	0.342
polarity-HJHL-3	2.775	2.695	1.030	0.377	0.371	0.343
profiling-GATE-1	2.982	2.982	1.000	0.373	0.335	0.333
profiling-OXY-4	2.982	2.947	1.012	0.369	0.339	0.337
profiling-ilps-2	2.982	2.960	1.008	0.369	0.338	0.336
polarity-HJHL-2	2.775	2.697	1.029	0.369	0.371	0.343
profiling-uiowa-2	2.982	2.937	1.015	0.367	0.340	0.338
profiling-uiowa-5	2.982	2.940	1.014	0.367	0.340	0.338
profiling-OXY-5	2.982	2.967	1.005	0.365	0.337	0.335
profiling-uiowa-1	2.980	2.933	1.016	0.362	0.341	0.339
profiling-uiowa-4	2.982	2.974	1.003	0.360	0.336	0.334
profiling-GATE-2	2.982	2.971	1.004	0.357	0.337	0.335
profiling-uiowa-3	2.980	2.975	1.001	0.355	0.336	0.334
profiling-OXY-3	2.982	2.967	1.005	0.350	0.337	0.335

is more common in RA texts and can explain, to some extent, the remarkable differences in the results.

Table 4 shows the numeric results of the various metrics being compared. The interesting note here is that another system would actually have won the competition if the metric was EMA, specifically "polarity-HJHL-4". This is one of set of systems marked in green whose EMA is comparable to that which won the competition.

4 Conclusions: A Proposal

We have motivated and proposed a combination of two tools as an alternative or a complement to standard accuracy-based metrics for Sentiment Analytics tasks, testing them on two different evaluation runs of Sentiment Analysis (TASS-SEPLN) and Reputation Analysis (RepLab 2012).

On the one hand, EMA is a better motivated, although pessimistic, estimate of accuracy that takes into consideration the dataset being considered and how much a particular system has learnt in the training process. This is to be used for ranking purposes.

On the other hand, the NIT factor is a measure of how efficient the training process of the classifier was, that can be visualized directly with the help of the Entropy Triangle. This is intended as a mechanism for technology development under the heuristic of maximizing the information transmitted in the learning process. It is well-matched to EMA in the sense that maximizing the former maximizes the latter.

We have shown that using both in combination in postmortem system analysis detects incongruencies and shortcomings of rankings based in accuracy.

As future lines of work a more in depth analysis of the learning process can be pursued by interpreting the split entropy diagram of Fig. 1.

The MATLAB[1] code to draw the entropy triangles in Figs. 3 and 4 has been made available at: http://www.mathworks.com/matlabcentral/fileexchange/30914

Acknowledgments. We would like to thank the organizers of the TASS-SEPLN and RepLab12 evaluations for providing us with the evaluation data.

References

[1] Zhu, X., Davidson, I.: Knowledge discovery and data mining: challenges and realities. Premier reference source. Information Science Reference (2007)

[2] Thomas, C., Balakrishnan, N.: Improvement in minority attack detection with skewness in network traffic. In: Proc. of SPIE, vol. 6973, pp. 69730N–69730N–12 (2008)

[3] Fernandes, J.A., Irigoien, X., Goikoetxea, N., Lozano, J.A., Inza, I., Pérez, A., Bode, A.: Fish recruitment prediction, using robust supervised classification methods. Ecological Modelling 221, 338–352 (2010)

[4] Valverde-Albacete, F.J., Peláez-Moreno, C.: Two information-theoretic tools to assess the performance of multi-class classifiers. Pattern Recognition Letters 31, 1665–1671 (2010)

[5] Meila, M.: Comparing clusterings—an information based distance. Journal of Multivariate Analysis 28, 875–893 (2007)

[6] Valverde-Albacete, F.J., Peláez-Moreno, C.: 100% classification accuracy considered harmful: The Normalized Information Transfer explains the accuracy paradox (submitted, 2013)

[1] A registered trademark of The MathWorks, Inc.

[7] Mejía-Navarrete, D., Gallardo-Antolín, A., Peláez-Moreno, C., Valverde-Albacete, F.J.: Feature extraction assessment for an acoustic-event classification task using the entropy triangle. In: Interspeech 2010 (2011)

[8] Villena-Román, J., García-Morera, J., Moreno-García, C., Ferrer-Ureña, L., Lana-Serrano, S.: TASS - Workshop on sentiment analysis at SEPLN (2012)

[9] Amigó, E., Corujo, A., Gonzalo, J., Meij, E., Rijke, M.: Overview of RepLab 2012: Evaluating online management systems. In: CLEF (2012)

[10] Greenwood, M.A., Aswani, N., Bontcheva, K.: Reputation profiling with gate. In: CLEF (2012)

[11] Carrillo-de-Albornoz, J., Chugur, I., Amigó, E.: Using an emotion-based model and sentiment analysis techniques to classify polarity for reputation. In: CLEF (2012)

[12] Martín-Wanton, T., Carrillo-de-Albornoz, J.: UNED at TASS 2012: Polarity classification and trending topic system. In: Workshop on Sentiment Analysis at SEPLN (2012)

[13] Carrillo-de-Albornoz, J., Plaza, L., Gervás, P.: A hybrid approach to emotional sentence polarity and intensity classification. In: Conference on Computational Natural Language Learning, CoNLL 2010, pp. 153–161 (2010)

[14] Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in twitter. Language Resources and Evaluation 47, 239–268 (2013)

[15] Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: beyond a simple case of negation. In: Knowledge and Information Systems, 1–20 (2013)