Evaluation Measures for the SemEval-2016 Task 4 "Sentiment Analysis in Twitter" (Draft: Version 1.1)

Preslav Nakov♣, **Alan Ritter**♢, **Sara Rosenthal**♡, **Fabrizio Sebastiani**♣[*], **Veselin Stoyanov**♠
♣Qatar Computing Research Institute, Hamad bin Khalifa University, Qatar
♢Department of Computer Science and Engineering, Ohio State University, US
♡Department of Computer Science, Columbia University, US
♠Facebook, US

This informal document details the evaluation measures that will be used in SemEval-2016 Task 4 "Sentiment Analysis in Twitter", a revamped edition of SemEval-2015 Task 10 (Rosenthal et al., 2015).

Task 4 consists of five subtasks; the evaluation measures that we will use for them will be discussed in Sections 1 to 5. Subtasks B to E conceptually form a "2×2 matrix" (see Table 1), where the rows indicate the *goal* of the task (classification vs. quantification) and the columns indicate the *granularity* of the task (two-point scale vs. five-point scale).

Note that, for each of our five subtasks, the dataset is subdivided into a number of "topics", and the subtask needs to be carried out independently for each topic. As a result, each of the evaluation measures described below is "macroaveraged" across the topics, i.e., we compute the measure individually for each topic, and we then average the results across the topics.

		Granularity	
		Two-point	Five-point
Goal	Classification	Subtask B	Subtask C
Goal	Quantification	Subtask D	Subtask E

Table 1: The "2×2 matrix" of Subtasks B-E.

1 Subtask A: Message Polarity Classification

Subtask A consists of the following problem: *Given a tweet, predict whether the tweet is of positive, negative, or neutral sentiment.* It is thus a "single-label multi-class" classification (SLMCC) task, in which each tweet must be classified as belonging to exactly one of the three classes $\mathcal{C}$={Positive, Neutral, Negative}. This subtask is a rerun; it was also present in SemEval-2013 (Nakov et al., 2013), SemEval-2014 (Rosenthal et al., 2014), SemEval-2015 (Rosenthal et al., 2015) as Subtask B.

For reasons of continuity with the 2013-2015 editions of this subtask, we will adopt the same evaluation measure that was used then, i.e.,

$$F_1^{PN} = \frac{F_1^{Pos} + F_1^{Neg}}{2} \tag{1}$$

F_1^{Pos} is defined

- by taking ρ^{Pos} to be the fraction of Positive tweets that are predicted to be such; in terms of the confusion matrix of Table 2, this means that $\rho^{Pos} = \frac{PP}{PP+UP+NP}$;
- by taking π^{Pos} to be the fraction of tweets predicted to be Positive that are indeed Positive, i.e., $\pi^{Pos} = \frac{PP}{PP+PU+PN}$;
- by taking $F_1^{Pos} = \dfrac{2\pi^{Pos}\rho^{Pos}}{\pi^{Pos} + \rho^{Pos}}$.

F_1^{Neg} is defined similarly, and the evaluation measure we finally adopt is F_1^{PN} as from Equation 1.

2 Subtask B: Tweet classification according to a two-point scale

Subtask B consists of the following problem: *Given a tweet known to be about a given topic, classify whether the tweet conveys a positive or a negative sentiment towards the topic.* As such, it is thus a *binary classification* task, in which each tweet must be classified as belonging to exactly one of the two classes $\mathcal{C}$={Positive, Negative}. This subtask is a simplification of Subtask C as from SemEval-2015, which also required to filter out tweets that

[*] Fabrizio Sebastiani is currently on leave from Consiglio Nazionale delle Ricerche, Italy.

		Actual		
		Pos	Neu	Neg
Predicted	Pos	PP	PU	PN
	Neu	UP	UU	UN
	Neg	NP	NU	NN

Table 2: The confusion matrix for Subtask B. Cell XY stands for "the number of tweets that were labelled as X and should have been labelled as Y", where P U N stand for Positive Neutral Negative, respectively.

were not about the topic, and which (like Subtask A does now – see Section 1) also involved the Neutral class.

As an evaluation measure, for this task we will adopt *macroaveraged recall*, i.e.,

$$\rho^{PN} = \frac{\rho^{Pos} + \rho^{Neg}}{2} \quad (2)$$

where ρ^{Pos} and ρ^{Neg} are as defined in Section 1. ρ^{PN} ranges in $[0, 1]$, where 1 is achieved only by the perfect classifier (the classifier that correctly classifies all items), 0 is achieved only by the perverse classifier (the classifier that misclassifies all items), while 0.5 is

- the value obtained by a trivial classifier (i.e., the classifier that assigns all tweets to the same class – be it Positive or Negative), and
- the expected value of a random classifier.

The advantage of ρ^{PN} over "standard" accuracy is that it is more robust to class imbalance, since for standard accuracy the score of the majority-class classifier is the relative frequency (aka "prevalence") of the majority class, that may be much higher than 0.5 if the test set is imbalanced.

The advantage of ρ^{PN} over F_1 is that it is more robust to class imbalance, since for F_1 the score of the trivial acceptor may be much higher than 0.5 if the test set is imbalanced and the Positive class is the majority class. Another advantage of ρ^{PN} over F_1 is that ρ^{PN} is invariant with respect to switching Positive with Negative, while F_1 is not.

3 Subtask C: Tweet classification according to a five-point scale

Subtask C consists of the following problem: *Given a tweet known to be about a given topic, estimate the sentiment conveyed by the tweet towards the topic on a five-point scale.* As such, it is thus an *ordinal classification* (OC – also known as *ordinal regression*) task, in which each tweet must be classified in exactly one of the classes in $\mathcal{C}$={VeryPositive, Positive, OK, Negative, VeryNegative} (represented in our dataset by numbers in {+2,+1,0,-1,-2}), where there is a total order defined on $\mathcal{C}$. This subtask was not present in SemEval-2015.

The essential difference between SLMCC (see Section 1) and OC is that in the latter not all mistakes weigh equally; e.g., classifying as VeryNegative an item that should be classified as VeryPositive is a more serious mistake than classifying as VeryNegative an item that should be classified as Negative.

As our evaluation measure, we use *macroaveraged mean absolute error* (MAE^M):

$$MAE^M(h, Te) = \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} \frac{1}{|Te_j|} \sum_{\mathbf{x}_i \in Te_j} |h(\mathbf{x}_i) - y_i| \quad (3)$$

where y_i denotes the true label of item $\mathbf{x}_i$, $h(\mathbf{x}_i)$ denotes its predicted label, Te_j denotes the set of test documents whose true class is c_j, $|h(\mathbf{x}_i) - y_i|$ denotes the "distance" between classes $h(\mathbf{x}_i)$ and y_i (e.g., the distance between veryPositive and Negative is 3), and the "M" superscript indicates "macroaveraging".

The advantage of MAE^M over "standard" mean absolute error, which is defined as

$$MAE^\mu(h, Te) = \frac{1}{|Te|} \sum_{\mathbf{x}_i \in Te} |h(\mathbf{x}_i) - y_i| \quad (4)$$

where the "μ" superscript stands for "microaveraging", is that it is robust to class imbalance (which is useful, given the imbalanced nature of our dataset) while coinciding with MAE^μ on perfectly balanced datasets (i.e., datasets with exactly the same number of test documents for each class).

Note that, unlike the measures discussed in Sections 1 and 2, MAE^M is a measure of error, and not a measure of accuracy, so lower values are better. See (Baccianella et al., 2009) for more detail on MAE^M.

4 Subtask D: Tweet quantification according to a two-point scale

Subtask D consists of the following problem: *Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the* Positive *and* Negative *classes.* It is thus a *binary quantification* task, in which each tweet belongs exactly to one of the classes in $\mathcal{C}$={Positive, Negative} and the task is to compute an estimate $\hat{p}(c_j)$ of the relative frequency in the test set $p(c_j)$ of each of the classes in $\mathcal{C}$. This is subtask is related to (yet, different from) SemEval-2015 subtask E.

The essential difference between binary classification (as from Section 2) and binary quantification is that, in the latter, errors of different polarity (e.g., a false positive and a false negative for the same class) compensate each other.

The measure we are going to adopt is *normalized cross-entropy*, better known as *Kullback-Leibler Divergence* (KLD). KLD was proposed as a quantification measure in (Forman, 2005), and is defined as follows:

$$KLD(\hat{p}, p, \mathcal{C}) = \sum_{c_j \in \mathcal{C}} p(c_j) \log \frac{p(c_j)}{\hat{p}(c_j)} \qquad (5)$$

KLD is a measure of the error made in estimating a true distribution p over a set $\mathcal{C}$ of classes by means of a predicted distribution $\hat{p}$. Like MAE^M in Section 3, KLD is a measure of error, so lower values are better. KLD ranges between 0 (best) and $+\infty$ (worst).

Note that the upper bound of KLD is not finite since Equation 5 has predicted probabilities, and not true probabilities, at the denominator: that is, by making a predicted probability $\hat{p}(c_j)$ infinitely small we can make KLD infinitely large. To solve this problem, in computing KLD we smooth both $p(c_j)$ and $\hat{p}(c_j)$ via additive smoothing, i.e.,

$$p_s(c_j) = \frac{p(c_j) + \epsilon}{(\sum_{c_j \in \mathcal{C}} p(c_j)) + \epsilon \cdot |\mathcal{C}|} \qquad (6)$$

where $p_s(c_j)$ denotes the smoothed version of $p(c_j)$ and the denominator is just a normalizing factor (same for the $\hat{p}_s(c_j)$'s); the quantity $\epsilon = \frac{1}{2 \cdot |Te|}$ is used as a smoothing factor, where Te denotes the test set. The smoothed versions of $p(c_j)$ and $\hat{p}(c_j)$ are then used in place of their original versions in Equation 5; as a result, KLD is always defined and still returns a value of 0 when p and $\hat{p}$ coincide.

5 Subtask E: Tweet quantification according to a five-point scale

Subtask E consists of the following problem: *Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the five classes of a five-point scale.*

It is an *ordinal quantification* (OQ) task, in which (as in OC) each tweet belongs exactly to one of the classes in $\mathcal{C}$={VeryPositive, Positive, OK, Negative, VeryNegative}, where there is a total order on $\mathcal{C}$, and (as in binary quantification) the task is to compute an estimate $\hat{p}(c_j)$ of the relative frequency $p(c_j)$ in the test tweets of all the classes $c_j \in \mathcal{C}$. This subtask was not present in SemEval-2015.

The measure we adopt for OQ is the *Earth Mover's Distance* (Rubner et al., 2000), a measure well known in the field of computer vision. When there is a total order on the classes in $\mathcal{C}$, the Earth Mover's Distance is defined as

$$EMD(\hat{p}, p) = \sum_{j=1}^{|\mathcal{C}|-1} |\sum_{i=1}^{j} \hat{p}(c_i) - \sum_{i=1}^{j} p(c_i)| \quad (7)$$

and can be computed in $|\mathcal{C}|$ steps from the estimated and true class prevalences. Like KLD in Section 4, EMD is a measure of error, so lower values are better; EMD ranges between 0 (best) and $|\mathcal{C}| - 1$ (worst). See (Esuli and Sebastiani, 2010) for more detail on EMD.

A Appendix: Useful pointers

Quantification. Several publications in the literature discuss methods for binary quantification: see e.g., (Alaíz-Rodríguez et al., 2011; Barranquero et al., 2015; Esuli and Sebastiani, 2015; Forman, 2008; Hopkins and King, 2010; Milli et al., 2013; Saerens et al., 2002). Some of these papers, e.g., (Esuli and Sebastiani, 2015; Hopkins and King, 2010), contain links for downloading the software for performing quantification. Sentiment quantification is discussed in (Esuli and Sebastiani, 2010); tweet sentiment quantification is discussed in (Gao and Sebastiani, 2015).

Ordinal classification. Ordinal classification has a very rich literature; papers proposing OC methods include, e.g., (Chu and Keerthi, 2007; Dembczyński et al., 2007; Fouad and Tino, 2012; Herbrich et al., 2000; Li and Lin, 2007; Lin and Li, 2006; Lin and Li, 2012; Sun et al., 2010; Xia et al., 2006). A survey on ordinal classification methods can be found in (Gutiérrez et al., 2015). Some of these papers, e.g., (Chu and Keerthi, 2007), contain links for downloading software performing OC.

References

Rocío Alaíz-Rodríguez, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro. 2011. Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift. *Neurocomputing*, 74(16):2614–2623.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA 2009)*, pages 283–287, Pisa, IT.

Jose Barranquero, Jorge Díez, and Juan José del Coz. 2015. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition*, 48(2):591–604.

Wei Chu and S. Sathiya Keerthi. 2007. Support vector ordinal regression. *Neural Computation*, 19(3):145–152.

Krzysztof Dembczyński, Wojciech Kotłowski, and Roman Słowiński. 2007. Ordinal classification with decision rules. In *Proceedings of the ECML/PKDD 2007 workshop on Mining Complex Data*, pages 169–181, Warsaw, PL.

Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiment quantification. *IEEE Intelligent Systems*, 25(4):72–75.

Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery and Data*, 9(4):Article 27.

George Forman. 2005. Counting positives accurately despite inaccurate classification. In *Proceedings of the 16th European Conference on Machine Learning (ECML 2005)*, pages 564–575, Porto, PT.

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206.

Shereen Fouad and Peter Tino. 2012. Adaptive metric learning vector quantization for ordinal classification. *Neural Computation*, 24(11):2825–2851.

Wei Gao and Fabrizio Sebastiani. 2015. Tweet sentiment: From classification to quantification. In *Proceedings of the 7th International Conference on Advances in Social Network Analysis and Mining (ASONAM 2015)*, pages 97–104, Paris, FR.

Pedro Antonio Gutiérrez, María Pérez-Ortiz, Javier Sánchez-Monedero, Francisco Fernández-Navarro, and César Hervás-Martînez. 2015. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*. Forthcoming.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. The MIT Press, Cambridge, US.

Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Ling Li and Hsuan-Tien Lin. 2007. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems*, volume 19, pages 865–872. The MIT Press, Cambridge, US.

Hsuan-Tien Lin and Ling Li. 2006. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory (ALT 2006)*, pages 319–333, Barcelona, ES.

Hsuan-Tien Lin and Ling Li. 2012. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367.

Letizia Milli, Anna Monreale, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi, and Fabrizio Sebastiani. 2013. Quantification trees. In *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM 2013)*, pages 528–536, Dallas, US.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, US.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task

9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, IE.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, US.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41.

Bing-Yu Sun, Jiuyong Li, Desheng Dash Wu, Xiao-Ming Zhang, and Wen-Bo Li. 2010. Kernel discriminant learning for ordinal regression. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):906–910.

Fen Xia, Wensheng Zhang, and Jue Wang. 2006. An effective tree-based algorithm for ordinal regression. *IEEE Intelligent Informatics Bulletin*, 7(1):22–26.