SemEval-2016 Task 4: Sentiment Analysis in Twitter

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, The Ohio State University, USA
♡IBM Watson Health Research, USA
♠Johns Hopkins University, USA

Abstract

This paper discusses the fourth year of the "Sentiment Analysis in Twitter Task". SemEval-2016 Task 4 comprises five subtasks, three of which represent a significant departure from previous editions. The first two subtasks are reruns from prior years and ask to predict the overall sentiment, and the sentiment towards a topic in a tweet. The three new subtasks focus on two variants of the basic "sentiment classification in Twitter" task. The first variant adopts a five-point scale, which confers an *ordinal* character to the classification task. The second variant focuses on the correct estimation of the prevalence of each class of interest, a task which has been called *quantification* in the supervised learning literature. The task continues to be very popular, attracting a total of 43 teams.

1 Introduction

Sentiment classification is the task of detecting whether a textual item (e.g., a product review, a blog post, an editorial, etc.) expresses a POSITIVE or a NEGATIVE opinion in general or about a given entity, e.g., a product, a person, a political party, or a policy. Sentiment classification has become a ubiquitous enabling technology in the Twittersphere. Classifying tweets according to sentiment has many applications in political science, social sciences, market research, and many others (Martínez-Cámara et al., 2014; Mejova et al., 2015).

As a testament to the prominence of research on sentiment analysis in Twitter, the tweet sentiment classification (TSC) task has attracted the highest number of participants in the last three SemEval campaigns (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016b).

Previous editions of the SemEval task involved binary (POSITIVE vs. NEGATIVE) or *single-label multi-class* classification (SLMC) when a NEUTRAL[1] class is added (POSITIVE vs. NEGATIVE vs. NEUTRAL). SemEval-2016 Task 4 represents a significant departure from these previous editions. Although two of the subtasks (Subtasks A and B) are reincarnations of previous editions (SLMC classification for Subtask A, binary classification for Subtask B), SemEval-2016 Task 4 introduces two completely new problems, taken individually (Subtasks C and D) and in combination (Subtask E):

1.1 Ordinal Classification

We replace the two- or three-point scale with a five-point scale {HIGHLYPOSITIVE, POSITIVE, NEUTRAL, NEGATIVE, HIGHLYNEGATIVE}, which is now ubiquitous in the corporate world where human ratings are involved: e.g., Amazon, TripAdvisor, and Yelp, all use a five-point scale for rating sentiment towards products, hotels, and restaurants.

Moving from a categorical two/three-point scale to an ordered five-point scale means, in machine learning terms, moving from binary to *ordinal classification* (a.k.a. *ordinal regression*).

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

[1] We merged OBJECTIVE under NEUTRAL, as previous attempts to have annotators distinguish between the two have consistently resulted in very low inter-annotator agreement.

1.2 Quantification

We replace classification with *quantification*, i.e., supervised class prevalence estimation. With regard to Twitter, hardly anyone is interested in whether *a specific person* has a positive or a negative view of the topic. Rather, applications look at estimating the *prevalence* of positive and negative tweets about a given topic. Most (if not all) tweet sentiment classification studies conducted within political science (Borge-Holthoefer et al., 2015; Kaya et al., 2013; Marchetti-Bowick and Chambers, 2012), economics (Bollen et al., 2011; O'Connor et al., 2010), social science (Dodds et al., 2011), and market research (Burton and Soboleva, 2011; Qureshi et al., 2013), use Twitter with an interest in aggregate data and *not* in individual classifications.

Estimating prevalences (more generally, estimating the *distribution* of the classes in a set of unlabelled items) by leveraging training data is called *quantification* in data mining and related fields. Previous work has argued that quantification is not a mere byproduct of classification, since (a) a good classifier is not necessarily a good quantifier, and vice versa, see, e.g., (Forman, 2008); (b) quantification requires evaluation measures different from classification. Quantification-specific learning approaches have been proposed over the years; Sections 2 and 5 of (Esuli and Sebastiani, 2015) contain several pointers to such literature.

Note that, in Subtasks B to E, tweets come labelled with the *topic* they are about and participants need not classify whether a tweet is about a given topic. A topic can be anything that people express opinions about; for example, a product (e.g., iPhone6), a political candidate (e.g., Hillary Clinton), a policy (e.g., Obamacare), an event (e.g., the Pope's visit to Palestine), etc.

The rest of the paper is structured as follows. In Section 2, we give a general overview of SemEval-2016 Task 4 and the five subtasks. Section 3 focuses on the datasets, and on the data generation procedure. In Section 4, we describe in detail the evaluation measures for each subtask. Section 5 discusses the results of the evaluation and the techniques and tools that the top-ranked participants used. Section 6 concludes, discussing the lessons learned and some possible ideas for a followup at SemEval-2017.

2 Task Definition

SemEval-2016 Task 4 consists of five subtasks:

1. **Subtask A:** Given a tweet, predict whether it is of positive, negative, or neutral sentiment.

2. **Subtask B:** Given a tweet known to be about a given topic, predict whether it conveys a positive or a negative sentiment towards the topic.

3. **Subtask C:** Given a tweet known to be about a given topic, estimate the sentiment it conveys towards the topic on a five-point scale ranging from HIGHLYNEGATIVE to HIGHLYPOSITIVE.

4. **Subtask D:** Given a set of tweets known to be about a given topic, estimate the distribution of the tweets in the POSITIVE and NEGATIVE classes.

5. **Subtask E:** 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, ranging from HIGHLYNEGATIVE to HIGHLYPOSITIVE.

Subtask A is a rerun – it was present in all three previous editions of the task. In the 2013-2015 editions, it was known as Subtask B.[2] We ran it again this year because it was the most popular subtask in the three previous task editions. It was the most popular subtask this year as well – see Section 5.

Subtask B is a variant of SemEval-2015 Task 10 Subtask C (Rosenthal et al., 2015; Nakov et al., 2016b), with POSITIVE, NEUTRAL, and NEGATIVE as the classification labels.

Subtask E is similar to SemEval-2015 Task 10 Subtask D, which consisted of the following problem: *Given a set of messages on a given topic from the same period of time, classify the overall sentiment towards the topic in these messages as strongly positive, weakly positive, neutral, weakly negative, or strongly negative*. Note that in SemEval-2015 Task 10 Subtask D, exactly one of the five classes had to be chosen, while in our Subtask E, a distribution across the five classes has to be estimated.

[2]Note that we retired the expression-level subtask A, which was present in SemEval 2013–2015 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016b).

As per the above discussion, Subtasks B to E are new. Conceptually, they form a 2×2 matrix, as shown in Table 1, where the rows indicate the *goal* of the task (classification vs. quantification) and the columns indicate the *granularity* of the task (two- vs. five-point scale).

		Granularity	
		Two-point (binary)	Five-point (ordinal)
Goal	Classification	Subtask B	Subtask C
	Quantification	Subtask D	Subtask E

Table 1: A 2×2 matrix summarizing the similarities and the differences between Subtasks B-E.

3 Datasets

In this section, we describe the process of collection and annotation of the training, development and testing tweets for all five subtasks. We dub this dataset the *Tweet 2016* dataset in order to distinguish it from datasets generated in previous editions of the task.

3.1 Tweet Collection

We provided the datasets from the previous editions[3] (see Table 2) of this task (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016b) for training and development. In addition we created new training and testing datasets.

Dataset	POSITIVE	NEGATIVE	NEUTRAL	Total
Twitter2013-train	3,662	1,466	4,600	9,728
Twitter2013-dev	575	340	739	1,654
Twitter2013-test	1,572	601	1,640	3,813
SMS2013-test	492	394	1,207	2,093
Twitter2014-test	982	202	669	1,853
Twitter2014-sarcasm	33	40	13	86
LiveJournal2014-test	427	304	411	1,142
Twitter2015-test	1,040	365	987	2,392

Table 2: Statistics about data from the 2013-2015 editions of the SemEval task on Sentiment Analysis in Twitter, which could be used for training and development for SemEval-2016 Task 4.

[3]For Subtask A, we did not allow training on the testing datasets from 2013–2015, as we used them for progress testing.

We employed the following annotation procedure. As in previous years, we first gathered tweets that express sentiment about popular topics. For this purpose, we extracted named entities from millions of tweets, using a Twitter-tuned named entity recognition system (Ritter et al., 2011). The collected tweets were greatly skewed towards the neutral class. In order to reduce the class imbalance, we removed those that contained no sentiment-bearing words. We used SentiWordNet 3.0 (Baccianella et al., 2010) as a repository of sentiment words. Any word listed in SentiWordNet 3.0 with at least one sense having a positive or a negative sentiment score greater than 0.3 was considered sentiment-bearing.[4]

The training and development tweets were collected from July to October 2015. The test tweets were collected from October to December 2015. We used the public streaming Twitter API to download the tweets.[5]

We then manually filtered the resulting tweets to obtain a set of 200 meaningful topics with at least 100 tweets each (after filtering out near-duplicates). We excluded topics that were incomprehensible, ambiguous (e.g., *Barcelona*, which is the name both of a city and of a sports team), or too general (e.g., *Paris*, which is the name of a big city). We then discarded tweets that were just mentioning the topic but were not really about the topic.

Note that the topics in the training and in the test sets do not overlap, i.e., the test set consists of tweets about topics different from the topics the training and development tweets are about.

3.2 Annotation

The 2016 data consisted of four parts: TRAIN (for training models), DEV (for tuning models), DEVTEST (for development-time evaluation), and TEST (for the official evaluation). The first three datasets were annotated using Amazon's Mechanical Turk, while the TEST dataset was annotated on CrowdFlower.

[4]Filtering based on an existing lexicon does bias the dataset to some degree; however, the text still contains sentiment expressions outside those in the lexicon.

[5]We distributed the datasets to the task participants in a similar way: we only released the annotations and the tweet IDs, and the participants had to download the actual tweets by themselves via the API, for which we provided a script: `https://github.com/aritter/twitter_download`

Instructions: Given a Twitter message and a topic, identify whether the message is highly positive, positive, neutral, negative, or highly negative (a) in general and (b) with respect to the provided topic. If a tweet is sarcastic, please select the checkbox "The tweet is sarcastic". Please read the examples and the invalid responses before beginning if this is the first time you are working on this HIT.

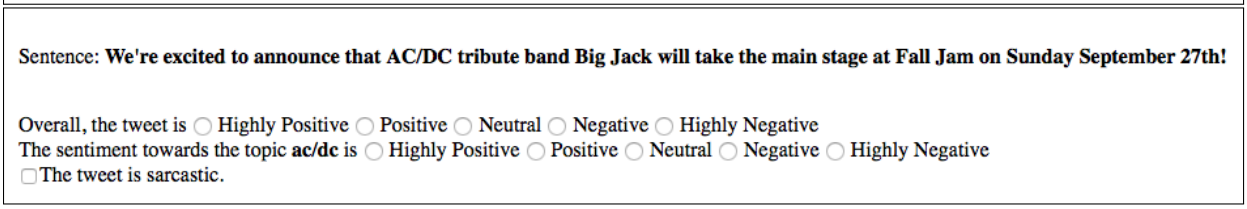

Figure 1: The instructions provided to the Mechanical Turk annotators, followed by a screenshot.

Annotation with Amazon's Mechanical Turk. A Human Intelligence Task (HIT) consisted of providing all required annotations for a given tweet message. In order to qualify to work on our HITs, a Mechanical Turk annotator (a.k.a. "Turker") had to have an approval rate greater than 95% and to have completed at least 50 approved HITs. Each HIT was carried out by five Turkers and consisted of five tweets to be annotated. A Turker had to indicate the overall polarity of the tweet message (on a five-point scale) as well as the overall polarity of the message towards the given target topic (again, on a five-point scale). The annotation instructions along with an example are shown in Figure 1. We made available to the Turkers several additional examples, which are shown in Table 3.

We rejected HITs with the following problems:

- one or more responses do not have the overall sentiment marked;
- one or more responses do not have the sentiment towards the topic marked;
- one or more responses appear to be randomly selected.

Annotation with CrowdFlower. We annotated the TEST data using CrowdFlower, as it allows better quality control of the annotations across a number of dimensions. Most importantly, it allows us to find and exclude unreliable annotators based on hidden tests, which we created starting with the highest-confidence and highest-agreement annotations from Mechanical Turk. We added some more tests manually. Otherwise, we setup the annotation task giving exactly the same instructions and examples as in Mechanical Turk.

Consolidation of annotations. In previous years, we used majority voting to select the true label (and discarded cases where a majority had not emerged, which amounted to about 50% of the tweets). As this year we have a five-point scale, where the expected agreement is lower, we used a two-step procedure. If three out of the five annotators agreed on a label, we accepted the label. Otherwise, we first mapped the categorical labels to the integer values -2, -1, 0, 1, 2. Then we calculated the average, and finally we mapped that average to the closest integer value. In order to counter-balance the tendency of the average to stay away from -2 and 2, and also to prefer 0, we did not use rounding at ± 0.5 and ± 1.5, but at ± 0.4 and ± 1.4 instead.

To give the reader an idea about the degree of agreement, we will look at the TEST dataset as an example. It included 20,632 tweets. For 2,760, all five annotators assigned the same value, and for another 9,944 there was a majority value. For the remaining 7,928 cases, we had to perform averaging as described above.

The consolidated statistics from the five annotators on a three-point scale for Subtask A are shown in Table 4. Note that, for consistency, we annotated the data for Subtask A on a five-point scale, which we then converted to a three-point scale.

The topic annotations on a two-point scale for Subtasks B and D are shown in Table 5, while those on a five-point scale for Subtasks C and E are in Table 6. Note that, as for Subtask A, the two-point scale annotation counts for Subtasks B and D derive from summing the POSITIVEs with the HIGHLYPOSITIVEs, and the NEGATIVEs with the HIGHLYNEGATIVEs from Table 6; moreover, this time we also remove the NEUTRALs.

Tweet	Overall Sentiment	Topic Sentiment
Why would you still wear shorts when it's this cold?! I love how Britain see's a bit of sun and they're like 'OOOH LET'S STRIP!	POSITIVE	Britain: NEGATIVE
Saturday without Leeds United is like Sunday dinner it doesn't feel normal at all (Ryan)	NEGATIVE	Leeds United: HIGHLYPOSITIVE
Who are you tomorrow? Will you make me smile or just bring me sorrow? #HottieOfTheWeek Demi Lovato	NEUTRAL	Demi Lovato: POSITIVE

Table 3: List of example tweets and annotations that were provided to the annotators.

	POSITIVE	NEUTRAL	NEGATIVE	**Total**
TRAIN	3,094	863	2,043	6,000
DEV	844	765	391	2,000
DEVTEST	994	681	325	2,000
TEST	7,059	10,342	3,231	20,632

Table 4: 2016 data statistics (Subtask A).

	Topics	POSITIVE	NEGATIVE	**Total**
TRAIN	60	3,591	755	4,346
DEV	20	986	339	1,325
DEVTEST	20	1,153	264	1,417
TEST	100	8,212	2,339	10,551

Table 5: 2016 data statistics (Subtasks B and D).

	Topics	HIGHLYPOSITIVE	POSITIVE	NEUTRAL	NEGATIVE	HIGHLYNEGATIVE	**Total**
TRAIN	60	437	3,154	1,654	668	87	6,000
DEV	20	53	933	675	296	43	2,000
DEVTEST	20	148	1,005	583	233	31	2,000
TEST	100	382	7,830	10,081	2,201	138	20,632

Table 6: 2016 data statistics (Subtasks C and E).

As we use the same test tweets for all subtasks, the submission of results by participating teams was subdivided in two stages: (*i*) participants had to submit results for Subtasks A, C, E, and (*ii*) only after the submission deadline for A, C, E had passed, we distributed to participants the unlabelled test data for Subtasks B and D.

Otherwise, since for Subtasks B and D we filter out the NEUTRALs, we would have leaked information about which the NEUTRALs are, and this information could have been used in Subtasks C and E.

Finally, as the same tweets can be selected for different topics, we ended up with some duplicates; arguably, these are true duplicates for Subtask A only, as for the other subtasks the topics still differ. This includes 25 duplicates in TRAIN, 3 in DEV, 2 in DEVTEST, and 116 in TEST. There is a larger number in TEST, as TEST is about twice as large as TRAIN, DEV, and DEVTEST combined. This is because we wanted a large TEST set with 100 topics and 200 tweets per topic on average for Subtasks C and E.

4 Evaluation Measures

This section discuss the evaluation measures for the five subtasks of our SemEval-2016 Task 4. A document describing the evaluation measures in detail[6] (Nakov et al., 2016a), and a scoring software implementing all the five "official" measures, were made available to the participants via the task website together with the training data.[7]

For Subtasks B to E, the datasets are each 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 will be "macroaveraged" across the topics, i.e., we compute the measure individually for each topic, and we then average the results across the topics.

[6] http://alt.qcri.org/semeval2016/task4/

[7] An earlier version of the scoring script contained a bug, to the effect that for Subtask B it was computing F_1^{PN}, and not ρ^{PN}. This was detected only after the submissions were closed, which means that participants to Subtask B who used the scoring system (and not their own implementation of ρ^{PN}) for parameter optimization, may have been penalized in the ranking as a result.

4.1 Subtask A: Message polarity classification

Subtask A is a *single-label multi-class* (SLMC) classification task. Each tweet must be classified as belonging to exactly one of the following three classes $\mathcal{C}$={POSITIVE, NEUTRAL, NEGATIVE}.

We adopt the same evaluation measure as the 2013-2015 editions of this subtask, F_1^{PN}:

$$F_1^{PN} = \frac{F_1^P + F_1^N}{2} \quad (1)$$

F_1^P is the F_1 score for the POSITIVE class:

$$F_1^P = \frac{2\pi^P \rho^P}{\pi^P + \rho^P} \quad (2)$$

Here, π^P and ρ^P denote precision and recall for the POSITIVE class, respectively:

$$\pi^P = \frac{PP}{PP + PU + PN} \quad (3)$$

$$\rho^P = \frac{PP}{PP + UP + NP} \quad (4)$$

where PP, UP, NP, PU, PN are the cells of the confusion matrix shown in Table 7.

		Gold Standard		
		POSITIVE	NEUTRAL	NEGATIVE
Predicted	POSITIVE	PP	PU	PN
	NEUTRAL	UP	UU	UN
	NEGATIVE	NP	NU	NN

Table 7: The confusion matrix for Subtask A. Cell XY stands for "the number of tweets that the classifier labeled X and the gold standard labells as Y". P, U, N stand for POSITIVE, NEUTRAL, NEGATIVE, respectively.

F_1^N is defined analogously, and the measure we finally adopt is F_1^{PN} as from Equation 1.

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

Subtask B is a *binary classification* task. Each tweet must be classified as either POSITIVE or NEGATIVE.

For this subtask we adopt *macroaveraged recall*:

$$\begin{aligned} \rho^{PN} &= \frac{1}{2}(\rho^P + \rho^N) \\ &= \frac{1}{2}(\frac{PP}{PP + NP} + \frac{NN}{NN + PN}) \end{aligned} \quad (5)$$

In the above formula, ρ^P and ρ^N are the positive and the negative class recall, respectively. Note that U terms are entirely missing in Equation 5; this is because we do not have the NEUTRAL class for SemEval-2016 Task 4, subtask A.

ρ^{PN} ranges in $[0, 1]$, where a value of 1 is achieved only by the perfect classifier (i.e., the classifier that correctly classifies all items), a value of 0 is achieved only by the perverse classifier (the classifier that misclassifies all items), while 0.5 is both (*i*) 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 (*ii*) the expected value of a random classifier. The advantage of ρ^{PN} over "standard" accuracy is that it is more robust to class imbalance. The accuracy 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. Standard F_1 is also sensitive to class imbalance for the same reason. 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. See (Sebastiani, 2015) for more details on ρ^{PN}.

As we noted before, the training dataset, the development dataset, and the test dataset are each subdivided into a number of topics, and Subtask B needs to be carried out independently for each topic. As a result, the evaluation measures discussed in this section are computed individually for each topic, and the results are then averaged across topics to yield the final score.

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

Subtask C is an *ordinal classification* (OC – also known as *ordinal regression*) task, in which each tweet must be classified into exactly one of the classes in $\mathcal{C}$={HIGHLYPOSITIVE, POSITIVE, NEUTRAL, NEGATIVE, HIGHLYNEGATIVE}, represented in our dataset by numbers in $\{+2,+1,0,-1,-2\}$, with a total order defined on $\mathcal{C}$. The essential difference between SLMC (see Section 4.1 above) and OC is that not all mistakes weigh equally in the latter. For example, misclassifying a HIGHLYNEGATIVE example as HIGHLYPOSITIVE is a bigger mistake than misclassifying it as NEGATIVE or NEUTRAL.

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

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

where y_i denotes the true label of item $\mathbf{x}_i$, $h(\mathbf{x}_i)$ is 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 HIGHLYPOSITIVE and NEGATIVE is 3), and the "M" superscript indicates "macroaveraging".

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

$$MAE^{\mu}(h, Te) = \frac{1}{|Te|}\sum_{\mathbf{x}_i \in Te}|h(\mathbf{x}_i) - y_i| \quad (6)$$

is that it is robust to class imbalance (which is useful, given the imbalanced nature of our dataset). On perfectly balanced datasets MAE^M and MAE^{μ} are equivalent.

Unlike the measures discussed in Sections 4.1 and 4.2, MAE^M is a measure of error, and not accuracy, and thus lower values are better. See (Baccianella et al., 2009) for more detail on MAE^M.

Similarly to Subtask B, Subtask C needs to be carried out independently for each topic. As a result, MAE^M is computed individually for each topic, and the results are then averaged across all topics to yield the final score.

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

Subtask D also assumes a *binary quantification* setup, in which each tweet is classified as POSITIVE or NEGATIVE. The task is to compute an estimate $\hat{p}(c_j)$ of the relative frequency (in the test set) of each of the classes.

The difference between binary classification (as from Section 4.2) and binary quantification is that errors of different polarity (e.g., a false positive and a false negative for the same class) can compensate each other in the latter. Quantification is thus a more lenient task since a perfect classifier is also a perfect quantifier, but a perfect quantifier is not necessarily a perfect classifier.

We adopt *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:

$$KLD(\hat{p}, p, \mathcal{C}) = \sum_{c_j \in \mathcal{C}} p(c_j)\log_e\frac{p(c_j)}{\hat{p}(c_j)} \quad (7)$$

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 4.3, KLD is a measure of error, which means that lower values are better. KLD ranges between 0 (best) and $+\infty$ (worst).

Note that the upper bound of KLD is not finite since Equation 7 has predicted prevalences, and not true prevalences, at the denominator: that is, by making a predicted prevalence $\hat{p}(c_j)$ 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.,

$$\begin{aligned} p^s(c_j) &= \frac{p(c_j) + \epsilon}{(\sum_{c_j \in \mathcal{C}} p(c_j)) + \epsilon \cdot |\mathcal{C}|} \\ &= \frac{p(c_j) + \epsilon}{1 + \epsilon \cdot |\mathcal{C}|} \end{aligned} \quad (8)$$

where $p^s(c_j)$ denotes the smoothed version of $p(c_j)$ and the denominator is just a normalizer (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 used in place of their original versions in Equation 7; as a result, KLD is always defined and still returns a value of 0 when p and $\hat{p}$ coincide.

KLD is computed individually for each topic, and the results are averaged to yield the final score.

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

Subtask E is an *ordinal quantification* (OQ) task, in which (as in OC) each tweet belongs exactly to one of the classes in $\mathcal{C}$={HIGHLYPOSITIVE, POSITIVE, NEUTRAL, NEGATIVE, HIGHLYNEGATIVE}, where there is a total order on $\mathcal{C}$. 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 $c_j \in \mathcal{C}$.

The measure we adopt for OQ is the *Earth Mover's Distance* (Rubner et al., 2000) (also known as the *Vaserštein metric* (Rüschendorf, 2001)), a measure well-known in the field of computer vision. EMD is currently the only known measure for ordinal quantification. It is defined for the general case in which a distance $d(c', c'')$ is defined for each $c', c'' \in \mathcal{C}$. When there is a total order on the classes in $\mathcal{C}$ and $d(c_i, c_{i+1}) = 1$ for all $i \in \{1, ..., (\mathcal{C}-1)\}$ (as in our application), the Earth Mover's Distance is defined as

$$EMD(\hat{p}, p) = \sum_{j=1}^{|\mathcal{C}|-1} |\sum_{i=1}^{j} \hat{p}(c_i) - \sum_{i=1}^{j} p(c_i)| \quad (9)$$

and can be computed in $|\mathcal{C}|$ steps from the estimated and true class prevalences.

Like KLD in Section 4.4, 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 details on EMD.

As before, EMD is computed individually for each topic, and the results are then averaged across all topics to yield the final score.

5 Participants and Results

A total of 43 teams (see Table 15 at the end of the paper) participated in SemEval-2016 Task 4, representing 25 countries; the country with the highest participation was China (5 teams), followed by Italy, Spain, and USA (4 teams each). The subtask with the highest participation was Subtask A (34 teams), followed by Subtask B (19 teams), Subtask D (14 teams), Subtask C (11 teams), and Subtask E (10 teams).

It was not surprising that Subtask A proved to be the most popular – it was a rerun from previous years; conversely, none among Subtasks B to E had previously been offered in precisely the same form. Quantification-related subtasks (D and E) generated 24 participations altogether, while subtasks with an ordinal nature (C and E) attracted 21 participations. Only three teams participated in all five subtasks; conversely, no less than 23 teams took part in one subtask only (with a few exceptions, Subtask A). Many teams that participated in more than one subtask used essentially the same system for all of them, with little tuning to the specifics of each subtask.

Few trends stand out among the participating systems. In terms of the supervised learning methods used, there is a clear dominance of methods based on deep learning, including convolutional neural networks and recurrent neural networks (and, in particular, long short-term memory networks); the software libraries for deep learning most frequently used by the participants are Theano and Keras. Conversely, kernel machines seem to be less frequently used than in the past, and the use of learning methods other than the ones mentioned above is scarce.

The use of distant supervision is ubiquitous; this is natural, since there is an abundance of freely available tweets labelled according to sentiment (possibly with silver labels only, e.g., emoticons), and it is intuitive that their use as additional training data could be helpful. Another ubiquitous technique is the use of word embeddings, usually generated via either word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014); most authors seem to use general-purpose, pre-trained embeddings, while some authors also use customized word embeddings, trained either on the Tweet 2016 dataset or on tweet datasets of some sort.

Nothing radically new seems to have emerged with respect to text preprocessing; as in previous editions of this task, participants use a mix of by now obvious techniques, such as negation scope detection, elongation normalization, detection of amplifiers and diminishers, plus the usual extraction of word n-grams, character n-grams, and POS n-grams. The use of sentiment lexicons (alone or in combination with each other; general-purpose or Twitter-specific) is obviously still frequent.

In the next five subsections, we discuss the results of the participating systems in the five subtasks, focusing on the techniques and tools that the top-ranked participants have used. We also focus on how the participants tailored (if at all) their approach to the specific subtask. When discussing a specific subtask, we will adopt the convention of adding to a team name a subscript which indicates the position in the ranking for that subtask that the team obtained; e.g., when discussing Subtask E, "Finki$_2$" indicates team "Finki, which placed 2nd in the ranking for Subtask E". The papers describing the participants' approach are quoted in Table 15.

5.1 Subtask A: Message polarity classification

Table 8 ranks the systems submitted by the 34 teams who participated in Subtask A "Message Polarity Classification" in terms of the official measure F_1^{PN}. We further show the result for two other measures, ρ^{PN} (the measure that we adopted for Subtask B) and accuracy ($Acc = \frac{TP+TN}{TP+TN+FP+FN}$). We also report the result for a baseline classifier that assigns to each tweet the POSITIVE class. For Subtask A evaluated using F_1^{PN}, this is the equivalent of the majority class classifier for (binary or SLMC) classification evaluated via vanilla accuracy, i.e., this is the "smartest" among the trivial policies that attempt to maximize F_1^{PN}.

#	System	F_1^{PN}	ρ^{PN}	Acc
1	SwissCheese	**0.633**$_1$	0.667$_2$	0.646$_1$
2	SENSEI-LIF	**0.630**$_2$	0.670$_1$	0.617$_7$
3	UNIMELB	**0.617**$_3$	0.641$_5$	0.616$_8$
4	INESC-ID	**0.610**$_4$	0.662$_3$	0.600$_{10}$
5	aueb.twitter.sentiment	**0.605**$_5$	0.644$_4$	0.629$_6$
6	SentiSys	**0.598**$_6$	0.641$_5$	0.609$_9$
7	I2RNTU	**0.596**$_7$	0.637$_7$	0.593$_{12}$
8	INSIGHT-1	**0.593**$_8$	0.616$_{11}$	0.635$_5$
9	TwiSE	**0.586**$_9$	0.598$_{16}$	0.528$_{24}$
10	ECNU (*)	**0.585**$_{10}$	0.617$_{10}$	0.571$_{16}$
11	NTNUSentEval	**0.583**$_{11}$	0.619$_8$	0.643$_2$
12	MDSENT	**0.580**$_{12}$	0.592$_{18}$	0.545$_{20}$
	CUFE	**0.580**$_{12}$	0.619$_8$	0.637$_4$
14	THUIR	**0.576**$_{14}$	0.605$_{15}$	0.596$_{11}$
	PUT	**0.576**$_{14}$	0.607$_{13}$	0.584$_{14}$
16	LYS	**0.575**$_{16}$	0.615$_{12}$	0.585$_{13}$
17	IIP	**0.574**$_{17}$	0.579$_{19}$	0.537$_{23}$
18	UniPI	**0.571**$_{18}$	0.607$_{13}$	0.639$_3$
19	DIEGOLab16 (*)	**0.554**$_{19}$	0.593$_{17}$	0.549$_{19}$
20	GTI	**0.539**$_{20}$	0.557$_{21}$	0.518$_{26}$
21	OPAL	**0.505**$_{21}$	0.560$_{20}$	0.541$_{22}$
22	DSIC-ELIRF	**0.502**$_{22}$	0.511$_{25}$	0.513$_{27}$
23	UofL	**0.499**$_{23}$	0.537$_{22}$	0.572$_{15}$
	ELiRF	**0.499**$_{23}$	0.516$_{24}$	0.543$_{21}$
25	ISTI-CNR	**0.494**$_{25}$	0.529$_{23}$	0.567$_{17}$
26	SteM	**0.478**$_{26}$	0.496$_{27}$	0.452$_{31}$
27	Tweester	**0.455**$_{27}$	0.503$_{26}$	0.523$_{25}$
28	Minions	**0.415**$_{28}$	0.485$_{28}$	0.556$_{18}$
29	Aicyber	**0.402**$_{29}$	0.457$_{29}$	0.506$_{28}$
30	mib	**0.401**$_{30}$	0.438$_{30}$	0.480$_{29}$
31	VCU-TSA	**0.372**$_{31}$	0.390$_{32}$	0.382$_{32}$
32	SentimentalITists	**0.339**$_{32}$	0.424$_{31}$	0.480$_{29}$
33	WR	**0.330**$_{33}$	0.333$_{34}$	0.298$_{34}$
34	CICBUAPnlp	**0.303**$_{34}$	0.377$_{33}$	0.374$_{33}$
	Baseline	**0.255**	0.333	0.342

Table 8: Results for Subtask A "Message Polarity Classification" on the Tweet 2016 dataset. The systems are ordered by their F_1^{PN} score. In each column the rankings according to the corresponding measure are indicated with a subscript. Teams marked as "(*)" are late submitters, i.e., their original submission was deemed irregular by the organizers, and a revised submission was entered after the deadline.

All 34 participating systems were able to outperform the baseline on all three measures, with the exception of one system that scored below the baseline on Acc. The top-scoring team (SwissCheese$_1$) used an ensemble of convolutional neural networks, differing in their choice of filter shapes, pooling shapes and usage of hidden layers. Word embeddings generated via word2vec were also used, and the neural networks were trained by using distant supervision. Out of the 10 top-ranked teams, 5 teams (SwissCheese$_1$, SENSEI-LIF$_2$, UNIMELB$_3$, INESC-ID$_4$, INSIGHT-1$_8$) used deep NNs of some sort, and 7 teams (SwissCheese$_1$, SENSEI-LIF$_2$, UNIMELB$_3$, INESC-ID$_4$, aueb.twitter.sentiment$_5$, I2RNTU$_7$, INSIGHT-1$_8$) used either general-purpose or task-specific word embeddings, generated via word2vec or GloVe.

Historical results. We also tested the participating systems on the test sets from the three previous editions of this subtask. Participants were not allowed to use these test sets for training. Results (measured on F_1^{PN}) are reported in Table 9. The top-performing systems on Tweet 2016 are also top-ranked on the test datasets from previous years. There is a general pattern: the top-ranked system in year x outperforms the top-ranked system in year $(x-1)$ on the official dataset of year $(x-1)$. Top-ranked systems tend to use approaches that are universally strong, even when tested on out-of-domain test sets such as SMS, LiveJournal, or sarcastic tweets (yet, for sarcastic tweets, there are larger differences in rank compared to systems rankings on Tweet 2016). It is unclear where improvements come from: (a) the additional training data that we made available this year (in addition to Tweet-train-2013, which was used in 2013–2015), thus effectively doubling the amount of training data, or (b) because of advancement of learning methods.

We further look at the top scores achieved by any system in the period 2013–2016. The results are shown in Table 10. Interestingly, the results for a test set improve in the second year it is used (i.e., the year after it was used as an official test set) by 1–3 points absolute, but then do not improve further and stay stable, or can even decrease a bit. This might be due to participants optimizing their systems primarily on the test set from the preceding year.

#	System	2013 Tweet	2013 SMS	2014 Tweet	2014 Tweet sarcasm	2014 Live-Journal	2015 Tweet	2016 Tweet
1	SwissCheese	0.700_{4}	0.637_{2}	0.716_{4}	0.566_{1}	0.695_{7}	0.671_{1}	$\mathbf{0.633}_{1}$
2	SENSEI-LIF	0.706_{3}	0.634_{3}	0.744_{1}	0.467_{8}	0.741_{1}	0.662_{2}	$\mathbf{0.630}_{2}$
3	UNIMELB	0.687_{6}	0.593_{9}	0.706_{6}	0.449_{11}	0.683_{9}	0.651_{4}	$\mathbf{0.617}_{3}$
4	INESC-ID	0.723_{1}	0.609_{6}	0.727_{2}	0.554_{2}	0.702_{4}	0.657_{3}	$\mathbf{0.610}_{4}$
5	aueb.twitter.sentiment	0.666_{7}	0.618_{5}	0.708_{5}	0.410_{17}	0.695_{7}	0.623_{7}	$\mathbf{0.605}_{5}$
6	SentiSys	0.714_{2}	0.633_{4}	0.723_{3}	0.515_{4}	0.726_{2}	0.644_{5}	$\mathbf{0.598}_{6}$
7	I2RNTU	0.693_{5}	0.597_{7}	0.680_{7}	0.469_{6}	0.696_{6}	0.638_{6}	$\mathbf{0.596}_{7}$
8	INSIGHT-1	0.602_{16}	0.582_{12}	0.644_{15}	0.391_{23}	0.559_{23}	0.595_{16}	$\mathbf{0.593}_{8}$
9	TwiSE	0.610_{15}	0.540_{16}	0.645_{13}	0.450_{10}	0.649_{13}	0.621_{8}	$\mathbf{0.586}_{9}$
10	ECNU (*)	0.643_{9}	0.593_{9}	0.662_{8}	0.425_{14}	0.663_{10}	0.606_{11}	$\mathbf{0.585}_{10}$
11	NTNUSentEval	0.623_{11}	0.641_{1}	0.651_{10}	0.427_{13}	0.719_{3}	0.599_{13}	$\mathbf{0.583}_{11}$
12	MDSENT	0.589_{19}	0.509_{20}	0.587_{20}	0.386_{24}	0.606_{18}	0.593_{17}	$\mathbf{0.580}_{12}$
	CUFE	0.642_{10}	0.596_{8}	0.662_{8}	0.466_{9}	0.697_{5}	0.598_{14}	$\mathbf{0.580}_{12}$
14	THUIR	0.616_{12}	0.575_{14}	0.648_{11}	0.399_{20}	0.640_{15}	0.617_{10}	$\mathbf{0.576}_{14}$
	PUT	0.565_{21}	0.511_{19}	0.614_{19}	0.360_{27}	0.648_{14}	0.597_{15}	$\mathbf{0.576}_{14}$
16	LYS	0.650_{8}	0.579_{13}	0.647_{12}	0.407_{18}	0.655_{11}	0.603_{12}	$\mathbf{0.575}_{16}$
17	IIP	0.598_{17}	0.465_{23}	0.645_{13}	0.405_{19}	0.640_{15}	0.619_{9}	$\mathbf{0.574}_{17}$
18	UniPI	0.592_{18}	0.585_{11}	0.627_{17}	0.381_{25}	0.654_{12}	0.586_{18}	$\mathbf{0.571}_{18}$
19	DIEGOLab16 (*)	0.611_{14}	0.506_{21}	0.618_{18}	0.497_{5}	0.594_{20}	0.584_{19}	$\mathbf{0.554}_{19}$
20	GTI	0.612_{13}	0.524_{17}	0.639_{16}	0.468_{7}	0.623_{17}	0.584_{19}	$\mathbf{0.539}_{20}$
21	OPAL	0.567_{20}	0.562_{15}	0.556_{23}	0.395_{21}	0.593_{21}	0.531_{21}	$\mathbf{0.505}_{21}$
22	DSIC-ELIRF	0.494_{25}	0.404_{26}	0.546_{26}	0.342_{29}	0.517_{24}	0.531_{21}	$\mathbf{0.502}_{22}$
23	UofL	0.490_{26}	0.443_{24}	0.547_{25}	0.372_{26}	0.574_{22}	0.502_{25}	$\mathbf{0.499}_{23}$
	ELiRF	0.462_{28}	0.408_{25}	0.514_{28}	0.310_{33}	0.493_{25}	0.493_{26}	$\mathbf{0.499}_{23}$
25	ISTI-CNR	0.538_{22}	0.492_{22}	0.572_{21}	0.327_{30}	0.598_{19}	0.508_{24}	$\mathbf{0.494}_{25}$
26	SteM	0.518_{23}	0.315_{29}	0.571_{22}	0.320_{32}	0.405_{28}	0.517_{23}	$\mathbf{0.478}_{26}$
27	Tweester	0.506_{24}	0.340_{28}	0.529_{27}	0.540_{3}	0.379_{29}	0.479_{28}	$\mathbf{0.455}_{27}$
28	Minions	0.489_{27}	0.521_{18}	0.554_{24}	0.420_{16}	0.475_{26}	0.481_{27}	$\mathbf{0.415}_{28}$
29	Aicyber	0.418_{29}	0.361_{27}	0.457_{29}	0.326_{31}	0.440_{27}	0.432_{29}	$\mathbf{0.402}_{29}$
30	mib	0.394_{30}	0.310_{30}	0.415_{31}	0.352_{28}	0.359_{31}	0.413_{31}	$\mathbf{0.401}_{30}$
31	VCU-TSA	0.383_{31}	0.307_{31}	0.444_{30}	0.425_{14}	0.336_{32}	0.416_{30}	$\mathbf{0.372}_{31}$
32	SentimentalITists	0.339_{33}	0.238_{33}	0.393_{32}	0.288_{34}	0.323_{34}	0.343_{33}	$\mathbf{0.339}_{32}$
33	WR	0.355_{32}	0.284_{32}	0.393_{32}	0.430_{12}	0.366_{30}	0.377_{32}	$\mathbf{0.330}_{33}$
34	CICBUAPnlp	0.193_{34}	0.193_{34}	0.335_{34}	0.393_{22}	0.326_{33}	0.303_{34}	$\mathbf{0.303}_{34}$

Table 9: Historical results for Subtask A "Message Polarity Classification". The systems are ordered by their score on the Tweet 2016 dataset; the rankings on the individual datasets are indicated with a subscript. The meaning of "(*)" is as in Table 8.

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

Table 11 ranks the 19 teams who participated in Subtask B "Tweet classification according to a two-point scale" in terms of the official measure ρ^{PN}. Two other measures are reported, F_1^{PN} (the measure adopted for Subtask A) and accuracy (Acc). We also report the result of a baseline that assigns to each tweet the positive class. This is the "smartest" among the trivial policies that attempt to maximize ρ^{PN}. This baseline always returns $\rho^{PN} = 0.500$.

Note however that this is also (*i*) the value returned by the classifier that assigns to each tweet the negative class, and (*ii*) the expected value returned by the random classifier; for more details see (Sebastiani, 2015, Section 5), where ρ^{PN} is called K.

The top-scoring team (Tweester$_1$) used a combination of convolutional neural networks, topic modeling, and word embeddings generated via word2vec. Similar to Subtask A, the main trend among all participants is the widespread use of deep learning techniques.

Year	2013		2014			2015	2016
	Tweet	SMS	Tweet	Tweet sarcasm	Live-Journal	Tweet	Tweet
Best in 2016	0.723	0.641	0.744	0.566	0.741	0.671	0.633
Best in 2015	0.728	0.685	0.744	0.591	0.753	0.648	–
Best in 2014	0.721	0.703	0.710	0.582	0.748	–	–
Best in 2013	0.690	0.685	–	–	–	–	–

Table 10: Historical results for the best systems for Subtask A "Message Polarity Classification" over the years 2013–2016.

#	System	ρ^{PN}	F_1^{PN}	Acc
1	Tweester	$\mathbf{0.797}_1$	0.799_1	0.862_3
2	LYS	$\mathbf{0.791}_2$	0.720_{10}	0.762_{17}
3	thecerealkiller	$\mathbf{0.784}_3$	0.762_5	0.823_9
4	ECNU (*)	$\mathbf{0.768}_4$	0.770_4	0.843_5
5	INSIGHT-1	$\mathbf{0.767}_5$	0.786_3	0.864_2
6	PUT	$\mathbf{0.763}_6$	0.732_8	0.794_{14}
7	UNIMELB	$\mathbf{0.758}_7$	0.788_2	0.870_1
8	TwiSE	$\mathbf{0.756}_8$	0.752_6	0.826_8
9	GTI	$\mathbf{0.736}_9$	0.731_9	0.811_{11}
10	Finki	$\mathbf{0.720}_{10}$	0.748_7	0.848_4
11	pkudblab	$\mathbf{0.689}_{11}$	0.716_{11}	0.832_7
12	CUFE	$\mathbf{0.679}_{12}$	0.708_{12}	0.834_6
13	ISTI-CNR	$\mathbf{0.671}_{13}$	0.690_{13}	0.811_{11}
14	SwissCheese	$\mathbf{0.648}_{14}$	0.674_{14}	0.820_{10}
15	SentimentalITists	$\mathbf{0.624}_{15}$	0.643_{15}	0.802_{13}
16	PotTS	$\mathbf{0.618}_{16}$	0.610_{17}	0.712_{18}
17	OPAL	$\mathbf{0.616}_{17}$	0.633_{16}	0.792_{15}
18	WR	$\mathbf{0.522}_{18}$	0.502_{18}	0.577_{19}
19	VCU-TSA	$\mathbf{0.502}_{19}$	0.448_{19}	0.775_{16}
	Baseline	**0.500**	0.438	0.778

Table 11: Results for Subtask B "Tweet classification according to a two-point scale" on the Tweet 2016 dataset. The systems are ordered by their ρ^{PN} score (higher is better). The meaning of "(*)" is as in Table 8.

Out of the 10 top-ranked participating teams, 5 teams (Tweester$_1$, LYS$_2$, INSIGHT-1$_5$, UNIMELB$_7$, Finki$_{10}$) used convolutional neural networks; 3 teams (thecerealkiller$_3$, UNIMELB$_7$, Finki$_{10}$) submitted systems using recurrent neural networks; and 7 teams (Tweester$_1$, LYS$_2$, INSIGHT-1$_5$, UNIMELB$_7$, Finki$_{10}$) incorporated in their participating systems either general-purpose or task-specific word embeddings (generated via toolkits such as GloVe or word2vec).

Conversely, the use of classifiers such as support vector machines, which were dominant until a few years ago, seems to have decreased, with only one team (TwiSE$_8$) in the top 10 using them.

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

Table 12 ranks the 11 teams who participated in Subtask C "Tweet classification according to a five-point scale" in terms of the official measure MAE^M; we also show MAE^μ (see Equation 6). We also report the result of a baseline system that assigns to each tweet the middle class (i.e., NEUTRAL); for ordinal classification evaluated via MAE^M, this is the majority-class classifier for (binary or SLMC) classification evaluated via vanilla accuracy, i.e., this is (Baccianella et al., 2009) the "smartest" among the trivial policies that attempt to maximize MAE^M.

#	System	MAE^M	MAE^μ
1	TwiSE	$\mathbf{0.719}_1$	0.632_5
2	ECNU (*)	$\mathbf{0.806}_2$	0.726_8
3	PUT	$\mathbf{0.860}_3$	0.773_9
4	LYS	$\mathbf{0.864}_4$	0.694_7
5	Finki	$\mathbf{0.869}_5$	0.672_6
6	INSIGHT-1	$\mathbf{1.006}_6$	0.607_3
7	ISTI-CNR	$\mathbf{1.074}_7$	0.580_1
8	YZU-NLP	$\mathbf{1.111}_8$	0.588_2
9	SentimentalITists	$\mathbf{1.148}_9$	0.625_4
10	PotTS	$\mathbf{1.237}_{10}$	0.860_{10}
11	pkudblab	$\mathbf{1.697}_{11}$	1.300_{11}
	Baseline	**1.200**	0.537

Table 12: Results for Subtask C "Tweet classification according to a five-point scale" on the Tweet 2016 dataset. The systems are ordered by their MAE^M score (lower is better). The meaning of "(*)" is as in Table 8.

The top-scoring team (TwiSE$_1$) used a single-label multi-class classifier to classify the tweets according to their overall polarity. In particular, they used logistic regression that minimizes the multinomial loss across the classes, with weights to cope with class imbalance. Note that they ignored the given topics altogether.

#	System	KLD	AE	RAE
1	Finki	$\mathbf{0.034}_{1}$	0.074_{1}	0.707_{3}
2	LYS	$\mathbf{0.053}_{2}$	0.099_{4}	0.844_{5}
	TwiSE	$\mathbf{0.053}_{2}$	0.101_{5}	0.864_{6}
4	INSIGHT-1	$\mathbf{0.054}_{4}$	0.085_{2}	0.423_{1}
5	GTI	$\mathbf{0.055}_{5}$	0.104_{6}	1.200_{10}
	QCRI	$\mathbf{0.055}_{5}$	0.095_{3}	0.864_{6}
7	NRU-HSE	$\mathbf{0.084}_{7}$	0.120_{8}	0.767_{4}
8	PotTS	$\mathbf{0.094}_{8}$	0.150_{12}	1.838_{12}
9	pkudblab	$\mathbf{0.099}_{9}$	0.109_{7}	0.947_{8}
10	ECNU (*)	$\mathbf{0.121}_{10}$	0.148_{11}	1.171_{9}
11	ISTI-CNR	$\mathbf{0.127}_{11}$	0.147_{9}	1.371_{11}
12	SwissCheese	$\mathbf{0.191}_{12}$	0.147_{9}	0.638_{2}
13	UDLAP	$\mathbf{0.261}_{13}$	0.274_{13}	2.973_{13}
14	HSENN	$\mathbf{0.399}_{14}$	0.336_{14}	3.930_{14}
	Baseline_1	**0.175**	0.184	2.110
	Baseline_2	**0.887**	0.242	1.155

Table 13: Results for Subtask D "Tweet quantification according to a two-point scale" on the Tweet 2016 dataset. The systems are ordered by their KLD score (lower is better). The meaning of "(*)" is as in Table 8.

Only 2 of the 11 participating teams tuned their systems to exploit the ordinal (as opposed to binary, or single-label multi-class) nature of this subtask. The two teams who did exploit the ordinal nature of the problem are PUT_3, which uses an ensemble of ordinal regression approaches, and ISTI-CNR_7, which uses a tree-based approach to ordinal regression. All other teams used general-purpose approaches for single-label multi-class classification, in many cases relying (as for Subtask B) on convolutional neural networks, recurrent neural networks, and word embeddings.

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

Table 13 ranks the 14 teams who participated in Subtask D "Tweet quantification according to a two-point scale" on the official measure KLD. Two other measures are reported, *absolute error* (AE):

$$AE(p, \hat{p}, \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |\hat{p}(c) - p(c)| \qquad (10)$$

and *relative absolute error* (RAE):

$$RAE(p, \hat{p}, \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|\hat{p}(c) - p(c)|}{p(c)} \qquad (11)$$

where the notation is the same as in Equation 7.

We also report the result of a "maximum likelihood" baseline system (dubbed Baseline_1). This system assigns to each test topic the distribution of the training tweets (the union of TRAIN, DEV, DEVTEST) across the classes. This is the "smartest" among the trivial policies that attempt to maximize KLD. We also report the result of a further (less smart) baseline system (dubbed Baseline_2), i.e., one that assigns a prevalence of 1 to the majority class (which happens to be the POSITIVE class) and a prevalence of 0 to the other class.

The top-scoring team (Finki_1) adopts an approach based on "classify and count", a classification-oriented (instead of quantification-oriented) approach, using recurrent and convolutional neural networks, and GloVe word embeddings.

Indeed, only 5 of the 14 participating teams tuned their systems to the fact that it deals with quantification (as opposed to classification). Among the teams who do rely on quantification-oriented approaches, teams LYS_2 and HSENN_{14} used an existing structured prediction method that directly optimizes KLD; teams QCRI_5 and ISTI-CNR_{11} use existing probabilistic quantification methods; team NRU-HSE_7 uses an existing iterative quantification method based on cost-sensitive learning. Interestingly, team TwiSE_2 uses a "classify and count" approach after comparing it with a quantification-oriented method (similar to the one used by teams LYS_2 and HSENN_{14}) on the development set, and concluding that the former works better than the latter. All other teams used "classify and count" approaches, mostly based on convolutional neural networks and word embeddings.

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

Table 14 lists the results obtained by the 10 participating teams on Subtask E "Tweet quantification according to a five-point scale". We also report the result of a "maximum likelihood" baseline system (dubbed Baseline_1), i.e., one that assigns to each test topic the same distribution, namely the distribution of the training tweets (the union of TRAIN, DEV, DEVTEST) across the classes; this is the "smartest" among the trivial policies (i.e., those that do not require any genuine work) that attempt to maximize EMD.

We further report the result of less smart baseline system (dubbed Baseline_2) – one that assigns a prevalence of 1 to the majority class (which coincides with the POSITIVE class) and a prevalence of 0 to all other classes.

#	System	EMD
1	QCRI	$\mathbf{0.243}_1$
2	Finki	$\mathbf{0.316}_2$
3	pkudblab	$\mathbf{0.331}_3$
4	NRU-HSE	$\mathbf{0.334}_4$
5	ECNU (*)	$\mathbf{0.341}_5$
6	ISTI-CNR	$\mathbf{0.358}_6$
7	LYS	$\mathbf{0.360}_7$
8	INSIGHT-1	$\mathbf{0.366}_8$
9	HSENN	$\mathbf{0.545}_9$
10	PotTS	$\mathbf{0.818}_{10}$
	Baseline_1	**0.474**
	Baseline_2	**0.734**

Table 14: Results for Subtask E "Tweet quantification according to a five-point scale" on the Tweet 2016 dataset. The systems are ordered by their EMD score (lower is better). The meaning of "(*)" is as in Table 8.

Only 3 of the 10 participants tuned their systems to the specific characteristics of this subtask, i.e., to the fact that it deals with quantification (as opposed to classification) *and* to the fact that it has an ordinal (as opposed to binary) nature.

In particular, the top-scoring team (QCRI_1) used a novel algorithm explicitly designed for ordinal quantification, that leverages an ordinal hierarchy of binary probabilistic quantifiers.

Team NRU-HSE_4 uses an existing quantification approach based on cost-sensitive learning, and adapted it to the ordinal case.

Team ISTI-CNR_6 instead used a novel adaptation to quantification of a tree-based approach to ordinal regression.

Teams LYS_7 and HSENN_9 also used an existing quantification approach, but did not exploit the ordinal nature of the problem.

The other teams mostly used approaches based on "classify and count" (see Section 5.4), and viewed the problem as single-label multi-class (instead of ordinal) classification; some of these teams (notably, team Finki_2) obtained very good results, which testifies to the quality of the (general-purpose) features and learning algorithm they used.

6 Conclusion and Future Work

We described SemEval-2016 Task 4 "Sentiment Analysis in Twitter", which included five subtasks including three that represent a significant departure from previous editions. The three new subtasks focused, individually or in combination, on two variants of the basic "sentiment classification in Twitter" task that had not been previously explored within SemEval. The first variant adopts a five-point scale, which confers an *ordinal* character to the classification task. The second variant focuses on the correct estimation of the prevalence of each class of interest, a task which has been called *quantification* in the supervised learning literature. In contrast, previous years' subtasks have focused on the correct labeling of individual tweets. As in previous years (2013–2015), the 2016 task was very popular and attracted a total of 43 teams.

A general trend that emerges from SemEval-2016 Task 4 is that most teams who were ranked at the top in the various subtasks used deep learning, including convolutional NNs, recurrent NNs, and (general-purpose or task-specific) word embeddings. In many cases, the use of these techniques allowed the teams using them to obtain good scores even without tuning their system to the specifics of the subtask at hand, e.g., even without exploiting the ordinal nature of the subtask – for Subtasks C and E – or the quantification-related nature of the subtask – for Subtasks D and E. Conversely, several teams that have indeed tuned their system to the specifics of the subtask at hand, but have not used deep learning techniques, have performed less satisfactorily. This is a further confirmation of the power of deep learning techniques for tweet sentiment analysis.

Concerning Subtasks D and E, if quantification-based subtasks are proposed again, we think it might be a good idea to generate, for each test topic t_i, multiple "artificial" test topics $t_i^1, t_i^2, ...$, where class prevalences are altered with respect to the ones of t_i by means of selectively removing from t_i tweets belonging to a certain class. In this way, the evaluation can take into consideration (*i*) class prevalences in the test set and (*ii*) levels of distribution drift (i.e., of the divergence of the test distribution from the training distribution) that are not present in the "naturally occurring" data.

By varying the amount of removed tweets at will, one may obtain *many* test topics, thus augmenting the magnitude of the experimentation at will while at the same time keeping constant the amount of manual annotation needed.

In terms of possible follow-ups of this task, it might be interesting to have a subtask whose goal is to distinguish tweets that are NEUTRAL about the topic (i.e., do not express any opinion about the topic) from tweets that express a FAIR opinion (i.e., lukewarm, intermediate between POSITIVE and NEGATIVE) about the topic.

Another possibility is to have a multi-lingual tweet sentiment classification subtask, where training examples are provided for the same topic for two languages (e.g., English and Arabic), and where participants can improve their performance on one language by leveraging the training examples for the other language via transfer learning. Alternatively, it might be interesting to include a cross-lingual tweet sentiment classification subtask, where training examples are provided for a given language (e.g., English) but not for the other (e.g., Arabic); the second language could be also a surprise language, which could be announced at the last moment.

References

[Abdelwahab and Elmaghraby2016] Omar Abdelwahab and Adel Elmaghraby. 2016. UofL at SemEval-2016 Task 4: Multi domain word2vec for Twitter sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Astudillo and Amir2016] Ramón Astudillo and Silvio Amir. 2016. INESC-ID at SemEval-2016 Task 4: Reducing the problem of out-of-embedding words. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Attardi and Sartiano2016] Giuseppe Attardi and Daniele Sartiano. 2016. UniPI at SemEval-2016 Task 4: Convolutional neural networks for sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Baccianella et al.2009] 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.

[Baccianella et al.2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, MT.

[Balahur2016] Alexandra Balahur. 2016. OPAL at SemEval-2016 Task 4: the Challenge of Porting a Sentiment Analysis System to the "Real" World. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Balikas and Amini2016] Georgios Balikas and Massih-Reza Amini. 2016. TwiSE at SemEval-2016 Task 4: Twitter sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Bollen et al.2011] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

[Borge-Holthoefer et al.2015] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2015)*, pages 700–711, Vancouver, CA.

[Briones and Amarasinghe2016] Gerard Briones and Kasun Amarasinghe. 2016. VCU-TSA at SemEval-2016 Task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Burton and Soboleva2011] S. Burton and A. Soboleva. 2011. Interactive or reactive? Marketing with Twitter. *Journal of Consumer Marketing*, 28(7):491–499.

[Castillo et al.2016] Esteban Castillo, Ofelia Cervantes, Darnes Vilariño, and David Báez. 2016. UDLAP at SemEval-2016 Task 4: Sentiment quantification using a graph based representation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Ciubotariu et al.2016] Calin-Cristian Ciubotariu, Marius-Valentin Hrisca, Mihail Gliga, Diana Darabana, Diana Trandabat, and Adrian Iftene. 2016. Minions at SemEval-2016 Task 4: Or how to boost a student's self esteem. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Cozza and Petrocchi2016] Vittoria Cozza and Marinella Petrocchi. 2016. mib at SemEval-2016 Task 4: Exploiting lexicon based features for sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Da San Martino et al.2016] Giovanni Da San Martino, Wei Gao, and Fabrizio Sebastiani. 2016. QCRI at SemEval-2016 Task 4: Probabilistic methods for binary and ordinal quantification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US. Forthcoming.

[Deriu et al.2016] Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. SwissCheese at SemEval-2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Dodds et al.2011] Peter S. Dodds, Kameron D. Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, 6(12).

[Du and Zhang2016] Steven Du and Xi Zhang. 2016. Aicyber at SemEval-2016 Task 4: i-vector based sentence representation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Esuli and Sebastiani2010] Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiment quantification. *IEEE Intelligent Systems*, 25(4):72–75.

[Esuli and Sebastiani2015] Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery and Data*, 9(4):Article 27.

[Esuli2016] Andrea Esuli. 2016. ISTI-CNR at SemEval-2016 Task 4: Quantification on an ordinal scale. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Florean et al.2016] Cosmin Florean, Oana Bejenaru, Eduard Apostol, Octavian Ciobanu, Adrian Iftene, and Diana Trandabat. 2016. SentimentalITists at SemEval-2016 Task 4: Building a Twitter sentiment analyzer in your backyard. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Forman2005] 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.

[Forman2008] George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206.

[Friedrichs2016] Jasper Friedrichs. 2016. IIP at SemEval-2016 Task 4: Prioritizing classes in ensemble classification for sentiment analysis of tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Gao and Oates2016] Hang Gao and Tim Oates. 2016. MDSENT at SemEval-2016 Task 4: Supervised system for message polarity classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Giorgis et al.2016] Stavros Giorgis, Apostolos Rousas, John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. aueb.twitter.sentiment at SemEval-2016 Task 4: A weighted ensemble of SVMs for Twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Gomez et al.2016] Helena Gomez, Darnes Vilariño, Grigori Sidorov, and David Pinto Avendaño. 2016. CICBUAPnlp at SemEval-2016 Task 4: Discovering Twitter polarity using enhanced embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Hamdan2016] Hussam Hamdan. 2016. SentiSys at SemEval-2016 Task 4: Feature-based system for sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[He et al.2016] Yunchao He, Liang-Chih Yu, Chin-Sheng Yang, K. Robert Lai, and Weiyi Liu. 2016. YZU-NLP at SemEval-2016 Task 4: Ordinal sentiment classification using a recurrent convolutional network. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Jahren et al.2016] Brage Ekroll Jahren, Valerij Fredriksen, Björn Gambäck, and Lars Bungum. 2016. NTNUSentEval at SemEval-2016 Task 4: Combining general classifiers for fast Twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Juncal-Martínez et al.2016] Jonathan Juncal-Martínez, Tamara Álvarez-López, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. 2016. GTI at SemEval-2016 Task 4: Training a naive Bayes classifier using features of an unsupervised system. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Karpov et al.2016] Nikolay Karpov, Alexander Porshnev, and Kirill Rudakov. 2016. NRU-HSE at SemEval-2016 Task 4: The open quantification library with two iterative methods. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Kaya et al.2013] Mesut Kaya, Guven Fidan, and Ismail Hakki Toroslu. 2013. Transfer learning using Twitter data for improving sentiment classification of Turkish political news. In *Proceedings of the 28th In-*

ternational Symposium on Computer and Information Sciences (ISCIS 2013), pages 139–148, Paris, FR.

[Lango et al.2016] Mateusz Lango, Dariusz Brzezinski, and Jerzy Stefanowski. 2016. PUT at SemEval-2016 Task 4: The ABC of Twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Marchetti-Bowick and Chambers2012] Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 603–612, Avignon, FR.

[Martínez-Cámara et al.2014] Eugenio Martínez-Cámara, Maria Teresa Martín-Valdivia, Luis Alfonso Ureña López, and Arturo Montejo Ráez. 2014. Sentiment analysis in Twitter. *Natural Language Engineering*, 20(1):1–28.

[Mejova et al.2015] Yelena Mejova, Ingmar Weber, and Michael W. Macy, editors. 2015. *Twitter: A Digital Socioscope*. Cambridge University Press, Cambridge, UK.

[Mikolov et al.2013] Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, pages 746–751, Atlanta, US.

[Morant et al.2016] Victor Martinez Morant, Lluís-F. Hurtado, and Ferran Pla. 2016. DSIC-ELIRF at SemEval-2016 Task 4: Message polarity classification in Twitter using a support vector machine approach. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Nabil et al.2016] Mahmoud Nabil, Amir Atyia, and Mohamed Aly. 2016. CUFE at SemEval-2016 Task 4: A gated recurrent model for sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Nakov et al.2013] 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.

[Nakov et al.2016a] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016a. Evaluation measures for the SemEval-2016 Task 4 "Sentiment analysis in Twitter". Available from http://alt.qcri.org/semeval2016/task4/.

[Nakov et al.2016b] Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M. Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2016b. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

[O'Connor et al.2010] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, Washington, US.

[Palogiannidi et al.2016] Elisavet Palogiannidi, Athanasia Kolovou, Fenia Christopoulou, Filippos Kokkinos, Elias Iosif, Nikolaos Malandrakis, Haris Papageorgiou, Shrikanth Narayanan, and Alexandros Potamianos. 2016. Tweester at SemEval-2016 Task 4: Sentiment analysis in Twitter using semantic-affective model adaptation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543, Doha, QA.

[Qureshi et al.2013] Muhammad A. Qureshi, Colm O'Riordan, and Gabriella Pasi. 2013. Clustering with error estimation for monitoring reputation of companies on Twitter. In *Proceedings of the 9th Asia Information Retrieval Societies Conference (AIRS 2013)*, pages 170–180, Singapore, SN.

[Räbiger et al.2016] Stefan Räbiger, Mishal Kazmi, Yücel Saygın, Peter Schüller, and Myra Spiliopoulou. 2016. SteM at SemEval-2016 Task 4: Applying active learning to improve sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Ritter et al.2011] Alan Ritter, Sam Clark, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1524–1534, Edinburgh, UK.

[Rosenthal et al.2014] 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.

[Rosenthal et al.2015] 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.

[Rouvier and Favre2016] Mickael Rouvier and Benoit Favre. 2016. SENSEI-LIF at SemEval-2016 Task 4: Polarity embedding fusion for robust sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Rubner et al.2000] 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.

[Ruder et al.2016] Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. INSIGHT-1 at SemEval-2016 Task 4: Convolutional Neural Networks for Sentiment Classification and Quantification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Rüschendorf2001] Ludger Rüschendorf. 2001. Wasserstein metric. In Michiel Hazewinkel, editor, *Encyclopaedia of Mathematics*. Kluwer Academic Publishers, Dordrecht, NL.

[Sarker2016] Abeed Sarker. 2016. DIEGOLab16 at SemEval-2016 Task 4: Sentiment analysis in Twitter using centroids, clusters, and sentiment lexicons. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Sebastiani2015] Fabrizio Sebastiani. 2015. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 5th ACM International Conference on the Theory of Information Retrieval (ICTIR 2015)*, pages 11–20, Northampton, US.

[Sidarenka2016] Uladzimir Sidarenka. 2016. PotTS at SemEval-2016 Task 4: Sentiment analysis of Twitter using character-level convolutional neural networks. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Stojanovski et al.2016] Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. 2016. Finki at SemEval-2016 Task 4: Deep learning architecture for Twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Vilares et al.2016] David Vilares, Yerai Doval, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2016. LYS at SemEval-2016 Task 4: Exploiting neural activation values for Twitter sentiment classification and quantification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Xu et al.2016] Steven Xu, HuiZhi Liang, and Tim Baldwin. 2016. UNIMELB at SemEval-2016 Task 4: An ensemble of neural networks and a word2vec based model for sentiment classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Yadav2016] Vikrant Yadav. 2016. thecerealkiller at SemEval-2016 Task 4: Deep learning based system for classifying sentiment of tweets on two point scale. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Zhang et al.2016] Zhengchen Zhang, Chen Zhang, Dongyan Huang, Wu Fuxiang, Weisi Lin, and Minghui Dong. 2016. I2RNTU at SemEval-2016 Task 4: Classifier fusion for polarity classification in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

[Zhou et al.2016] Yunxiao Zhou, Zhihua Zhang, and Man Lan. 2016. ECNU at SemEval-2016 Task 4: An empirical investigation of traditional NLP features and word embedding features for sentence-level and topic-level sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, US.

Subtasks	Team ID	Affiliation	Nation	Paper
A	Aicyber	Aicyber.com	Singapore; China	(Du and Zhang, 2016)
A	aueb.twitter.sentiment	Department of Informatics, Athens University of Economics and Business	Greece	(Giorgis et al., 2016)
A	CICBUAPnlp	Instituto Politècnico Nacional Benemèrita Universidad Autonoma de Puebla	Mexico	(Gomez et al., 2016)
A B	CUFE	Cairo University	Egypt	(Nabil et al., 2016)
A	DIEGOLab16	Arizona State University	USA	(Sarker, 2016)
A	DSIC-ELIRF	Universitat Politècnica de València	Spain	(Morant et al., 2016)
A B C D E	ECNU	East China Normal University	China	(Zhou et al., 2016)
A	ELiRF	Universitat Politècnica de València	Spain	
B C D E	Finki	Saints Cyril and Methodius University, Skopje	Macedonia	(Stojanovski et al., 2016)
A B D	GTI	AtlantTIC Centre, University of Vigo	Spain	(Juncal-Martínez et al., 2016)
D E	HSENN	National Research University Higher School of Economics	Russia	
A	I2RNTU	Institute for Infocomm Research, A*STAR School of Computer Engineering, Nanyang Technological University	Singapore	(Zhang et al., 2016)
A	IIP	Infosys Limited	India	(Friedrichs, 2016)
A	INESC-ID	INESC-ID, Lisboa Instituto Superior Técnico, Universidade de Lisboa	Portugal	(Astudillo and Amir, 2016)
A B C D E	INSIGHT-1	INSIGHT Research Centre, National University of Ireland, Galway AYLIEN Inc.	Ireland	(Ruder et al., 2016)
A B C D E	LYS	Universidade da Coruña Universidade de Vigo	Spain	(Vilares et al., 2016)
A	MDSENT	University of Maryland Baltimore County	USA	(Gao and Oates, 2016)
A	mib	Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche	Italy	(Cozza and Petrocchi, 2016)
A	Minions	University of Iasi	Romania	(Ciubotariu et al., 2016)
A B C D E	ISTI-CNR	Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche	Italy	(Esuli, 2016)
D E	NRU-HSE	National Research University Higher School of Economics	Russia	(Karpov et al., 2016)
A	NTNUSentEval	Norwegian University of Science and Technology	Norway	(Jahren et al., 2016)
A B	OPAL	European Commission Joint Research Centre	Italy	(Balahur, 2016)
B C D E	pkudblab	Peking University	China	
B C D E	PotTS	University of Potsdam Retresco GmbH	Germany	(Sidarenka, 2016)
A B C	PUT	Poznan University of Technology	Poland	(Lango et al., 2016)
D E	QCRI (**)	Qatar Computing Research Institute	Qatar	(Da San Martino et al., 2016)
A	SENSEI-LIF	Aix-Marseille University - CNRS - LIF	France	(Rouvier and Favre, 2016)
A B C	SentimentallTists	University of Iasi	Romania	(Florean et al., 2016)
A	SentiSys	Aix-Marseille University	France	(Hamdan, 2016)
A	SteM	Sabanci University Marmara University Otto-von-Guericke University Magdeburg	Turkey Germany	(Räbiger et al., 2016)
A B D	SwissCheese	ETH Zürich	Switzerland	(Deriu et al., 2016)
B	thecerealkiller	Amazon.in	India	(Yadav, 2016)
A	THUIR	Tsinghua University	China	
A B	Tweester	School of ECE, National Technical University of Athens School of ECE, Technical University of Crete Department of Informatics, University of Athens Signal Analysis and Interpretation Laboratory (SAIL) Institute for Language & Speech Processing - ILSP	Greece	(Palogiannidi et al., 2016)
A B C D	TwiSE	University of Grenoble-Alpes	France	(Balikas and Amini, 2016)
D	UDLAP	Universidad de las Américas Puebla (UDLAP)	Mexico	(Castillo et al., 2016)
A B	UNIMELB	University of Melbourne	Australia	(Xu et al., 2016)
A	UniPI	Università di Pisa	Italy	(Attardi and Sartiano, 2016)
A	UofL	University of Louisville	USA	(Abdelwahab and Elmaghraby, 2016)
A B	VCU-TSA	Virginia Commonwealth University	USA	(Briones and Amarasinghe, 2016)
A B	WR	WR	Hong Kong	
C	YZU-NLP	Yuan Ze University, Taoyuan Yunnan University, Kunming	Taiwan China	(He et al., 2016)
34 19 11 14 10	**Total**			

Table 15: Participating teams (Column 2), their affiliation (Column 3) and nationality (Column 4), the subtasks they have participated in (Column 1), and the paper they have contributed (Column 5). Teams whose "Affiliation" column is typeset on more that one row include researchers with different affiliations. Teams marked with a (**) include some of the SemEval 2016 Task 4 organizers. An empty entry for the "Paper" column indicates that the team have not contributed a system description paper.