# SemEval-2017 Task 4: Sentiment Analysis in Twitter

**Sara Rosenthal**[♣], **Noura Farra**[♢], **Preslav Nakov**[♡]
[♡]Qatar Computing Research Institute, Hamad bin Khalifa University, Qatar
[♢]Department of Computer Science, Columbia University
[♣]IBM Research, USA

## Abstract

This paper describes the fifth year of the *Sentiment Analysis in Twitter task*. SemEval-2017 Task 4 continues with a rerun of the subtasks of SemEval-2016 Task 4, which include identifying the *overall sentiment* of the tweet, *sentiment towards a topic* with classification on a two-point and on a five-point ordinal scale, and *quantification* of the distribution of sentiment towards a topic across a number of tweets: again on a two-point and on a five-point ordinal scale. Compared to 2016, we made two changes: (*i*) we introduced a new language, Arabic, for all subtasks, and (*ii*) we made available information from the profiles of the Twitter users who posted the target tweets. The task continues to be very popular, with a total of 48 teams participating this year.

## 1 Introduction

The identification of sentiment in text is an important field of study, with social media platforms such as Twitter garnering the interest of researchers in language processing as well as in political and social sciences. The task usually involves detecting whether a piece of text expresses a POSITIVE, a NEGATIVE, or a NEUTRAL sentiment; the sentiment can be general or about a specific topic, e.g., a person, a product, or an event.

The *Sentiment Analysis in Twitter* task has been run yearly at SemEval since 2013 (Nakov et al., 2013; Rosenthal et al., 2014; Nakov et al., 2016b), with the 2015 task introducing sentiment towards a topic (Rosenthal et al., 2015) and the 2016 task introducing tweet quantification and five-point ordinal classification (Nakov et al., 2016a).

SemEval is the International Workshop on Semantic Evaluation, formerly SensEval. It is an ongoing series of evaluations of computational semantic analysis systems, organized under the umbrella of SIGLEX, the Special Interest Group on the Lexicon of the Association for Computational Linguistics. Other related tasks at SemEval have explored sentiment analysis of product review and their aspects (Pontiki et al., 2014, 2015, 2016), sentiment analysis of figurative language on Twitter (Ghosh et al., 2015), implicit event polarity (Russo et al., 2015), detecting stance in tweets (Mohammad et al., 2016a), out-of-context sentiment intensity of words and phrases (Kiritchenko et al., 2016), and emotion detection (Strapparava and Mihalcea, 2007). Some of these tasks featured languages other than English, such as Arabic (Pontiki et al., 2016; Mohammad et al., 2016a); however, they did not target tweets, nor did they focus on sentiment towards a topic.

This year, we performed a re-run of the subtasks in SemEval-2016 Task 4, which, in addition to the overall sentiment of a tweet, featured classification, ordinal regression, and quantification with respect to a topic. Furthermore, we introduced a new language, Arabic. Finally, we made available to the participants demographic information about the users who posted the tweets, which we extracted from the respective public profiles.

**Ordinal Classification** As last year, SemEval-2017 Task 4 includes sentiment analysis on a five-point scale {HIGHLYPOSITIVE, POSITIVE, NEUTRAL, NEGATIVE, HIGHLYNEGATIVE}, which is in line with product ratings occurring in the corporate world, e.g., Amazon, TripAdvisor, and Yelp. In machine learning terms, moving from a categorical two-point scale to an ordered five-point scale means moving from binary to *ordinal classification* (aka *ordinal regression*).



*Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 502–518,
Vancouver, Canada, August 3 - 4, 2017. 

**Tweet Quantification** SemEval-2017 Task 4 includes *tweet quantification* tasks along with tweet classification tasks, also on 2-point and 5-point scales. While the tweet classification task is concerned with whether a specific tweet expresses a given sentiment towards a topic, the tweet quantification task looks at estimating the *distribution* of tweets about a given topic across the different sentiment classes. Most (if not all) tweet sentiment classification studies 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), study Twitter with an interest in aggregate statistics about sentiment and are *not* interested in the sentiment expressed in individual tweets. We should also note that quantification is not a mere byproduct of classification, as it can be addressed using different approaches and it also needs different evaluation measures (Forman, 2008; Esuli and Sebastiani, 2015).

**Analysis in Arabic** This year, we added a new language, Arabic, in order to encourage participants to experiment with multilingual and cross-lingual approaches for sentiment analysis. Our objective was to expand the Twitter sentiment analysis resources available to the research community, not only for general multilingual sentiment analysis, but also for multilingual sentiment analysis *towards a topic*, which is still a largely unexplored research direction for many languages and in particular for morphologically complex languages such as Arabic.

Arabic has become an emergent language for sentiment analysis, especially as more resources and tools for it have recently become available. It is also both interesting and challenging due to its rich morphology and abundance of dialectal use in Twitter. Early Arabic studies focused on sentiment analysis in newswire (Abdul-Mageed and Diab, 2011; Elarnaoty et al., 2012), but recently there has been a lot more work on social media, especially Twitter (Mourad and Darwish, 2013; Abdul-Mageed et al., 2014; Refaee and Rieser, 2014; Salameh et al., 2015), where the challenges of sentiment analysis are compounded by the presence of multiple dialects and orthographical variants, which are frequently used in conjunction with the formal written language. Some work studied the utility of machine translation for sentiment analysis of Arabic texts (Salameh et al., 2015; Mohammad et al., 2016b; Refaee and Rieser, 2015), identification of sentiment holders (Elarnaoty et al., 2012), and sentiment targets (Al-Smadi et al., 2015; Farra et al., 2015; Farra and McKeown, 2017). We believe that the development of a standard Arabic Twitter dataset for sentiment, and particularly with respect to topics, will encourage further research in this regard.

**User Information** Demographic information in Twitter has been studied and analyzed using network analysis and natural language processing (NLP) techniques (Mislove et al., 2011; Nguyen et al., 2013; Rosenthal and McKeown, 2016). Recent work has shown that user information and information from the network can help sentiment analysis in other corpora (Hovy, 2015) and in Twitter (Volkova et al., 2013; Yang and Eisenstein, 2015). Thus, this year we encouraged participants to use information from the public profiles of Twitter users such as demographics (e.g., age, location) as well as information from the rest of the social network (e.g., sentiment of the tweets of friends), with the goal of analyzing the impact of this information on improving sentiment analysis.

The rest of this paper is organized as follows. Section 2 presents in more detail the five subtasks of SemEval-2017 Task 4. Section 3 describes the English and the Arabic datasets and how we created them. Section 4 introduces and motivates the evaluation measures for each subtask. Section 5 presents the results of the evaluation and discusses the techniques and the tools that the participants used. Finally, Section 6 concludes and points to some possible directions for future work.

## 2 Task Definition

SemEval-2017 Task 4 consists of five subtasks, each offered for both Arabic and English:

1. **Subtask A:** Given a tweet, decide whether it expresses POSITIVE, NEGATIVE or NEUTRAL sentiment.

2. **Subtask B:** Given a tweet and a topic, classify the sentiment conveyed towards that topic on a two-point scale: POSITIVE vs. NEGATIVE.

3. **Subtask C:** Given a tweet and a topic, classify the sentiment conveyed in the tweet towards that topic on a five-point scale: STRONGLYPOSITIVE, WEAKLYPOSITIVE, NEUTRAL, WEAKLYNEGATIVE, and STRONGLYNEGATIVE.

4. **Subtask D:** Given a set of tweets about a topic, estimate the *distribution* of tweets across the POSITIVE and NEGATIVE classes.

5. **Subtask E:** Given a set of tweets about a topic, estimate the *distribution* of tweets across the five classes: STRONGLYPOSITIVE, WEAKLYPOSITIVE, NEUTRAL, WEAKLYNEGATIVE, and STRONGLYNEGATIVE.

| **Languages: English and Arabic** | | | |
|---|---|---|---|
| | **Goal** | **Granularity** | **Topic** |
| A | Classification | 3-point | No |
| B | Classification | 2-point | Yes |
| C | Classification | 5-point | Yes |
| D | Quantification | 2-point | Yes |
| E | Quantification | 5-point | Yes |

Table 1: Summary of the subtasks.

Each subtask is run for both English and Arabic. Subtask A has been run in all previous editions of the task and continues to be the most popular one (see section 5.) Subtasks B-E have all been run at SemEval-2016 Task 4 (Nakov et al., 2016a), with variants running in 2015 (Rosenthal et al., 2015). Table 1 shows a summary of the subtasks.

## 3 Datasets

Our datasets consist of tweets annotated for sentiment on a 2-point, 3-point, and 5-point scales. We made available to participants all the data from previous years (Nakov et al., 2016a) for the English training sets, and we collected new training data for Arabic, as well as new test sets for both English and Arabic. The annotation scheme remained the same as last year (Nakov et al., 2016a), with the key new contribution being to apply the task and instructions to Arabic as well as providing a script to download basic user information. All annotations were performed on CrowdFlower. Note that we release all our datasets to the research community to be used freely beyond SemEval.

### 3.1 Tweet Collection

We chose English and Arabic topics based on popular current events that were trending on Twitter, both internationally and in specific Arabic-speaking countries, using local and global Twitter trends.[1] The topics included a range of named entities (e.g., *Donald Trump*, *iPhone*), geopolitical entities (e.g., *Aleppo*, *Palestine*), and other entities (e.g., *Syrian refugees*, *Dakota Access Pipeline*, *Western media*, *gun control*, and *vegetarianism*). We then used the Twitter API to download tweets, along with corresponding user information, containing mentions of these topics in the specified language. We intentionally chose to use some overlapping topics between the two languages in order to encourage cross-language approaches.

We automatically filtered the tweets for duplicates and we removed those for which the bag-of-words cosine similarity exceeded 0.6. We then retained only the topics for which at least 100 tweets remained. The training tweets for Arabic were collected over the period of September-November 2016 and all test tweets were collected over the period of December 2016-January 2017.

For both English and Arabic, the topics for the test dataset were different from those in the training and in the development datasets.

### 3.2 Annotation using CrowdFlower

We used CrowdFlower to annotate the new training and testing tweets. The annotators were asked to indicate the overall polarity of the tweet (on a five-point scale) as well as the polarity of the tweet towards the given target topic (again, on a five-point scale), as described in (Nakov et al., 2016a). We also provided additional examples, some of which are shown in Tables 2 and 3. In particular, we stressed that topic-level positive or negative sentiment needed to express an opinion about the topic itself rather than about a positive or a negative event occurring in the context of the topic (see for example, the third row of Table 3).

Each tweet was annotated by at least five people, and we created many hidden tests for quality control, which we used to reject annotations by contributors who missed a large number of the hidden tests. We also created pilot runs, which helped us adjust the annotation instructions until we found, based on manual inspection, the quality of the annotated tweets to be satisfactory.

[1] https://trends24.in/

| Tweet | Overall Sentiment | Topic-level Sentiment |
|---|---|---|
| Who are you tomorrow? Will you make me smile or just bring me sorrow? #HottieOfTheWeek Demi Lovato | NEUTRAL | Demi Lovato: POSITIVE |
| Saturday without Leeds United is like Sunday dinner it doesn't feel normal at all (Ryan) | WEAKLYNEGATIVE | Leeds United: HIGHLYPOSITIVE |
| Apple releases a new update of its OS | NEUTRAL | Apple: NEUTRAL |

Table 2: Some English example annotations that we provided to the annotators.

| Tweet | Overall Sentiment | Topic-level Sentiment |
|---|---|---|
| أبل تطلق النسخة التجريبية الرابعة لنظام التشغيل<br>*Apple releases a fourth beta of its OS* | NEUTRAL | أبل *Apple*: NEUTRAL |
| المايسترو ... الاسطورة روجر فدرر ملك اللّعب الخلفي من اجمل لقطاته<br>*The maestro ... the legend Roger Federer king of the backhand game one of his best shots* | HIGHLYPOSITIVE | فدرر *Federer*: HIGHLYPOSITIVE |
| اللّاجئون يواجهون الصعوبات<br>*Refugees are facing difficulties* | WEAKLYNEGATIVE | اللّاجئون *Refugees* : NEUTRAL |

Table 3: Some Arabic example annotations that we provided to the annotators.

For Arabic, the contributors tended to annotate somewhat conservatively, and thus a very small number of HIGHLYPOSITIVE and HIGHLYNEGATIVE annotations were consolidated, despite us having provided examples of such annotations.

### 3.3 Consolidating the Annotations

As the annotations are on 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 the extreme values $-2$ and 2, and also to prefer 0, we did not use rounding at $\pm 0.5$ and $\pm 1.5$, but at $\pm 0.4$ and $\pm 1.4$ instead. Finally, note that the values $-2$, $-1$, 0, 1, 2 are to be interpreted as STRONGLYNEGATIVE, WEAKLYNEGATIVE, NEUTRAL, WEAKLYPOSITIVE, and STRONGLYPOSITIVE, respectively.

### 3.4 Data Statistics

The English training and development data this year consisted of the data from all previous editions of this task (Nakov et al., 2013; Rosenthal et al., 2014, 2015; Nakov et al., 2016b). Unlike in previous years, we did not set aside data to assess progress compared to prior years. Therefore, we allowed all data to be used for training and development.

For evaluation, we used the newly-created data described in the previous subsection. Tables 4 and 5 show the statistics for the English and Arabic data. For English, we only show the aggregate statistics for the training data; the breakdown from prior years can be found in (Nakov et al., 2016a). Note that the same tweets were annotated for multiple subtasks, so there is overlap between the tweets across the tasks. Duplicates may have occurred where the same tweet was extracted for multiple topics.

As Arabic is a new language this year, we created for it a default train-development split of the Arabic data for the participants to use if they wished to do so.

### 3.5 Data Distribution

As in previous years, we provided the participants with a script[2] to download the training tweets given IDs. In addition, this year we also included in the script the option to download basic user information for the author of each tweet: user id, follower count, status count, description, friend count, location, language, name, and time zone. To ensure a fair evaluation, the test set was provided via download and included the tweets as well as the basic user information provided by the download script. The training and the test data is available for downloaded on our task page.[3]

[2] https://github.com/seirasto/twitter_download

[3] http://alt.qcri.org/semeval2017/task4/index.php?id=data-and-tools

| Dataset | Subtask | Topics | Positive | | Neutral | Negative | | Total |
|---|---|---|---|---|---|---|---|---|
| | | | 2 | 1 | 0 | −1 | −2 | |
| **Train** | A | N/A | 19,902 | | 22,591 | 7,840 | | 50,333 |
| | B, D | 373 | 14,951 | | 1,544 | 4,013 | | 20,508 |
| | C, E | 200 | 1,020 | 12,922 | 12,993 | 3,398 | 299 | 30,632 |
| **Test** | A | N/A | 2375 | | 5,937 | 3,972 | | 12,284 |
| | B, D | 125 | 2,463 | | — | 3,722 | | 6,185 |
| | C, E | 125 | 131 | 2,332 | 6,194 | 3,545 | 177 | 12,379 |

Table 4: Statistics about the English training and testing datasets. The training data is the aggregate of all data from prior years, while the testing data is new.

| Dataset | Subtask | Topics | Positive | | Neutral | Negative | | Total |
|---|---|---|---|---|---|---|---|---|
| | | | 2 | 1 | 0 | −1 | −2 | |
| **Train** | A | N/A | 743 | | 1,470 | 1,142 | | 3,355 |
| | B, D | 34 | 885 | | — | 771 | | 1,656 |
| | C, E | 34 | 1 | 884 | 1699 | 770 | 1 | 3,355 |
| **Test** | A | N/A | 1,514 | | 2,364 | 2,222 | | 6,100 |
| | B, D | 61 | 1,561 | | — | 1,196 | | 2,757 |
| | C, E | 61 | 13 | 1,548 | 3,343 | 1,175 | 21 | 6,100 |

Table 5: Statistics about the newly collected Arabic training and testing datasets.

## 4 Evaluation Measures

This section describes the evaluation measures for our five subtasks. Note that 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.

### 4.1 Subtask A: Overall Sentiment of a Tweet

Our primary measure is AvgRec, or *average recall*, which is recall averaged across the POSITIVE (P), NEGATIVE (N), and NEUTRAL (U) classes. This measure has desirable theoretical properties (Sebastiani, 2015), and is also the one we use as primarily for Subtask B. It is computed as follows:

$$AvgRec = \frac{1}{3}(R^P + R^N + R^U) \quad (1)$$

where $R^P$, $R^N$ and $R^U$ refer to recall with respect to the POSITIVE, the NEGATIVE, and the NEUTRAL class, respectively. See (Nakov et al., 2016a) for more detail.

*AvgRec* 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 0.3333 is both (*i*) the value for a trivial classifier (i.e., one that assigns all tweets to the same class – be it POSITIVE, NEGATIVE, or NEUTRAL), and (*ii*) the expected value of a random classifier.

The advantage of *AvgRec* 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 *AvgRec* over $F_1$ is that *AvgRec* is invariant with respect to switching POSITIVE with NEGATIVE, while $F_1$ is not. See (Sebastiani, 2015) for more detail on *AvgRec*.

We further use two secondary measures: accuracy and $F_1^{PN}$. The latter was the primary evaluation measure for Subtask A in previous editions of the task. It is macro-average $F_1$, calculated over the POSITIVE and the NEGATIVE classes (note the exclusion of NEUTRAL). This year, we demoted $F_1^{PN}$ to a secondary evaluation measure. It is calculated as follows:

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

where $F_1^P$ and $F_1^N$ refer to $F_1$ with respect to the POSITIVE and the NEGATIVE class, respectively.

### 4.2 Subtask B: Topic-Based Classification on a 2-point Scale

As in 2016, our primary evaluation measure for subtask B is average recall, or AvgRec (note that there are only two classes for this subtask):

$$AvgRec = \frac{1}{2}(R^P + R^N) \quad (3)$$

We further use accuracy and $F_1$ as secondary measures for subtask B. Finally, as subtask B is topic-based, we computed each metric individually for each topic, and we then averaged the result across the topics to yield the final score.

### 4.3 Subtask C: Topic-based Classification on a 5-point Scale

Subtask C is an *ordinal classification* (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 $\mathcal{C}$.

We adopt an evaluation measure that takes the order of the five classes into account. For instance, misclassifying a HIGHLYNEGATIVE example as HIGHLYPOSITIVE is a bigger mistake than misclassifying it as NEGATIVE or as NEUTRAL.

As in SemEval-2016 Task 4, we use *macro-average mean absolute error* ($MAE^M$) as the main ordinal classification measure:

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

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

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

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 $MAE^{\mu}$ are equivalent.

$MAE^M$ is an extension of macro-average recall for ordinal regression; yet, it is a measure of error, and thus lower values are better. We also use $MAE^{\mu}$ as a secondary measure, in order to provide better consistency with Subtasks A and B. These measures are computed for each topic, and the results are then averaged across all topics to yield the final score. See (Baccianella et al., 2009) for more detail about $MAE^M$ and $MAE^{\mu}$.

### 4.4 Subtask D: Tweet Quantification on a 2-point Scale

Subtask D assumes a binary quantification setup, in which each tweet is classified as POSITIVE or NEGATIVE, and the distribution across classes must be estimated. The difference with binary classification is that errors of different polarity (e.g., a false positive and a false negative for the same class) can compensate for each other in quantification. Quantification is thus a more lenient task than classification, since a perfect classifier is also a perfect quantifier, but a perfect quantifier is not necessarily a perfect classifier.

For evaluating binary quantification, we keep the *Kullback-Leibler Divergence (KLD)* measure used in 2016 along with additive smoothing (Nakov et al., 2016a; Forman, 2005). KLD was proposed as a quantification measure in (Forman, 2005), and is defined as follows:

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

$KLD$ 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 $MAE^M$, $KLD$ 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 5 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.,

$$\begin{aligned} p^s(c_j) &= \frac{p(c_j) + \epsilon}{(\sum_{c_j \in \mathcal{C}} p(c_j)) + \epsilon \cdot |\mathcal{C}|} \\ &= \frac{p(c_j) + \epsilon}{1 + \epsilon \cdot |\mathcal{C}|} \end{aligned} \quad (6)$$

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 dataset.

The smoothed versions of $p(c_j)$ and $\hat{p}(c_j)$ are 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.

Like $MAE^M$, $KLD$ is a measure of error, which means that lower values are better. We further use two secondary error-based evaluation measures: *absolute error* (AE), and *relative absolute error* (RAE).

Again, the measures are computed individually for each topic, and the results are averaged across the topics to yield the final score.

### 4.5 Subtask E: Tweet Quantification on a 5-point Scale

Subtask E is an ordinal quantification task. As in binary quantification, the goal is to compute the distribution across classes, this time assuming a quantification setup.

Here 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 ordinal quantification 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 $i \in \{1, ..., (\mathcal{C} - 1)\}$, 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 $KLD$, $EMD$ 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 $EMD$.

As before, $EMD$ is computed individually for each topic, and the results are then averaged across all topics to yield the final score. For more detail on $EMD$, the reader is referred to (Esuli and Sebastiani, 2010) and to last year's task description paper (Nakov et al., 2016a).

## 5 Participants and Results

A total of 48 teams participated in SemEval-2017 Task 4 this year. As in previous years, the most popular subtask this year was Subtask A, with 38 teams participating in the English subtask A, and 8 teams participating in the Arabic subtask A. Overall, there were 46 teams who participated in some English subtask and 9 teams that participated in some Arabic subtask. There were 28 teams that participated in a subtask other than subtask A. Moreover, two teams (OMAM and ELiRF-UPV) participated in all English and in all Arabic subtasks. There were 9 teams that participated in the topic versions of the subtasks but not in subtask A, reflecting a growing interest among researchers in developing systems for topic-specific analysis.

### 5.1 Common Resources and Methods

In terms of methods, the use of deep learning stands out in particular, and we also see an increase over the last year. There were at least 20 teams who used deep learning and neural network methods such as CNN and LSTM networks. Supervised SVM and Liblinear were also very popular, with several participants combining SVM with neural network methods or SVM with dense word embedding features. Other teams used classifiers such as Maximum Entropy, Logistic Regression, Random Forest, Naïve Bayes classifier, and Conditional Random Fields.

Common software used included Python (with the sklearn and numpy libraries), Java, Tensorflow, Weka, NLTK, Keras, Theano, and Stanford CoreNLP. The most common external datasets used were sentiment140 as a lexicon, pre-trained word2vec embeddings. Many teams further gathered additional tweets using the Twitter API that were not annotated for sentiment. These were used for distant supervision, lexicon building, and word vector training.

In the following subsections, we present the results and the ranking for each subtask, and we highlight the best-performing systems for each subtask.

| # | System | *AvgRec* | $F_1^{PN}$ | *Acc* |
|---|---|---|---|---|
| **1** | DataStories | $\mathbf{0.681}_{1}$ | $0.677_{2}$ | $0.651_{5}$ |
| | BB_twtr | $\mathbf{0.681}_{1}$ | $0.685_{1}$ | $0.658_{3}$ |
| **3** | LIA | $\mathbf{0.676}_{3}$ | $0.674_{3}$ | $0.661_{2}$ |
| **4** | Senti17 | $\mathbf{0.674}_{4}$ | $0.665_{4}$ | $0.652_{4}$ |
| **5** | NNEMBs | $\mathbf{0.669}_{5}$ | $0.658_{5}$ | $0.664_{1}$ |
| **6** | Tweester | $\mathbf{0.659}_{6}$ | $0.648_{6}$ | $0.648_{6}$ |
| **7** | INGEOTEC | $\mathbf{0.649}_{7}$ | $0.645_{7}$ | $0.633_{11}$ |
| **8** | SiTAKA | $\mathbf{0.645}_{8}$ | $0.628_{9}$ | $0.643_{9}$ |
| **9** | TSA-INF | $\mathbf{0.643}_{9}$ | $0.620_{11}$ | $0.616_{17}$ |
| **10** | UCSC-NLP | $\mathbf{0.642}_{10}$ | $0.624_{10}$ | $0.565_{30}$ |
| **11** | HLP@UPENN | $\mathbf{0.637}_{11}$ | $0.632_{8}$ | $0.646_{8}$ |
| **12** | YNU-HPCC | $\mathbf{0.633}_{12}$ | $0.612_{15}$ | $0.647_{7}$ |
| | SentiME++ | $\mathbf{0.633}_{12}$ | $0.613_{13}$ | $0.601_{23}$ |
| **14** | ELiRF-UPV | $\mathbf{0.632}_{14}$ | $0.619_{12}$ | $0.599_{24}$ |
| **15** | ECNU | $\mathbf{0.628}_{15}$ | $0.613_{13}$ | $0.630_{12}$ |
| **16** | TakeLab | $\mathbf{0.627}_{16}$ | $0.607_{16}$ | $0.628_{14}$ |
| **17** | DUTH | $\mathbf{0.621}_{17}$ | $0.605_{17}$ | $0.640_{10}$ |
| **18** | CrystalNest | $\mathbf{0.619}_{18}$ | $0.593_{19}$ | $0.629_{13}$ |
| **19** | deepSA | $\mathbf{0.618}_{19}$ | $0.587_{20}$ | $0.616_{17}$ |
| **20** | NILC-USP | $\mathbf{0.612}_{20}$ | $0.595_{18}$ | $0.617_{16}$ |
| **21** | Ti-Senti | $\mathbf{0.607}_{21}$ | $0.577_{22}$ | $0.627_{15}$ |
| **22** | BUSEM | $\mathbf{0.605}_{22}$ | $0.587_{20}$ | $0.603_{22}$ |
| **23** | EICA | $\mathbf{0.595}_{23}$ | $0.555_{24}$ | $0.599_{24}$ |
| **24** | OMAM | $\mathbf{0.590}_{24}$ | $0.542_{26}$ | $0.615_{19}$ |
| **25** | Adullam | $\mathbf{0.589}_{25}$ | $0.552_{25}$ | $0.614_{20}$ |
| **26** | NileTMRG | $\mathbf{0.578}_{26}$ | $0.515_{32}$ | $0.606_{21}$ |
| **27** | Amobee-C-137 | $\mathbf{0.575}_{27}$ | $0.520_{30}$ | $0.587_{27}$ |
| **28** | ej-za-2017 | $\mathbf{0.571}_{28}$ | $0.539_{27}$ | $0.582_{28}$ |
| | LSIS | $\mathbf{0.571}_{28}$ | $0.561_{23}$ | $0.521_{34}$ |
| **30** | XJSA | $\mathbf{0.556}_{30}$ | $0.519_{31}$ | $0.575_{29}$ |
| **31** | Neverland-THU | $\mathbf{0.555}_{31}$ | $0.507_{33}$ | $0.597_{26}$ |
| **32** | MI&T-Lab | $\mathbf{0.551}_{32}$ | $0.522_{29}$ | $0.561_{31}$ |
| **33** | diegoref | $\mathbf{0.546}_{33}$ | $0.527_{28}$ | $0.540_{33}$ |
| **34** | xiwu | $\mathbf{0.479}_{34}$ | $0.365_{34}$ | $0.547_{32}$ |
| **35** | SSN_MLRG1 | $\mathbf{0.431}_{35}$ | $0.344_{35}$ | $0.439_{35}$ |
| **36** | YNUDLG | $\mathbf{0.340}_{36}$ | $0.201_{37}$ | $0.387_{36}$ |
| **37** | WarwickDCS | $\mathbf{0.335}_{37}$ | $0.221_{36}$ | $0.382_{37}$ |
| | Avid | $\mathbf{0.335}_{37}$ | $0.163_{38}$ | $0.206_{38}$ |
| B1 | All POSITIVE | **0.333** | 0.162 | 0.193 |
| B2 | All NEGATIVE | **0.333** | **0.244** | 0.323 |
| B3 | All NEUTRAL | **0.333** | 0.000 | **0.483** |

Table 6: **Results for Subtask A "Message Polarity Classification", English.** The systems are ordered by average recall *AvgRec* (higher is better). In each column, the rankings according to the corresponding measure are indicated with a subscript. B$x$ indicates a baseline.

### 5.2 Results for Subtask A: Overall Sentiment in a Tweet

Tables 6 and 7 show the results for Subtask A in English and Arabic, respectively, where the teams are ranked by macro-average recall.

**For English** the best ranking teams were *BB_twtr* and *DataStories*, both achieving a macro-average recall of 0.681. Both top teams used deep learning; *BB_twtr* used an ensemble of LSTMs and CNNs with multiple convolution operations, while *DataStories* used deep LSTM networks with an attention mechanism.

| # | System | *AvgRec* | $F_1^{PN}$ | *Acc* |
|---|---|---|---|---|
| **1** | NileTMRG | $\mathbf{0.583}_{1}$ | $0.610_{1}$ | $0.581_{1}$ |
| **2** | SiTAKA | $\mathbf{0.550}_{2}$ | $0.571_{2}$ | $0.563_{2}$ |
| **3** | ELiRF-UPV | $\mathbf{0.478}_{3}$ | $0.467_{4}$ | $0.508_{3}$ |
| **4** | INGEOTEC | $\mathbf{0.477}_{4}$ | $0.455_{5}$ | $0.499_{4}$ |
| **5** | OMAM | $\mathbf{0.438}_{5}$ | $0.422_{6}$ | $0.430_{8}$ |
| | LSIS | $\mathbf{0.438}_{5}$ | $0.469_{3}$ | $0.445_{6}$ |
| **7** | Tw-StAR | $\mathbf{0.431}_{7}$ | $0.416_{7}$ | $0.454_{5}$ |
| **8** | HLP@UPENN | $\mathbf{0.415}_{8}$ | $0.320_{8}$ | $0.443_{7}$ |
| B1 | All POSITIVE | **0.333** | 0.199 | 0.248 |
| B2 | All NEGATIVE | **0.333** | **0.267** | 0.364 |
| B3 | All NEUTRAL | **0.333** | 0.000 | **0.388** |

Table 7: **Results for Subtask A "Message Polarity Classification", Arabic.** The systems are ordered by average recall *AvgRec* (higher is better). In each column, the rankings according to the corresponding measure are indicated with a subscript. B$x$ indicates a baseline.

Both teams participated in all English subtasks and were also ranked in first (*BB_twtr*) and second (*DataStories*) place for subtasks B-D; *BB_twtr* was also ranked first for subtask E.

The top 5 teams for English were very closely scored. The following four best-ranked teams all used deep learning or deep learning ensembles. Three of the top-10 scoring teams (*INGEOTEC*, *SiTAKA*, and *UCSC-NLP*) used SVM classifiers instead, with various surface, lexical, semantic, and dense word embedding features. The use of ensembles clearly stood out, with five of the top-10 scoring systems (*BB_twtr*, *LIA*, *NNEMBs*, *Tweester*, and *INGEOTEC*) using ensembles, hybrid, stacking or some kind of mix of learning methods. All teams beat the baseline on macro-average recall; however, a few teams did not beat the harsher average F-measure and accuracy baselines.

**For Arabic** the best-ranked team was *NileTMRG*, and it achieved a score of 0.583. They used a Naïve Bayes classifier with a combination of lexical and sentiment features; they further augmented the training dataset to about 13K examples using external tweets. The *SiTAKA* team was ranked second with a score of 0.55. Their system used a feature-rich SVM with lexical features and embedding representations. Except for *EliRF-UPV*, who used multi-layer neural networks (CRNNs), the remaining teams used SVM and Naïve Bayes classifiers, genetic algorithms, or conditional random fields (CRFs). All teams managed to beat all baselines for all metrics.

The difference in the absolute scores for the two languages is probably partially due to the difference in the amount of training data available for Arabic, which was much smaller compared English, even when external datasets were taken into account. The results also reflect the linguistic complexity of Arabic as it is used in social media, which is characterized by the abundant use of dialectal forms and spelling variants. Overall, participants preferred to focus on developing Arabic-specific systems (varying in the extent to which they applied Arabic-specific preprocessing) rather than trying to leverage cross-language models that would enable them to use English data to augment their Arabic models.

| # | System | *AvgRec* | $F_1$ | *Acc* |
|---|---|---|---|---|
| **1** | BB_twtr | $\mathbf{0.882}_{1}$ | $0.890_{1}$ | $0.897_{1}$ |
| **2** | DataStories | $\mathbf{0.856}_{2}$ | $0.861_{2}$ | $0.869_{2}$ |
| **3** | Tweester | $\mathbf{0.854}_{3}$ | $0.856_{3}$ | $0.863_{3}$ |
| **4** | TopicThunder | $\mathbf{0.846}_{4}$ | $0.847_{4}$ | $0.854_{4}$ |
| **5** | TakeLab | $\mathbf{0.845}_{5}$ | $0.836_{5}$ | $0.840_{6}$ |
| **6** | funSentiment | $\mathbf{0.834}_{6}$ | $0.824_{8}$ | $0.827_{8}$ |
| | YNU-HPCC | $\mathbf{0.834}_{6}$ | $0.816_{10}$ | $0.818_{10}$ |
| **8** | WarwickDCS | $\mathbf{0.829}_{8}$ | $0.834_{6}$ | $0.843_{5}$ |
| **9** | CrystalNest | $\mathbf{0.827}_{9}$ | $0.822_{9}$ | $0.827_{8}$ |
| **10** | Ti-Senti | $\mathbf{0.826}_{10}$ | $0.830_{7}$ | $0.838_{7}$ |
| **11** | Amobee-C-137 | $\mathbf{0.822}_{11}$ | $0.801_{12}$ | $0.802_{12}$ |
| **12** | SINAI | $\mathbf{0.818}_{12}$ | $0.806_{11}$ | $0.809_{11}$ |
| **13** | NRU-HSE | $\mathbf{0.798}_{13}$ | $0.787_{13}$ | $0.790_{13}$ |
| **14** | EICA | $\mathbf{0.790}_{14}$ | $0.775_{14}$ | $0.777_{16}$ |
| **15** | OMAM | $\mathbf{0.779}_{15}$ | $0.762_{17}$ | $0.764_{17}$ |
| **16** | NileTMRG | $\mathbf{0.769}_{16}$ | $0.774_{15}$ | $0.789_{15}$ |
| **17** | ELiRF-UPV | $\mathbf{0.766}_{17}$ | $0.773_{16}$ | $0.790_{13}$ |
| **18** | DUTH | $\mathbf{0.663}_{18}$ | $0.600_{18}$ | $0.607_{18}$ |
| **19** | ej-za-2017 | $\mathbf{0.594}_{19}$ | $0.486_{21}$ | $0.518_{19}$ |
| **20** | SSN_MLRG1 | $\mathbf{0.586}_{20}$ | $0.494_{20}$ | $0.518_{19}$ |
| **21** | YNUDLG | $\mathbf{0.516}_{21}$ | $0.499_{19}$ | $0.499_{21}$ |
| **22** | TM-Gist | $\mathbf{0.499}_{22}$ | $0.428_{22}$ | $0.444_{22}$ |
| **23** | SSK_JNTUH | $\mathbf{0.483}_{23}$ | $0.372_{23}$ | $0.412_{23}$ |
| B1 | All POSITIVE | **0.500** | 0.285 | 0.398 |
| B2 | All NEGATIVE | **0.500** | **0.376** | **0.602** |

Table 8: **Results for Subtask B "Tweet classification according to a two-point scale", English.** The systems are ordered by average recall *AvgRec* (higher is better). B$x$ indicates a baseline.

| # | System | *AvgRec* | $F_1$ | *Acc* |
|---|---|---|---|---|
| **1** | NileTMRG | $\mathbf{0.768}_{1}$ | $0.767_{1}$ | $0.770_{1}$ |
| **2** | ELiRF-UPV | $\mathbf{0.721}_{2}$ | $0.724_{2}$ | $0.734_{2}$ |
| **3** | ASA | $\mathbf{0.693}_{3}$ | $0.670_{4}$ | $0.672_{4}$ |
| **4** | OMAM | $\mathbf{0.687}_{4}$ | $0.678_{3}$ | $0.679_{3}$ |
| B1 | All POSITIVE | **0.500** | **0.362** | **0.566** |
| B2 | All NEGATIVE | 0.500 | 0.303 | 0.434 |

Table 9: **Results for Subtask B "Tweet classification according to a two-point scale", Arabic.** The systems are ordered by average recall *AvgRec* (higher is better). B$x$ indicates a baseline.

### 5.3 Results for Subtasks B and C: Topic-Based Classification

The results of Subtasks B and C are shown in Tables 8–11. We can see that the system scores for subtask B are higher than those for subtask A, with the best team achieving 0.882 accuracy for English (compared to 0.681 for subtask A) and 0.768 for Arabic (compared to 0.583 for subtask A). However, this is primarily due to the fact there are two classes for subtask B, while there are three classes for subtask A.

| # | System | $MAE^M$ | $MAE^\mu$ |
|---|---|---|---|
| **1** | BB_twtr | $\mathbf{0.481}_{1}$ | $0.554_{6}$ |
| **2** | DataStories | $\mathbf{0.555}_{2}$ | $0.543_{4}$ |
| **3** | Amobee-C-137 | $\mathbf{0.599}_{3}$ | $0.582_{10}$ |
| **4** | Tweester | $\mathbf{0.623}_{4}$ | $0.734_{13}$ |
| **5** | TwiSe | $\mathbf{0.640}_{5}$ | $0.616_{12}$ |
| **6** | CrystalNest | $\mathbf{0.698}_{6}$ | $0.571_{9}$ |
| **7** | ELiRF-UPV | $\mathbf{0.806}_{7}$ | $0.586_{11}$ |
| **8** | EICA | $\mathbf{0.823}_{8}$ | $0.509_{2}$ |
| **9** | funSentiment | $\mathbf{0.842}_{9}$ | $0.530_{3}$ |
| **10** | DUTH | $\mathbf{0.895}_{10}$ | $0.544_{5}$ |
| | OMAM | $\mathbf{0.895}_{10}$ | $0.475_{1}$ |
| **12** | YNU-HPCC | $\mathbf{0.925}_{12}$ | $0.567_{8}$ |
| **13** | NRU-HSE | $\mathbf{0.928}_{13}$ | $0.557_{7}$ |
| **14** | YNU-1510 | $\mathbf{1.262}_{14}$ | $0.764_{14}$ |
| **15** | SSN_MLRG1 | $\mathbf{1.325}_{15}$ | $0.985_{15}$ |
| B1 | HIGHLYNEGATIVE | 2.000 | 1.895 |
| B2 | NEGATIVE | 1.400 | 0.923 |
| B3 | NEUTRAL | **1.200** | **0.525** |
| B4 | POSITIVE | 1.400 | 1.127 |
| B5 | HIGHLYPOSITIVE | 2.000 | 2.105 |

Table 10: **Results for Subtask C "Tweet classification according to a five-point scale", English.** The systems are ordered by their $MAE^M$ score (lower is better). B$x$ indicates a baseline.

| # | System | $MAE^M$ | $MAE^\mu$ |
|---|---|---|---|
| **1** | OMAM | $\mathbf{0.943}_{1}$ | $0.646_{1}$ |
| **2** | ELiRF-UPV | $\mathbf{1.264}_{2}$ | $0.787_{2}$ |
| B1 | HIGHLYNEGATIVE | 2.000 | 2.059 |
| B2 | NEGATIVE | 1.400 | 1.065 |
| B3 | NEUTRAL | **1.200** | **0.458** |
| B4 | POSITIVE | 1.400 | 0.946 |
| B5 | HIGHLYPOSITIVE | 2.000 | 1.941 |

Table 11: **Results for Subtask C "Tweet classification according to a five-point scale", Arabic.** The systems are ordered by their $MAE^M$ score (lower is better). B$x$ indicates a baseline.

**For English** the *BB_twtr* system, ranked first, modeled topics by concatenating the topical information at the word level. The second-best system, *DataStories*, also accounted for topics by producing topic annotations and a context-aware attention mechanism.

| # | System | $KLD$ | $AE$ | $RAE$ |
|---|---|---|---|---|
| **1** | BB_twtr | $\mathbf{0.036}_{1}$ | $0.080_{1}$ | $0.598_{1}$ |
| **2** | DataStories | $\mathbf{0.048}_{2}$ | $0.095_{2}$ | $0.848_{2}$ |
| **3** | TakeLab | $\mathbf{0.050}_{3}$ | $0.096_{3}$ | $1.057_{5}$ |
| **4** | CrystalNest | $\mathbf{0.056}_{4}$ | $0.104_{5}$ | $1.202_{6}$ |
| **5** | Tweester | $\mathbf{0.057}_{5}$ | $0.103_{4}$ | $1.051_{4}$ |
| **6** | funSentiment | $\mathbf{0.060}_{6}$ | $0.109_{6}$ | $0.939_{3}$ |
| **7** | NileTMRG | $\mathbf{0.077}_{7}$ | $0.120_{7}$ | $1.228_{7}$ |
| **8** | NRU-HSE | $\mathbf{0.078}_{8}$ | $0.132_{8}$ | $1.528_{8}$ |
| **9** | EICA | $\mathbf{0.092}_{9}$ | $0.143_{9}$ | $1.922_{9}$ |
| **10** | THU_HCSI_IDU | $\mathbf{0.129}_{10}$ | $0.179_{10}$ | $2.428_{11}$ |
| **11** | Amobee-C-137 | $\mathbf{0.149}_{11}$ | $0.179_{10}$ | $2.168_{10}$ |
| **12** | OMAM | $\mathbf{0.164}_{12}$ | $0.204_{12}$ | $2.790_{12}$ |
| **13** | SSK_JNTUH | $\mathbf{0.421}_{13}$ | $0.314_{13}$ | $2.983_{13}$ |
| **14** | ELiRF-UPV | $\mathbf{1.060}_{14}$ | $0.593_{15}$ | $7.991_{15}$ |
| **15** | YNU-HPCC | $\mathbf{1.142}_{15}$ | $0.592_{14}$ | $7.859_{14}$ |
| B1 | (0 1) | 1.518 | 0.422 | **2.645** |
| B2 | macro-avg on 2016 data | 0.554 | 0.423 | 6.061 |
| B3 | micro-avg on 2016 data | 0.591 | 0.432 | 6.169 |
| B4 | macro-avg on 2015-6 data | **0.534** | **0.418** | 6.000 |
| B5 | micro-avg on 2015-6 data | 0.587 | 0.431 | 6.157 |

Table 12: **Results for Subtask D "Tweet quantification according to a two-point scale", English.** The systems are ordered by their $KLD$ score (lower is better). B$x$ indicates a baseline.

| # | System | $KLD$ | $AE$ | $RAE$ |
|---|---|---|---|---|
| **1** | NileTMRG | $\mathbf{0.127}_{1}$ | $0.170_{1}$ | $2.462_{1}$ |
| **2** | OMAM | $\mathbf{0.202}_{2}$ | $0.238_{2}$ | $4.835_{2}$ |
| **3** | ELiRF-UPV | $\mathbf{1.183}_{3}$ | $0.537_{3}$ | $11.434_{3}$ |
| B1 | (0 1) | 1.518 | 0.422 | **2.645** |
| B2 | macro-avg on train-2017 | 0.296 | 0.322 | 6.600 |
| B3 | micro-avg on train-2017 | **0.295** | **0.321** | 6.692 |

Table 13: **Results for Subtask D "Tweet quantification according to a two-point scale", Arabic.** The systems are ordered by their $KLD$ score (lower is better). B$x$ indicates a baseline.

*funSentiment*, ranked 6th and 9th for subtasks B and C, respectively, modeled the sentiment towards the topic using the left and the right context around a topic mention in the tweet. *WarwickDCS*, ranked 8th, used simple tweet-level classification, while ignoring the topic. Overall, almost all teams managed to outperform the majority class baseline for subtask B, but only two teams outperformed the NEUTRAL class baseline for subtask C.

**For Arabic** four teams participated in Subtask B and two teams in Subtask C. *NileTMRG* was once again ranked first for Subtask B, with a system based on ensembles of topic-specific and topic-agnostic models. For subtask C, *OMAM* also used combinations of such models applied in succession. All teams easily outperformed the baselines for Subtask B, but only the *OMAM* team managed to do so for Subtask C.

| # | System | $EMD$ |
|---|---|---|
| 1 | BB_twtr | 0.245 |
| 2 | TwiSe | 0.269 |
| 3 | funSentiment | 0.273 |
| 4 | ELiRF-UPV | 0.306 |
| 5 | NRU-HSE | 0.317 |
| 6 | Amobee-C-137 | 0.345 |
| 7 | OMAM | 0.350 |
| 8 | Tweester | 0.365 |
| 9 | THU_HCSI_IDU | 0.385 |
| 10 | YNU-HPCC | 0.447 |
| 11 | DataStories | 0.595 |
| 12 | EICA | 1.461 |
| B1 | (0 0 0 1 0) | 1.123 |
| B2 | macro-avg on 2016 data | 0.583 |
| B3 | micro-avg on 2016 data | **0.552** |

Table 14: **Results for Subtask E "Tweet quantification according to a five-point scale", English.** The systems are ordered by their $EMD$ score (lower is better). B$x$ indicates a baseline.

| # | System | $EMD$ |
|---|---|---|
| 1 | OMAM | 0.548 |
| 2 | ELiRF-UPV | 0.564 |
| B1 | (0 0 1 0 0) | 0.458 |
| B2 | macro-avg on train-2017 | **0.440** |
| B3 | micro-avg on train-2017 | **0.440** |

Table 15: **Results for Subtask E "Tweet quantification according to a five-point scale", Arabic.** The systems are ordered by their $EMD$ score (lower is better). B$x$ indicates a baseline.

### 5.4 Results for Subtasks D and E: Tweet Quantification

Tables 12–15 show the results for the tweet quantification subtasks. The bottom of the tables report the result of a baseline system, B1, that assigns a prevalence of 1 to the majority class (which is the POSITIVE class for subtask D, and the WEAKLYPOSITIVE/NEUTRAL class for subtask E, English/Arabic) and 0 to the other class(es).

We further show the results for a smarter "maximum likelihood" baseline, which 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 $KLD$. For this baseline, for English we use for training either (*i*) the 2016 data only, or (*ii*) data from both 2015 and 2016; we also experiment with (*i*) micro-averaging and (*ii*) macro-averaging over the topics. It turns out that macro-averaging over 2015+2016 data is the strongest baseline in terms of $KLD$. For Arabic, we use the train-2017 data, and micro-averaging works better there.

There were 15 participating teams competing in Subtask D: 15 for English and 3 for Arabic (these 3 teams all participated in English). As in the other subtasks, *BB_twtr* was ranked first in English. They achieved an improvement of .50 points absolute in KLD over the best baseline, and a .01 improvement over the next best team, *DataStories*. For Arabic, the best team was *NileTMRG* With improvement of .17 over the best baseline and of .08 over the next best team, *OMAM*. All but the last two teams in English and the last team for Arabic outperformed all baselines.

In Subtask E, there were 12 participating teams, with *OMAM* and *EliRF-UPV* competing for both English and Arabic. Once again, *BB_twtr* was the best for English, improving over the best baseline by .31 EMD points absolute. Interestingly, this is the first subtask where *DataStories* was not the second-ranked team. *BB_twtr* outperformed the second-best team, *TwiSe*, by .02 points. For English, all but the last two teams outperformed the baselines. However, for Arabic, none of the two participating teams could do so.

### 5.5 User Information

This year, we encouraged teams to explore using in their models information about the user who wrote the tweet, which can be extracted from the public user profiles of the respective Twitter users. Participants could also try features about following relations and the structure of the social network in general, as well as could make use of other tweets by the target user when analyzing one particular tweet. Four teams tried that: *SINAI*, *ECNU*, *TakeLab*, and *OMAM*. *OMAM* and *TakeLab* did not find any improvements, and ultimately decided not to use any user information. *ECNU* used profile information such as favorited, favorite count, retweeted, and retweet count. They ended up 15th in Subtask A. *SINAI* used the last 200 tweets from the person's timeline. They ranked 12th in Subtask B. They generated a user model from the timeline of a given target user. They built a general SVM model on word2vec embeddings. Then, for each user in the test set, they downloaded the last 200 tweets published by the user and classified their sentiment using that SVM classifier. If the classified user tweets achieved an accuracy above a threshold (0.7), the user model was applied on the authored tweets from the test set. If not, the general SVM model was used.

It is difficult to judge whether and by how much user information could help the best approaches as they did not try to use such information. However, we believe that building and using a Twitter user profile is a promising research direction, and that participants should learn how to make this work in the future. Thus, we would like to encourage more teams to try to explore using this information. We would also like to provide more user information such as age and gender, which we can predict automatically (Rosenthal and McKeown, 2016), when it is not directly available from the user profile. Another promising direction is to make use of "conversations" in Twitter, i.e., take into account the replies to tweets in Twitter. For example, previous work (Vanzo et al., 2014) has shown that it is beneficial to model the polarity detection problem as a sequential classification task over streams of tweets, where the stream is a "conversation" on Twitter containing tweets, replies to these tweets, replies to these replies, etc.

## 6 Conclusion and Future Work

Sentiment Analysis in Twitter continues to be a very popular task, attracting 48 teams this year. The task provides immense value to the sentiment community by providing a large accessible benchmark dataset containing over 70,000 tweets across two languages for researchers to evaluate and compare their method to the state of the art. This year, we introduced a new language for the first time and also encouraged the use of user information. These additions drew new participants and ideas to the task. The Arabic tasks drew nine participants and four teams took advantage of user information. Although a respectable amount of participants for its inaugural year, further exploration into both of these areas would be useful in the future, such as collecting more training data for Arabic and encouraging the use of cross-lingual training data. In the future, we would like to include exploring additional languages, providing further user information, and other related tasks such as irony and emotion detection. Finally, deep learning continues to be popular and employed by the state of the art approaches. We expect this trend to continue in sentiment analysis research, but also look forward to new innovative ideas that are discovered.

| Team ID | Affiliation | Country | Subtasks | | | | | | | | | | Paper |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | English | | | | | Arabic | | | | | |
| Adullam | Korea University | South Korea | A | | | | | | | | | | (Yoon et al., 2017) |
| Amobee C-137 | Amobee | USA | A | B | C | D | E | | | | | | (Rozental and Fleischer, 2017) |
| ASA | Al-Imam Muhammad Ibn Saud Islamic University. | Saudi Arabia | | | | | | | B | | | | N/A |
| Avid | N/A | N/A | A | | | | | | | | | | N/A |
| BB_twtr | Bloomberg | USA | A | B | C | D | E | | | | | | (Cliche, 2017) |
| BUSEM | Bogazici University | Turkey | A | | | | | | | | | | (Ayata et al., 2017) |
| CrystalNest | Institute of High Performance Computing (IHPC) | Singapore | A | B | C | D | | | | | | | (Gupta and Yang, 2017) |
| DataStories | Data Science Lab at University of Piraeus | Greece | A | B | C | D | E | | | | | | (Baziotis et al., 2017) |
| deepSA | National Sun Yat-sen University | Taiwan | A | | | | | | | | | | (Yang et al., 2017) |
| diegoref | N/A | N/A | A | | | | | | | | | | N/A |
| DUTH | Democritus University of Thrace | Greece | A | B | C | | | | | | | | (Symeonidis et al., 2017) |
| ECNU | East China Normal University | China | A | | | | | | | | | | (Zhou et al., 2017) |
| EICA | East China Normal University | China | A | B | C | D | E | | | | | | (Maoquan et al., 2017) |
| ej-sa-2017 | University of Evora | Portugal | A | B | | | | | | | | | (Dovdon and Saias, 2017) |
| ELiRF-UPV | Universitat Politécnica de Valéncia | Spain | A | B | C | D | E | A | B | C | D | E | (González et al., 2017) |
| funSentiment | Thomson Reuters | USA | | B | C | D | E | | | | | | (Li et al., 2017) |
| HLP@UPENN | University of Pennsylvania | USA | A | | | | | A | | | | | (Sarker and Gonzalez, 2017) |
| INGEOTEC | CONACYT-INFOTEC/CENTROGEO | Mexico | A | | | | | A | | | | | (Miranda-Jiménez et al., 2017) |
| LIA | LIA | France | A | | | | | | | | | | (Rouvier, 2017) |
| LSIS | Aix-Marseille University | France | A | | | | | A | | | | | (Htait et al., 2017) |
| MI&T Lab | Harbin Institute of Technology | China | A | | | | | | | | | | (Zhao et al., 2017) |
| Neverland-THU | N/A | N/A | A | | | | | | | | | | N/A |
| NILC-USP | Institute of Mathematics and Computer Science, University of So Paulo | Brazil | A | | | | | | | | | | (Anselmo Corrêa Júnior et al., 2017) |
| NileTMRG | Nile University | Egypt | A | B | | D | | A | B | | D | | (El-Beltagy et al., 2017) |
| NNEMBs | Peking University | China | A | | | | | | | | | | (Yin et al., 2017) |
| NRU-HSE | National Research University Higher School of Economics | Russia | | B | C | D | E | | | | | | (Karpov, 2017) |
| OMAM | American University of Beirut, Universiti Teknologi Malaysia, Cairo University, New York University Abu Dhabi, Qatar University | Egypt, Lebanon, Malaysia, Qatar, United Arab Emirates | A | B | C | D | E | A | B | C | D | E | (Baly et al., 2017; Onyibe and Habash, 2017) |
| QUB | Queen's University Belfast | Ireland | A | | | | | | | | | | |
| senti17 | Lip6, UPMC | France | A | | | | | | | | | | (Hamdan, 2017) |
| SentiME++ | EURECOM | France | A | | | | | | | | | | (Troncy et al., 2017) |
| SINAI | Universidad de Jaén | Spain | | B | | | | | | | | | (Jiménez-Zafra et al., 2017) |
| SiTAKA | iTAKA, Universitat Rovira i Virgili; Hodeidah University | Spain, Yemen | A | | | | | A | | | | | (Jabreel and Moreno, 2017) |
| SSK_JNTUH | J.N.T.U.H College of Engg Jagtial and BVRIT Hyderabad College of Engineering for Women | India | | B | | D | | | | | | | N/A |
| SSN_MLRG1 | Department of CSE, SSN College of Engineering | India | A | B | C | | | | | | | | (Deborah et al., 2017) |
| TakeLab | TakeLab, University of Zagreb | Croatia | A | B | | D | | | | | | | (Lozić et al., 2017) |
| THU_HCSI_IDU | Human Computer Speech Interaction Research Group, Tsinghua University | China | | | | D | E | | | | | | |
| Ti-Senti | N/A | N/A | A | B | | | | | | | | | N/A |
| TM-Gist | N/A | | | B | | | | | | | | | N/A |
| TopicThunder | N/A | N/A | | B | | | | | | | | | N/A |
| TSA-INF | Infosys Limited | India | A | | | | | | | | | | (Deshmane and Friedrichs, 2017) |
| Tw-StAR | Selcuk University, Universit Libre de Bruxelles (ULB) | Belgium, Turkey | | | | | | A | | | | | (Mulki et al., 2017) |
| Tweester | National Technical University of Athens, University of Athens, "Athena" Research and Innovation Center, Signal Analysis and Interpretation Laboratory (SAIL), USC | Greece, USA | A | B | C | D | E | | | | | | (Kolovou et al., 2017) |
| TwiSe | University of Grenoble-Alps | France | | | C | | E | | | | | | (Balikas, 2017) |
| UCSC-NLP | Catholic University of the Most Holy Conception | Chile | A | | | | | | | | | | (Castro et al., 2017) |
| WarwickDCS | Department of Computer Science, University of Warwick | UK | A | B | | | | | | | | | N/A |
| XJSA | Xi'an JiaoTong University | China | A | | | | | | | | | | (Hao et al., 2017) |
| YNU-HPCC | Yunnan University | China | A | B | C | D | E | | | | | | (Zhang et al., 2017) |
| YNUDLG | Yunnan University | China | A | B | C | | | | | | | | (Wang et al., 2017) |
| **TOTAL** | | | **38** | **23** | **15** | **15** | **12** | **8** | **4** | **2** | **3** | **2** | |

Table 16: Alphabetical list of the participating teams, their affiliation, country, the subtasks they participated in, and the system description paper that they contributed to SemEval-2017. Teams whose *Affiliation* column is typeset on more than one row include researchers from different institutions, which have collaborated to build a joint system submission. An *N/A* entry for the *Paper* column indicates that the team did not contribute a system description paper. Finally, the last row gives statistics about the total number of system submissions for each subtask.

## References


Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language* 28(1):20–37.

Muhammad Abdul-Mageed and Mona T. Diab. 2011. Subjectivity and sentiment annotation of modern standard Arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*. Portland, Oregon, USA, LAW '11, pages 110–118.

Mohammad Al-Smadi, Omar Qawasmeh, Bashar Talafha, and Muhannad Quwaider. 2015. Human annotated Arabic dataset of book reviews for aspect based sentiment analysis. In *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud*. Rome, Italy, FiCloud '15, pages 726–730.

Edilson Anselmo Corrêa Júnior, Vanessa Marinho, and Leandro Santos. 2017. NILC-USP at SemEval-2017 Task 4: A Multi-view Ensemble for Twitter Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 610–614.

Deger Ayata, Murat Saraclar, and Arzucan Ozgur. 2017. BUSEM at SemEval-2017 Task 4A: Sentiment Analysis with Word Embedding and Long Short Term Memory RNN Approaches. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 776–782.

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*. Pisa, Italy, ISDA '09, pages 283–287.

Georgios Balikas. 2017. TwiSe at SemEval-2017 Task 4: Five-point Twitter Sentiment Classification and Quantification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 754–758.

Ramy Baly, Gilbert Badaro, Ali Hamdi, Rawan Moukalled, Rita Aoun, Georges El-Khoury, Ahmad Al Sallab, Hazem Hajj, Nizar Habash, Khaled Shaban, and Wassim El-Hajj. 2017. OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval 17, pages 602–609.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 746–753.

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

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*. Vancouver, Canada, CSCW '15, pages 700–711.

Suzan Burton and Alena Soboleva. 2011. Interactive or reactive? Marketing with Twitter. *Journal of Consumer Marketing* 28(7):491–499.

Iván Castro, Sebastián Oliva, José Abreu, Claudia Martínez, and Yoan Gutiérrez. 2017. UCSC-NLP at SemEval-2017 Task 4: Sense n-grams for sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 806–810.

Mathieu Cliche. 2017. BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 572–579.

Angel Deborah, Milton Rajendram, and T. Mirnalinee. 2017. SSN_MLRG1 at SemEval-2017 Task 4: Sentiment Analysis in Twitter Using Multi-Kernel Gaussian Process Classifier . In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17.

Amit Ajit Deshmane and Jasper Friedrichs. 2017. TSA-INF at SemEval-2017 Task 4: An ensemble of deep learning architectures including lexicon features for Twitter sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 801–805.

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).

Enkhzol Dovdon and José Saias. 2017. ej-sa-2017 at SemEval-2017 Task 4: Experiments for target oriented sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 643–646.

Samhaa R. El-Beltagy, Mona El Kalamawy, and Abu Bakr Soliman. 2017. NileTMRG at SemEval-2017 Task 4: Arabic sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 789–794.

Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. 2012. A machine learning approach for opinion holder extraction in Arabic language. *arXiv preprint arXiv:1206.1011* .

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.

Noura Farra and Kathleen McKeown. 2017. SMARTies: Sentiment models for Arabic target entities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, EACL '17.

Noura Farra, Kathy McKeown, and Nizar Habash. 2015. Annotating targets of opinions in Arabic using crowdsourcing. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*. Beijing, China, ANLP '17, pages 89–98.

George Forman. 2005. Counting positives accurately despite inaccurate classification. In *Proceedings of the 16th European Conference on Machine Learning*. Porto, Portugal, ECML '05, pages 564–575.

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

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 Task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 470–478.

José Ángel González, Ferran Pla, and Lluís-F Hurtado. 2017. ELiRF-UPV at SemEval-2017 Task 4: Sentiment Analysis using Deep Learning. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 722–726.

Raj Kumar Gupta and Yinping Yang. 2017. CrystalNest at SemEval-2017 Task 4: Using sarcasm detection for enhancing sentiment classification and quantification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 625–632.

Hussam Hamdan. 2017. Senti17 at SemEval-2017 Task 4: Ten convolutional neural network voters for tweet polarity classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 699–702.

Yazhou Hao, YangYang Lan, Yufei Li, and Chen Li. 2017. XJSA at SemEval-2017 Task 4: A deep system for sentiment classification in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 727–730.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, ACL-IJCNLP '17, pages 752–762.

Amal Htait, Sébastien Fournier, and Patrice Bellot. 2017. LSIS at SemEval-2017 Task 4: Using adapted sentiment similarity seed words for English and Arabic tweet polarity classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 717–721.

Mohammed Jabreel and Antonio Moreno. 2017. SiTAKA at SemEval-2017 Task 4: Sentiment analysis in Twitter based on a rich set of features. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 693–698.

Salud María Jiménez-Zafra, Arturo Montejo-Ráez, M. Teresa Martín-Valdivia, and L. Alfonso Ureña López. 2017. SINAI at SemEval-2017 Task 4: User based classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 633–638.

Nikolay Karpov. 2017. NRU-HSE at SemEval-2017 Task 4: Tweet quantification using deep learning architecture. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 682–687.

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 International Symposium on Computer and Information Sciences*. Paris, France, ISCIS '13, pages 139–148.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. SemEval-2016 Task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 42–51.

Athanasia Kolovou, Filippos Kokkinos, Aris Fergadis, Pinelopi Papalampidi, Elias Iosif, Nikolaos Malandrakis, Elisavet Palogiannidi, Haris Papageorgiou, Shrikanth Narayanan, and Alexandro Potamianos. 2017. Tweester at SemEval-2017 Task 4: Fusion of semantic-affective and pairwise classification models for sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Seman-*

*tic Evaluation*. Vancouver, Canada, SemEval '17, pages 674–681.

Quanzhi Li, Armineh Nourbakhsh, Xiaomo Liu, Rui Fang, and Sameena Shah. 2017. funSentiment at SemEval-2017 Task 4: Topic-based message sentiment classification by exploiting word embeddings, text features and target contexts. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 740–745.

David Lozić, Doria Šarić, Ivan Tokić, Zoran Medić, and Jan Šnajder. 2017. TakeLab at SemEval-2017 Task 4: Recent deaths and the power of nostalgia in sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 783–788.

Wang Maoquan, Chen Shiyun, Xie Yufei, and Zhao Lu. 2017. EICA at SemEval-2017 Task 4: A simple convolutional neural network for topic-based sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 293–299.

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*. Avignon, France, EACL '12, pages 603–612.

Sabino Miranda-Jiménez, Mario Graff, Eric Sadit Tellez, and Daniela Moctezuma. 2017. INGEOTEC at SemEval 2017 Task 4: A B4MSA ensemble based on genetic programming for Twitter sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 770–775.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of Twitter users. In *Proceedings of the 9th AAAI International Conference on Web and Social Media*. Barcelona, Spain, ICWSM '11, pages 554–557.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016a. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 31–41.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016b. How translation alters sentiment. *J. Artif. Intell. Res.(JAIR)* 55:95–130.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*. Atlanta, Georgia, USA, WASSA '13, pages 55–64.

Hala Mulki, Hatem Haddad, Mourad Gridach, and Ismail Babaolu. 2017. Tw-StAR at SemEval-2017 Task 4: Sentiment classification of Arabic tweets. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 663–668.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016a. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 1–18.

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.

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*. Atlanta, Georgia, USA, SemEval '13, pages 312–320.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the Seventh International Conference on Weblogs and Social Media*. Cambridge, Massachusetts, USA, ICWSM '13, pages 439–448.

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 Fourth International Conference on Weblogs and Social Media*. Washington, DC, USA, ICWSM '10, pages 122–129.

Chukwuyem Onyibe and Nizar Habash. 2017. OMAM at SemEval-2017 Task 4: English sentiment analysis with conditional random fields. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 669–673.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, USA, SemEval '16, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect based sentiment

analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*. Dublin, Ireland, SemEval '14, pages 27–35.

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*. Singapore, AIRS '13, pages 170–180.

Eshrag Refaee and Verena Rieser. 2014. Subjectivity and sentiment analysis of Arabic Twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*. Reykjavik, Iceland, pages 16–21.

Eshrag Refaee and Verena Rieser. 2015. Benchmarking machine translated sentiment analysis for Arabic tweets. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, USA, NAACL-HLT '15, pages 71–78.

Sara Rosenthal and Kathy McKeown. 2016. Social proof: The impact of author traits on influence detection. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas, USA, pages 27–36.

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*. Denver, Colorado, USA, SemEval '15, pages 451–463.

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*. Dublin, Ireland, SemEval '14, pages 73–80.

Mickael Rouvier. 2017. LIA at SemEval-2017 Task 4: An ensemble of neural networks for sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 759–764.

Alon Rozental and Daniel Fleischer. 2017. Amobee at SemEval-2017 Task 4: Deep learning system for sentiment detection on Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 652–657.

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.

Ludger Rüschendorf. 2001. Wasserstein metric. In Michiel Hazewinkel, editor, *Encyclopaedia of Mathematics*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. SemEval-2015 Task 9: CLIPEval implicit polarity of events. In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado, USA, SemEval '15, pages 443–450.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, USA, NAACL-HLT '15, pages 767–777.

Abeed Sarker and Graciela Gonzalez. 2017. HLP@UPenn at SemEval-2017 Task 4A: A simple, self-optimizing text classification system combining dense and sparse vectors. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 639–642.

Fabrizio Sebastiani. 2015. An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. Northampton, Massachusetts, USA, ICTIR '15, pages 11–20.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*. Prague, Czech Republic, SemEval '07, pages 70–74.

Symeon Symeonidis, Dimitrios Effrosynidis, John Kordonis, and Avi Arampatzis. 2017. DUTH at SemEval-2017 Task 4: A voting classification approach for Twitter sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 703–707.

Raphael Troncy, Enrico Palumbo, Efstratios Sygkounas, and Giuseppe Rizzo. 2017. SentiME++ at SemEval-2017 Task 4: Stacking state-of-the-art classifiers to enhance sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 647–651.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in Twitter. In *Proceedings of the 25th International*

*Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, COLING '14, pages 2345–2354.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA, EMNLP '13, pages 1815–1827.

Ming Wang, Biao Chu, Qingxun Liu, and Xiaobing Zhou. 2017. YNUDLG at SemEval-2017 Task 4: A GRU-SVM model for sentiment classification and quantification in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 712–716.

Tzu-Hsuan Yang, Tzu-Hsuan Tseng, and Chia-Ping Chen. 2017. deepSA at SemEval-2017 Task 4: Interpolated deep neural networks for sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 615–619.

Yi Yang and Jacob Eisenstein. 2015. Putting things in context: Community-specific embedding projections for sentiment analysis. *CoRR* abs/1511.06052.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. NNEMBs at SemEval-2017 Task 4: Neural Twitter sentiment classification: a simple ensemble method with different embeddings. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 620–624.

Joosung Yoon, Hyeoncheol Kim, and Kigon Lyu. 2017. Adullam at SemEval-2017 Task 4: Sentiment analyzer using lexicon integrated convolutional neural networks with attention. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 731–735.

Haowei Zhang, Jin Wang, Jixian Zhang, and Xuejie Zhang. 2017. YNU-HPCC at SemEval 2017 Task 4: Using a multi-channel CNN-LSTM model for sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 795–800.

Jingjing Zhao, Yan Yang, and Bing Xu. 2017. MI&T Lab at SemEval-2017 Task 4: An integrated training method of word vector for sentiment classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, SemEval '17, pages 688–692.

Yunxiao Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 4: Evaluating effective features on machine learning methods for Twitter message polarity classification. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Vancouver, Canada, Semeval 2017, pages 811–815.