SemEval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases

Svetlana Kiritchenko and Saif M. Mohammad
National Research Council Canada
svetlana.kiritchenko@nrc-cnrc.gc.ca
saif.mohammad@nrc-cnrc.gc.ca

Mohammad Salameh
University of Alberta
msalameh@ualberta.ca

Abstract

We present a shared task on automatically determining sentiment intensity of a word or a phrase. The words and phrases are taken from three domains: general English, English Twitter, and Arabic Twitter. The phrases include those composed of negators, modals, and degree adverbs as well as phrases formed by words with opposing polarities. For each of the three domains, we assembled the datasets that include multi-word phrases and their constituent words, both manually annotated for real-valued sentiment intensity scores. The three datasets were presented as the test sets for three separate tasks (each focusing on a specific domain). Five teams submitted nine system outputs for the three tasks. All datasets created for this shared task are freely available to the research community.

1 Introduction

Words have prior associations with sentiment. For example, honest and competent are associated with positive sentiment, whereas dishonest and dull are associated with negative sentiment. Further, the degree of positivity (or negativity), also referred to as intensity, can vary. For example, most people will agree that succeed is more positive (or less negative) than improve, and fail is more negative (or less positive) than setback. We present a shared task where automatic systems are asked to predict a prior sentiment intensity score for a word or a phrase. The words and phrases are taken from three domains: general English, English Twitter, and Arabic Twitter.

For each domain, a separate task with its own development and test sets was set up. The phrases include those composed of negators (e.g., nothing wrong), modals (e.g., might be fun), and degree adverbs (e.g., fairly important) as well as phrases formed by words with opposing polarities (e.g., lazy sundays).

Lists of words and their associated sentiment are commonly referred to as sentiment lexicons. They are used in sentiment analysis. For example, a number of unsupervised classifiers rely primarily on sentiment lexicons to determine whether a piece of text is positive or negative. Supervised classifiers also often use features drawn from sentiment lexicons (Mohammad et al., 2013; Pontiki et al., 2014). Sentiment lexicons are also beneficial in stance detection (Mohammad et al., 2016a; Mohammad et al., 2016b), literary analysis (Hartner, 2013; Kleres, 2011; Mohammad, 2012), and for detecting personality traits (Minamikawa and Yokoyama, 2011; Mohammad and Kiritchenko, 2015).

Existing manually created sentiment lexicons tend to provide only lists of positive and negative words (Hu and Liu, 2004; Wilson et al., 2005; Mohammad and Turney, 2013). The coarse-grained distinctions may be less useful in downstream applications than having access to fine-grained (real-valued) sentiment association scores. Most of the existing sentiment resources are available only for English. Non-English resources are scarce and often based on automatic translation of the English lexicons (Abdul-Mageed and Diab, 2014; Eskander and Rambow, 2015). Manually created sentiment lexicons usually include only single words. Yet, the sentiment of a phrase can differ markedly from the
sentiment of its constituent words. Sentiment composition is the determining of sentiment of a multi-word linguistic unit, such as a phrase or a sentence, from its constituents. Lexicons that include sentiment associations for phrases and their constituents are useful in studying sentiment composition. We refer to them as sentiment composition lexicons.

Automatically created lexicons often have real-valued sentiment association scores, have a high coverage, can include longer phrases, and can easily be collected for a specific domain. However, due to the lack of manually annotated real-valued sentiment lexicons the quality of automatic lexicons are often assessed only extrinsically through their use in sentence-level sentiment prediction. In this shared task, we intrinsically evaluate automatic methods that estimate sentiment association scores for terms in English and Arabic. For this, we assembled three datasets of phrases and their constituent single words manually annotated for sentiment with real-valued scores (Kiritchenko and Mohammad, 2016a; Kiritchenko and Mohammad, 2016c).

We first introduced this task as part of the SemEval-2015 Task 10 ‘Sentiment Analysis in Twitter’ Subtask E (Rosenthal et al., 2015). The 2015 test set was restricted to English single words and simple two-word negated expressions commonly found in tweets. This year (2016), we broadened the scope of the task and included three different domains. Furthermore, we shifted the focus from single words to longer, more complex phrases to explore sentiment composition.

Five teams submitted nine system outputs for the three tasks. All submitted outputs correlated strongly with the gold term rankings (Kendall’s rank correlation above 0.35). The best results on all tasks were achieved with supervised methods by exploiting a variety of sentiment resources. The highest rank correlation was obtained by team ECNU on the General English test set (τ = 0.7). On the other two domains, the results were lower (τ of 0.4–0.5).

All datasets created as part of this shared task are freely available through the task website.¹ For ease of exploration, we also created online interactive visualizations for the two English datasets.²

## 2 Task Description

The task is formulated as follows: given a list of terms (single words and multi-word phrases), an automatic system needs to provide a score between 0 and 1 that is indicative of the term’s strength of association with positive sentiment. A score of 1 indicates maximum association with positive sentiment (or least association with negative sentiment) and a score of 0 indicates least association with positive sentiment (or maximum association with negative sentiment). If a term is more positive than another, then it should have a higher score than the other.

There are three tasks, one for each of the three domains:

- **General English Sentiment Modifiers Set:** This dataset comprises English single words and multi-word phrases from the general domain. The phrases are formed by combining a word and a modifier, where a modifier is a negator, an auxiliary verb, a degree adverb, or a combination of those, for example, *would be very easy, did not harm,* and *would have been nice.* The single word terms are chosen from the set of words that are part of the multi-word phrases, for example, *easy, harm,* and *nice.*

- **English Twitter Mixed Polarity Set:** This dataset focuses on English phrases made up of opposing polarity terms, for example, phrases such as *lazy sundays, best winter break, happy accident* and *couldn’t stop smiling.* The dataset also includes single word terms (as separate entries). These terms are chosen from the set of words that are part of the multi-word phrases. The multi-word phrases and single-word terms are drawn from a corpus of tweets, and include a small number of hashtag words (e.g., #wan-tit) and creatively spelled words (e.g., plssss). However, a majority of the terms are those that one would use in everyday English.

- **Arabic Twitter Set:** This dataset includes single words and phrases commonly found in Arabic tweets. The phrases in this set are formed only by combining a negator and a word.

Teams could participate in any one, two, or all three tasks; however, only one submission was al-

¹[http://alt.qcri.org/semeval2016/task7/](http://alt.qcri.org/semeval2016/task7/)
²[http://www.saifmohammad.com/WebPages/SCL.html](http://www.saifmohammad.com/WebPages/SCL.html)
Table 1: The number of single-word and multi-word terms in the development and test sets.

<table>
<thead>
<tr>
<th>Task</th>
<th>Total</th>
<th>Development set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>words</td>
<td>phrases all</td>
<td>words</td>
</tr>
<tr>
<td>General English Sentiment Modifiers</td>
<td>2,999</td>
<td>101 99 200</td>
<td>1,330</td>
</tr>
<tr>
<td>English Twitter Mixed Polarity</td>
<td>1,269</td>
<td>60 140 200</td>
<td>358</td>
</tr>
<tr>
<td>Arabic Twitter</td>
<td>1,366</td>
<td>167 33 200</td>
<td>1,001</td>
</tr>
</tbody>
</table>

In 2015, the task was set up similarly (Rosenthal et al., 2015). Single words and multi-word phrases from English Twitter comprised the development and test sets (1,515 terms in total). The phrases were simple two-word negated expressions (e.g., can't wait). Participants were allowed to use these datasets for the development of their systems.

### 3 Datasets of English and Arabic Terms Annotated for Sentiment Intensity

The three datasets, General English Sentiment Modifiers Set, English Twitter Mixed Polarity Set, and Arabic Twitter Set, were created through manual annotation using an annotation scheme known as Best–Worst Scaling (described below in Section 3.1). The terms for each set (domain) were chosen as described in Sections 3.2, 3.3, and 3.4, respectively. Note that the exact sources of data and the term selection procedures were not known to the participants. The total number of words and phrases included in each of the datasets can be found in Table 1. Table 2 shows a few example entries from each set.

#### 3.1 Best–Worst Scaling Method of Annotation

Best–Worst Scaling (BWS), also sometimes referred to as Maximum Difference Scaling (MaxDiff), is an annotation scheme that exploits the comparative approach to annotation (Louviere and Woodworth, 1990; Cohen, 2003; Louviere et al., 2015). Annotators are given four items (4-tuple) and asked which item is the Best (highest in terms of the property of interest) and which is the Worst (least in terms of the property of interest). These annotations can then be easily converted into real-valued scores of association between the items and the property, which eventually allows for creating a ranked list of items as per their association with the property of interest. The Best–Worst Scaling method has been shown to produce reliable annotations of terms for sentiment (Kiritchenko and Mohammad, 2016a).

Given \( n \) terms to be annotated, the first step is to randomly sample this set (with replacement) to obtain sets of four terms each, 4-tuples, that satisfy the following criteria:

1. no two 4-tuples have the same four terms;
2. no two terms within a 4-tuple are identical;

Table 2: Examples of entries with real-valued sentiment scores from the three datasets.
3. each term in the term list appears approximately in the same number of 4-tuples;
4. each pair of terms appears approximately in the same number of 4-tuples.

The terms for the three tasks were annotated separately. For each task, $2 \times n$ 4-tuples were generated, where $n$ is the total number of terms in the task.

Next, the sets of 4-tuples were annotated through a crowdsourcing platform, CrowdFlower. The annotators were presented with four terms at a time, and asked which term is the most positive (or least negative) and which is the most negative (or least positive). Below is an example annotation question.³ (The Arabic data was annotated through a similar questionnaire in Arabic.)

Focus terms:
1. shameless self promotion 2. happy tears 3. hug 4. major pain

Q1: Identify the term that is associated with the most amount of positive sentiment (or least amount of negative sentiment) – the most positive term:
1. shameless self promotion
2. happy tears
3. hug
4. major pain

Q2: Identify the term that is associated with the most amount of negative sentiment (or least amount of positive sentiment) – the most negative term:
1. shameless self promotion
2. happy tears
3. hug
4. major pain

Each 4-tuple was annotated by at least eight respondents. Let majority answer refer to the option most chosen for a question. For all three datasets, at least 80% of the responses matched the majority answer.

The responses were then translated into real-valued scores and also a ranking of terms by sentiment for all the terms through a simple counting procedure: For each term, its score is calculated as the percentage of times the term was chosen as the most positive minus the percentage of times the term was chosen as the most negative (Orme, 2009; Flynn and Marley, 2014). For this competition, we converted the scores into the range from 0 (the least positive) to 1 (the most positive). The resulting rankings constituted the gold annotations for the three datasets. Finally, random samples of 200 terms from each dataset with the corresponding gold annotations were released to the participants as development sets for the three tasks. The rest of the terms were kept as test sets.

3.2 General English Sentiment Modifiers Dataset

The terms for this dataset were taken from the Sentiment Composition Lexicon for Negators, Modals, and Degree Adverbs (SCL-NMA) (Kiritchenko and Mohammad, 2016b).⁴ SCL-NMA includes all 1,621 positive and negative words from Osgood’s seminal study on word meaning (Osgood et al., 1957) available in General Inquirer (Stone et al., 1966). In addition, it includes 1,586 high-frequency phrases formed by the Osgood words in combination with simple negators such as no, don’t, and never, modals such as can, might, and should, or degree adverbs such as very and fairly.⁵ The eligible adverbs were chosen manually from adverbs that appeared in combination with an Osgood word at least ten times in the British National Corpus (BNC)⁶. Each phrase includes at least one modal, one negator, or one adverb; a phrase can include several modifiers (e.g., would be very happy). Sixty-four different (single or multi-word) modifiers were used in the dataset.

For this shared task, we removed terms that were used in the SemEval-2015 dataset. The final SemEval-2016 General English Sentiment Modifiers dataset contains 2,999 terms.

3.3 English Twitter Mixed Polarity Dataset

The terms for this dataset were taken in part from the Sentiment Composition Lexicon for Opposing Polarity Phrases (SCL-OPP) (Kiritchenko and Mohammad).³ The complete lists of negators, modals, and degree adverbs used in this dataset are available on the task website: http://alt.qcri.org/semeval2016/task7/

⁴www.saifmohammad.com/WebPages/SCL.html#NMA
⁵The complete lists of negators, modals, and degree adverbs used in this dataset are available on the task website: http://alt.qcri.org/semeval2016/task7/
SCL-OPP was created as follows. We polled the Twitter API (from 2013 to 2015) to collect a corpus of tweets that contain emoticons: ‘:)' or ‘:('. From this corpus, we selected bigrams and trigrams that had at least one positive word and at least one negative word. The polarity labels (positive or negative) of the words were determined by simple look-up in existing sentiment lexicons: Hu and Liu lexicon (Hu and Liu, 2004), NRC Emotion lexicon (Mohammad and Turney, 2013), MPQA lexicon (Wilson et al., 2005), and NRC’s Twitter-specific lexicon (Kiritchenko et al., 2014). Apart from the requirement of having at least one positive and at least one negative word, an n-gram must satisfy the following criteria:

- the n-gram must have a clear meaning on its own, (for example, the n-gram should not start or end with ‘or’, ‘and’, etc.);
- the n-gram should not include a named entity;
- the n-gram should not include obscene language.

In addition, we ensured that there was a good variety of phrases—for example, even though there were a large number of bigrams of the form super w, where w is a negative adjective, only a small number of such bigrams were included. Finally, we aimed to achieve a good spread in terms of degree of sentiment association (from very negative terms to very positive terms, and all the degrees of polarity in between). For this, we estimated the sentiment score of each phrase using an automatic PMI-based method described in (Kiritchenko et al., 2014). Then, the full range of sentiment values was divided into 5 bins, and approximately the same number of terms were selected from each bin.

In total, 851 n-grams (bigrams and trigrams) were selected. We also chose for annotation all unigrams that appeared in the selected set of bigrams and trigrams. There were 810 such unigrams.

When selecting the terms, we used sentiment associations obtained from both manual and automatic lexicons. As a result, some unigrams had erroneous sentiment associations. After manually annotating the full set of 1,661 terms (that include unigrams, bigrams, and trigrams), we found that 114 bigrams and 161 trigrams had all their comprising unigrams of the same polarity. These 275 n-grams were discarded from SCL-OPP but are included in this task dataset. Further, for this task we removed terms that were used in the SemEval-2015 dataset or in the General English set. The final SemEval-2016 English Twitter Mixed Polarity dataset contains 1,269 terms.

### 3.4 Arabic Twitter Dataset

Mohammad et al. (2015) automatically generated three high-coverage sentiment lexicons from Arabic tweets using hashtags and emoticons: Arabic Emoticon Lexicon, Arabic Hashtag Lexicon, and Dialectal Arabic Hashtag Lexicon. In addition to Modern Standard Arabic (MSA), these three lexicons comprise terms in Dialectal Arabic as well as hashtagged compound words, e.g., 

السعادة.الزوجة (#MaritalHappiness), which do not usually appear in manually created lexicons. Apart from unigrams, they also include entries for bigrams. From these lexicons, we selected single words as well as bigrams representing negated expressions in the form of ‘negator w’, where negator is a negation trigger from a list of 16 common Arabic negation words. Words used in negated expressions, but missing from the original list were also included. The selected terms satisfied the following criteria:

- the terms must occur frequently in tweets;
- the terms should not be highly ambiguous.

We also wanted the set of terms as a whole to have these properties:

- the set should have a good spread in terms of degree of sentiment association (from very
negative terms to very positive terms, and all the degrees of polarity in between);

- the set should include both standard and dialectal Arabic, Romanized words, misspellings, hashtags, and other categories frequently used in Twitter. (We chose not to include URLs, user mentions, named entities, and obscene terms.)

The final SemEval-2016 Arabic Twitter dataset contains 1,366 terms.

4 Evaluation

Sentiment association scores are most meaningful when compared to each other; they indicate which term is more positive than the other. Therefore, the automatic systems were evaluated in terms of their abilities to correctly rank the terms by the degree of sentiment association.

For each task, the predicted sentiment intensity scores submitted by the participated systems were evaluated by first ranking the terms according to the proposed sentiment scores and then comparing this ranked list to the gold rankings. We used Kendall’s rank correlation coefficient (Kendall’s $\tau$) as the official evaluation metric to determine the similarity between the ranked lists (Kendall, 1938):

$$\tau = \frac{c - d}{n(n - 1)/2}$$

where $c$ is the number of concordant pairs, i.e., pairs of terms $w_i$ and $w_j$ for which both the gold ranked list and the predicted ranked list agree (either both lists rank $w_i$ higher than $w_j$ or both lists rank $w_i$ lower than $w_j$); $d$ is the number of discordant pairs, i.e., pairs of terms $w_i$ and $w_j$ for which the gold ranked list and the predicted ranked list disagree (one list ranks $w_i$ higher than $w_j$ and the other list ranks $w_i$ lower than $w_j$); and $n$ is the total number of terms. If any list ranks two terms $w_i$ and $w_j$ the same, this pair of terms is considered neither concordant nor discordant. The values of Kendall’s $\tau$ range from -1 to 1.

We also calculated scores for Spearman’s rank correlation (Spearman, 1904), as an additional (unofficial) metric.

5 Participated Systems

There were nine submissions from five teams—three submissions for each task. The team affiliations are shown in Table 3.

![Table 3: The participated teams and their affiliations.](image)

Tables 4 and 5 summarize the approaches and resources used by the participants in the (two) English and (one) Arabic tasks, respectively. Most teams applied supervised approaches and trained regression classifiers using a variety of features. Team ECNU treated the task as a rank prediction task instead of regression and trained a pair-wise ranking model with the Random Forest algorithm.

The development data available for each task was used as the training data by some teams. However, these data were limited (200 instances per task); therefore, other manually labeled resources were also explored. One commonly used resource was the LabMT lexicon—a set of over 100,000 frequent single words from 10 languages, including English and Arabic (about 10,000 words in each language), manually annotated for happiness through Mechanical Turk (Dodds et al., 2011; Dodds et al., 2015). For the Arabic task, two teams took advantage of the Arabic Twitter corpus collected by Refae and Rieser (2014). The features employed include sentiment scores obtained from different sentiment lexicons, general and sentiment-specific word embeddings, pointwise mutual information (PMI) scores between terms (single words and multi-word phrases) and sentiment classes, as well as lists of negators, intensifiers, and diminishers.
Only one team, *LSIS*, employed an unsupervised approach to all three tasks. To predict a sentiment intensity score for a term, they used the following three sources: existing sentiment lexicons, PMI scores between terms and sentiment classes computed on sentiment-annotated corpora, and PMI scores between terms and words *poor* and *excellent* computed on Google search results.

All teams heavily relied on existing sentiment lexicons: AFINN (Nielsen, 2011), ArabSenti (Abdul-Mageed et al., 2011), Hu and Liu (Hu and Liu, 2004), Dialectal Arabic Lexicon (Refaee and Rieser, 2014), JRC (Steinberger et al., 2012), MPQA (Wilson et al., 2005), NRC Emoticon (a.k.a. Sentiment140) (Kiriitchenko et al., 2014), NRC Emotion (Mohammad and Turney, 2013), NileULex (El-Beltagy, 2016b), and SentiWordNet (Esuli and Sebastiani, 2006). (Note that even though the NRC Emotion Lexicon was created for English terms, its translations in close to 40 languages, including Arabic, are available.\(^\text{12}\))

### 6 Results

The results for the three tasks are presented in Tables 6, 7, and 8. Team *ECNU* showed the best performance in both English-language tasks. In the Arabic task, the best performing system was developed by *iLab-Edinburgh*.

A few observations can be made from the results:

- On all three datasets, the team rankings based on the two metrics, Kendall’s \(\tau\) and Spearman’s \(\rho\), are the same.
- For most of the teams, the results obtained on the General English Sentiment Modifiers set are markedly higher than the results obtained on the other datasets.
- The English Twitter Mixed Polarity set proved to be a challenging task for all teams. We have further analyzed regularities present in different kinds of mixed polarity phrases and concluded that for most phrases the sentiment of the phrase cannot be reliably predicted only from the parts of speech and polarities of their

---

Table 6: Results for General English Sentiment Modifiers test set. The systems are ordered by their overall Kendall’s τ score, which was the official competition metric. The highest score is shown in bold.

<table>
<thead>
<tr>
<th>Team</th>
<th>Overall</th>
<th>Single words</th>
<th>Multi-word phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kendall’s τ</td>
<td>Spearman’s ρ</td>
<td>Kendall’s τ</td>
</tr>
<tr>
<td>ECNU</td>
<td>0.704</td>
<td>0.863</td>
<td>0.734</td>
</tr>
<tr>
<td>UWB</td>
<td>0.659</td>
<td>0.854</td>
<td>0.644</td>
</tr>
<tr>
<td>LSIS</td>
<td>0.350</td>
<td>0.508</td>
<td>0.421</td>
</tr>
</tbody>
</table>

Table 7: Results for English Twitter Mixed Polarity test set. The systems are ordered by their overall Kendall’s τ score, which was the official competition metric. The highest score is shown in bold.

<table>
<thead>
<tr>
<th>Team</th>
<th>Overall</th>
<th>Single words</th>
<th>Multi-word phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kendall’s τ</td>
<td>Spearman’s ρ</td>
<td>Kendall’s τ</td>
</tr>
<tr>
<td>iLab-Edinburgh</td>
<td>0.523</td>
<td>0.674</td>
<td>0.601</td>
</tr>
<tr>
<td>NileTMRG</td>
<td>0.475</td>
<td>0.591</td>
<td>0.384</td>
</tr>
<tr>
<td>LSIS</td>
<td>0.414</td>
<td>0.578</td>
<td>0.564</td>
</tr>
</tbody>
</table>

Table 8: Results for Arabic Twitter test set. The systems are ordered by their overall Kendall’s τ score, which was the official competition metric. The highest score is shown in bold.

<table>
<thead>
<tr>
<th>Team</th>
<th>Overall</th>
<th>Single words</th>
<th>Multi-word phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kendall’s τ</td>
<td>Spearman’s ρ</td>
<td>Kendall’s τ</td>
</tr>
<tr>
<td>iLab-Edinburgh</td>
<td>0.536</td>
<td>0.680</td>
<td>0.592</td>
</tr>
<tr>
<td>NileTMRG</td>
<td>0.475</td>
<td>0.658</td>
<td>0.510</td>
</tr>
<tr>
<td>LSIS</td>
<td>0.424</td>
<td>0.583</td>
<td>0.478</td>
</tr>
</tbody>
</table>

constituent words (Kiritchenko and Moham-mad, 2016d). For example, a positive adjective and a negative noun can form either a positive phrase (e.g., happy tears) or a negative phrase (e.g., great loss).

- The results achieved on the Arabic Twitter test set are substantially lower than the results achieved on a similar English Twitter data used in the 2015 competition.
- For most teams, the results obtained on single words are noticeably higher than the corresponding results on multi-word phrases. This is especially apparent on the Arabic Twitter data. The possible reason for this outcome is the lack of sufficient training data for phrases; none of the existing manually created English or Arabic real-valued sentiment lexicons provide annotations for multi-word phrases.

Overall, we observe strong correlations between the predicted and gold term rankings for terms in the general English domain as well as for single words in the other two domains. However, for multi-word phrases in the English Mixed Polarity set and Arabic Twitter set the correlations are markedly weaker, especially for the Arabic language. We hope that the availability of these datasets will foster further research towards automatic methods for sentiment composition in English and other languages.

7 Conclusions

We have created three sentiment composition lexicons that provide real-valued sentiment association scores for multi-word phrases and their constituent single words in three domains: the General English Sentiment Modifiers Set, the English Twitter Mixed Polarity Set, and the Arabic Twitter Set. The terms were annotated manually using the Best–Worst Scaling method of annotation. We included phrases composed of negators, modals, and degree adverbs—categories known to be challenging for sentiment analysis. Furthermore, we included phrases formed by words with opposing polarities. As future work, we would like to extend the task to cover more domains (e.g., biomedical, legal) and more languages. All datasets are freely available to the research community.
References


Svetlana Kiritchenko and Saif M. Mohammad. 2016b. The effect of negators, modals, and degree adverbs on sentiment composition. In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA).


