Noname manuscript No.
(will be inserted by the editor)

Developing a Successful SemEval Task in Sentiment Analysis of Twitter and Other Social Media Texts

Preslav Nakov · Sara Rosenthal · Svetlana Kiritchenko · Saif Mohammad · Zornitsa Kozareva · Alan Ritter · Veselin Stoyanov · Xiaodan Zhu

Received: date / Accepted: date

Abstract We present the development and evaluation of a semantic analysis task that lies at the intersection of two very trendy lines of research in contemporary computational linguistics: (*i*) sentiment analysis, and (*ii*) natural language processing of social media text. The task was part of SemEval, the International Workshop on Semantic Evaluation, a semantic evaluation forum previously known as SensEval.

P. Nakov
Qatar Computing Research Institute, HBKU Tornado Tower, floor 10, P.O. box 5825, Doha, Qatar
E-mail: pnakov@qf.org.qa

S. Rosenthal
Columbia University
E-mail: sara@cs.columbia.edu

S. Kiritchenko
National Research Council Canada, 1200 Montreal Rd., Ottawa, ON, Canada
E-mail: Svetlana.Kiritchenko@nrc-cnrc.gc.ca

S. Mohammad
National Research Council Canada, 1200 Montreal Rd., Ottawa, ON, Canada
E-mail: saif.mohammad@nrc-cnrc.gc.ca

Z. Kozareva
USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292-6695
E-mail: zornitsa@kozareva.com

A. Ritter
The Ohio State University
E-mail: aritter@cs.washington.edu

V. Stoyanov
Facebook
E-mail: vesko.st@gmail.com

X. Zhu
National Research Council Canada, 1200 Montreal Rd., Ottawa, ON, Canada
E-mail: Xiaodan.Zhu@nrc-cnrc.gc.ca

The task ran in 2013 and 2014, attracting the highest number of participating teams at SemEval in both years, and there is an ongoing edition in 2015. The task included the creation of a large contextual and message-level polarity corpus consisting of tweets, SMS messages, LiveJournal messages, and a special test set of sarcastic tweets. The evaluation attracted 44 teams in 2013 and 46 in 2014, who used a variety of approaches. The best teams were able to outperform several baselines by sizable margins with improvement across the two years the task has been run. We hope that the long-lasting role of this task and the accompanying datasets will be to serve as a test bed for comparing different approaches, thus facilitating research.

Keywords Sentiment analysis · Twitter · SemEval

1 Introduction

The Internet has democratized content creation enabling a number of new technologies, media and tools of communication, and ultimately leading to the rise of social media and an explosion in the availability of short informal text messages that are publicly available. Microblogs such as Twitter, weblogs such as LiveJournal, social networks such as Facebook, and instant messengers such as Skype and Whatsapp are now commonly used to share thoughts and opinions about anything in the surrounding world, along with the old-fashioned cell phone messages such as SMS. This proliferation of social media content has created new opportunities for studying public opinion, with Twitter being especially popular for research purposes due to its scale, representativeness, variety of topics discussed, as well as easy public access to its messages [27, 29, 37].

Despite all these opportunities, the rise of social media has also presented new challenges for natural language processing (NLP) applications, which had largely relied on NLP tools tuned for formal text genres such as newswire, and thus were not readily applicable to the informal language and style of social media. That language proved to be quite challenging with its use of creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, e.g., RT for re-tweet and #hashtags.[1] In addition to the genre difference, there is also a difference in length: social media messages are generally short, often length-limited by design as in Twitter, i.e., a sentence or a headline rather than a full document. How to handle such challenges has only recently been the subject of thorough research [3, 5, 14, 28, 36, 52, 53, 76].

The advance in NLP tools for processing social media text has enabled researchers to analyze people's *opinions and sentiments* on a variety of topics, especially in Twitter. Unfortunately, research in that direction was further hindered by the unavailability of suitable datasets and lexicons for system training, development and testing.

[1] Hashtags are a type of tagging for Twitter messages.

Until the rise of social media, research on opinion mining and sentiment analysis had focused primarily on learning about the language of sentiment in general, meaning that it was either genre-agnostic [2] or focused on newswire texts [80] and customer reviews (e.g., from web forums), most notably about movies [56] and restaurants, but also about hotels, digital cameras, cell phones, MP3 and DVD players [26], laptops, etc. This has given rise to several resources, mostly word and phrase polarity lexicons, which have proved to be very valuable for their respective domains and types of texts, but less useful for short social media messages such as tweets.

Over time, some Twitter-specific resources were developed, but initially they were either small and proprietary, such as the i-sieve corpus [36], were created only for Spanish like the TASS corpus [78], or relied on noisy labels obtained automatically based on emoticons and hashtags [22,45,46,56]. Moreover, they all focused on message-level sentiment only, instead of expression-level sentiment in the context of a tweet. In fact, the first large-scale freely available resource for sentiment analysis on Twitter were the datasets that we developed for SemEval-2013 task 2 [49], which we further extended for SemEval-2014 Task 9 [66] as well as for the upcoming SemEval-2015 Task 10 [65]. They offered both message-level and expression-level annotations.

The primary goal of these SemEval tasks was to serve as a test bed for comparing different approaches, thus facilitating research that will lead to a better understanding of how sentiment is conveyed in social media. These tasks have been highly successful, attracting wide interest at SemEval and beyond: they were the most popular SemEval tasks in both 2013 and 2014, attracting 44 and 46 participating teams, respectively, and they have further fostered the creation of additional freely available resources such as NRC's Hashtag Sentiment lexicon and the Sentiment140 lexicon [46], which the NRC team developed for their participation in SemEval-2013 task 2, and which were key for their winning the competition. Last but not least, even though named *Sentiment Analysis in Twitter*, the tasks also included evaluation on SMS and LiveJournal messages, as well as a special test set of sarcastic tweets.

In the remainder of this article, we first introduce the problem of contextual and message-level polarity classification (Section 2). We then describe the process of creating the training and the testing datasets (Section 3) and the evaluation setup (Section 4). Afterwards, we list and briefly discuss the participating systems, the results, and the lessons learned (Sections 5 and 6). Finally, we compare the task to other related efforts (Section 7), and we point to possible directions for future research (Section 9).

2 Task Description

SemEval-2013 task 2 [49] and SemEval-2014 Task 9 [66] had two subtasks: an expression-level subtask and a message-level subtask. Participants could choose to participate in either or both. Below we provide short descriptions of the objectives of these two subtasks.

Subtask A: Contextual Polarity Disambiguation: Given a message containing a marked instance of a word or a phrase, determine whether that instance is positive, negative or neutral in that context. The instance boundaries were provided: this was a classification task, not an entity recognition task.

Subtask B: Message Polarity Classification: Given a message, decide if it is of positive, negative, or neutral sentiment. For messages conveying both positive and negative sentiment, the stronger one is to be chosen.

Each participating team was allowed to submit results for two different systems per subtask: one constrained, and one unconstrained. A constrained system could only use the provided data for training, but it could also use other resources such as lexicons obtained elsewhere. An unconstrained system could use any additional data as part of the training process; this could be done in a supervised, semi-supervised, or unsupervised fashion.

Note that constrained/unconstrained refers to the data used to train a classifier. For example, if other data (excluding the test data) was used to develop a sentiment lexicon, and the lexicon was used to generate features, the system would still be constrained. However, if other, manually or automatically labeled data (excluding the test data) was used with the original data to train the classifier, then such a system would be considered unconstrained.[2]

3 Dataset Creation

In this section, we describe the process of collecting and annotating our datasets of short social media text messages. We will focus our discussion on general tweets as collected for SemEval-2013 Task 2, but our testing datasets also include sarcastic tweets, SMS messages and sentences from LiveJournal, which we will also describe.

[2] We should note that the distinction between constrained and unconstrained systems is quite subtle. For example, the creation of a dedicated lexicon obtained from other annotated data could be regarded by someone as a form of supervision beyond the dataset provided in the task. A similar argument could be also made about various NLP tools for Twitter processing such as Noah's ARK *Tweet NLP*, Alan Ritter's *twitter_nlp*, or GATE's *TwitIE*, which are commonly used for tweet tokenization, normalization, POS tagging [21], chunking, syntactic parsing [35], named entity recognition [62], information extraction [6], and event extraction [63]; all these tools are trained on additional tweets. Indeed, some participants in 2013 and 2014 did not understand well the constrained vs. unconstrained distinction, and we had to check the system descriptions, and to reclassify some submissions as constrained/unconstrained. This was a hard and tedious job, and thus for the 2015 edition of the task, we did not make a distinction between constrained and unconstrained systems, letting the participants to use any additional data, resources and tools they wished to. In any case, our constrained/unconstrained definition for the 2013 and 2014 editions of the task are clear, and the system descriptions for the individual systems are also available. Thus, researchers are free to see the final system ranking any way they like, e.g., as two separate constrained vs. unconstrained rankings or as one common ranking.

3.1 Data Collection

First, we gathered tweets that express sentiment about popular topics. For this purpose, we extracted named entities using a Twitter-tuned NER system [62] from millions of tweets, which we collected over a one-year period spanning from January 2012 to January 2013; for downloading, we used the public streaming Twitter API.

We then identified popular topics as those named entities that are frequently mentioned in association with a specific date [63]. Given this set of automatically identified topics, we gathered tweets from the same time period which mentioned the named entities. The testing messages had different topics from training and spanned later periods; this is true for both Twitter2013-test, which used tweets from later in 2013, and Twitter2014-test, which included tweets from 2014.

The collected tweet data were greatly skewed towards the neutral class. In order to reduce the class imbalance, we removed messages that contained no sentiment-bearing words using SentiWordNet as a repository of sentiment words. Any word listed in SentiWordNet 3.0 with at least one sense having a positive or a negative sentiment score greater than 0.3 was considered a sentiment-bearing word.[3]

We annotated the same Twitter messages for subtask A and subtask B. However, the final training and testing datasets overlap only partially between the two subtasks since we had to discard messages with low inter-annotator agreement, and this differed between the subtasks.

After the annotation process, we split the annotated tweets into training, development and testing datasets; for testing, we further annotated three additional out-of-domain datasets:[4]

- **SMS messages:** from the NUS SMS corpus[5] [10];
- **LiveJournal:** sentences from LiveJournal [64];
- **Sarcastic tweets:** a small set of tweets containing the #sarcasm hashtag.

3.2 Annotation Process

Our datasets were annotated for sentiment on Mechanical Turk.[6] Each sentence was annotated by five Mechanical Turk workers, also known as Turkers.

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

[4] We pre-filtered the SMS messages and the sarcastic tweets with SentiWordNet, but we did not do it for LiveJournal sentences.

[5] `http://wing.comp.nus.edu.sg/SMSCorpus/`

[6] The use of Amazon's Mechanical Turk has been criticised from an ethical (e.g., human exploitation) and a legal (e.g., tax evasion, minimal legal wage in some countries, absence of a work contract) perspective; see [19] for a broader discussion. We have tried our best to stay fair, adjusting the pay per HIT in such a way that the resulting hourly rate be on par with what is currently considered good pay on Mechanical Turk. Indeed, Turkers were eager to work on our HITs, and the annotations were completed quickly.

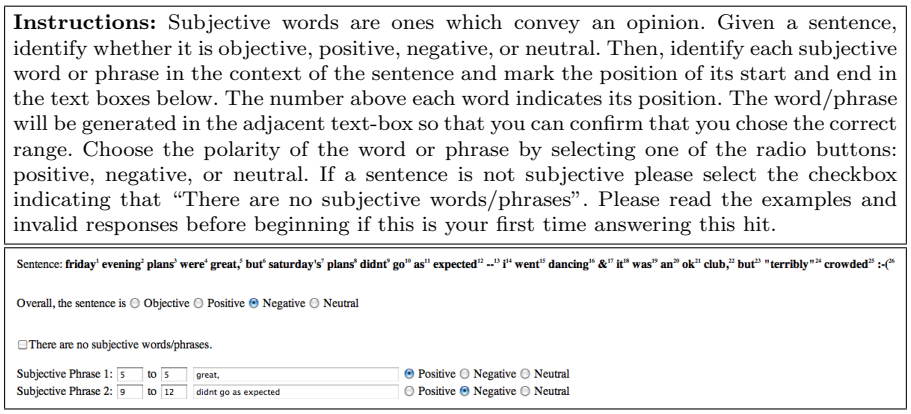

Fig. 1 Instructions given to workers on Mechanical Turk, followed by a screenshot.

The annotations for subtask A and subtask B were done concurrently. Each Turker had to mark all the subjective words/phrases in the tweet message by indicating their start and end positions and to say whether each subjective word/phrase was positive, negative, or neutral (subtask A). Turkers also had to indicate the overall polarity of the message (subtask B). The instructions we gave to the Turkers, along with an example, are shown in Figure 1. Several additional examples (Table 1) were also available to the annotators.

Providing all the required annotations for a given message (a tweet, an SMS, or a sentence from LiveJournal) constituted a Human Intelligence Task, or a HIT. In order to qualify for the task, a Turker had to have an approval rate greater than 95%, and should have completed 50 approved HITs. We further discarded the following types of annotations:[7]

- messages containing overlapping subjective phrases;
- messages marked as subjective but having no annotated subjective phrases;
- messages with every single word marked as subjective;
- messages with no overall sentiment marked.

For each message, the annotations provided by several Turkers were combined as follows. For subtask A, we combined the annotations using intersection as shown in the last row of Table 2. A word had to appear in 2/3 of the annotations in order to be considered subjective. Similarly, a word had to be labeled with a particular polarity (positive, negative, or neutral) 2/3 of the time in order to receive that label. We also experimented with other methods of combining annotations: (1) by computing the union of the annotations for the sentence, and (2) by taking the annotations provided by the worker who annotated the most HITs. However, we found that these methods were not as accurate. We plan to explore further alternatives in future work, e.g., using the MACE adjudication method [25].

[7] Note that this discarding only happened if a single Turker had created contradictory annotations; it was not done at the adjudication stage.

Authorities are *only too aware* that Kashgar is 4,000 kilometres (2,500 miles) from Beijing but *only* a tenth of the distance from the Pakistani border, and are *desperate* to *ensure instability or militancy* does not leak over the frontiers.
Taiwan-made products *stood a good chance* of becoming *even more competitive thanks to* wider access to overseas markets and lower costs for material imports, he said.
"March *appears* to be a *more reasonable* estimate while earlier admission *cannot be entirely ruled out*," according to Chen, also Taiwan's chief WTO negotiator.
friday evening plans were great, but saturday's plans *didnt go as expected* – i went dancing & it was an *ok* club, but *terribly crowded :-(*
WHY THE *HELL* DO YOU GUYS ALL HAVE MRS. KENNEDY! SHES A FUCKING DOUCHE
AT&T was *okay* but whenever they do something *nice* in the name of customer service it seems like a favor, while T-Mobile makes that a *normal everyday thin*
obama should be *impeached* on *TREASON* charges. Our Nuclear arsenal was TOP Secret. Till HE told our enemies what we had. *#Coward #Traitor*
My graduation speech: I'd like to *thanks* Google, Wikipedia and my computer! *:D* #iThingteens

Table 1 List of example sentences with annotations that were provided to the Turkers. All subjective phrases are italicized. Positive phrases are in green, negative phrases are in red, and neutral phrases are in blue.

I would love to watch Vampire Diaries ***:)*** and some Heroes! ***Great combination***	9/13
I would love to watch Vampire Diaries :) and some ***Heroes! Great*** combination	11/13
I ***would love*** to watch Vampire Diaries ***:)*** and some Heroes! ***Great*** combination	10/13
I would ***love*** to watch Vampire Diaries :) and some Heroes! ***Great*** combination	13/13
I would love to watch Vampire Diaries :) and some Heroes! ***Great*** combination	12/13
I would ***love*** to watch Vampire Diaries :) and some Heroes! ***Great*** combination	

Table 2 Example of a sentence annotated for subjectivity on Mechanical Turk. The words and the phrases that were marked as subjective are italicized and highlighted in bold. The first five rows show annotations provided by the Turkers, and the final row shows their intersection. The final column shows the accuracy for each annotation compared with respect to the intersection in the final row.

For subtask B, the polarity of the entire sentence was determined based on the majority of the labels. If there was a tie, the sentence was discarded (these are likely to be controversial cases). In order to reduce the number of rejected sentences, we combined the objective and the neutral labels, which Turkers tended to mix up.

For the sarcastic tweets, we slightly altered the annotation task. The tweets were shown to the Turkers without the #sarcasm hashtag, and the Turkers were asked to determine whether the tweet was sarcastic on its own. Furthermore, the Turkers had to indicate the degree of sarcasm as (a) definitely sarcastic, (b) probably sarcastic, and (c) not sarcastic. Although we do not use the degree of sarcasm at this time, it could be useful for analysis as well as possibly excluding tweets that do not appear to be sarcastic. For the SMS and the LiveJournal messages, the annotation task was the same as for tweets, but without the annotations for sarcasm.

Corpus	Positive	Negative	Objective / Neutral	Total
Twitter2013-train	5,895	3,131	471	9,497
Twitter2013-dev	648	430	57	1,135
Twitter2013-test	2,734	1,541	160	4,435
SMS2013-test	1,071	1,104	159	2,334
Twitter2014-test	1,807	578	88	2,473
Twitter2014-sarcasm	82	37	5	124
LiveJournal2014-test	660	511	144	1,315

Table 3 Dataset statistics for Subtask A.

The obtained annotations were used as gold labels for the corresponding subtasks. Consecutive tokens marked as subjective serve as target terms in subtask A. The statistics for all datasets are shown in Tables 3 and 4 for subtask A and B, respectively. Each dataset is marked with the year of the SemEval edition it was produced for. An annotated example from each source is shown in Table 5.

When building a system to solve a task, it is good to know how well we should expect it to perform. One good reference point is human performance and agreement between annotators. Unfortunately, as we derive annotations by agreement, we cannot calculate standard statistics such as Kappa directly. Instead, we decided to measure the agreement between our gold standard annotations (derived by agreement) and the annotations proposed by the best Turker, the worst Turker, and the average Turker (with respect to the gold/consensus annotation for a particular message). Given a HIT, we just calculate the overlaps as shown in the last column in Table 2, and then we calculate the best, the worst, and the average, which are respectively 13/13, 9/13 and 11/13, in the example. Finally, we average these statistics over all HITs that contributed to a given dataset, to produce lower, average, and upper averages for that dataset. The accuracy (with respect to the gold/consensus annotation) for different averages is shown in Table 6. Since the overall polarity of a message is chosen based on majority, the upper bound for subtask B is 100%. These averages give a good indication about how well we can expect the systems to perform. For example, we can see that even if we used the best annotator for each HIT, it would still not be possible to get perfect accuracy, and thus we should also not expect perfect accuracy for an automatic system.

3.3 Tweet Delivery

Due to Twitter's terms of service, we could not deliver the annotated tweets to the participants directly. Instead, we released annotation indexes and labels, a list of corresponding Twitter IDs, and a download script[8] that extracts the corresponding tweets via the Twitter API.

[8] `https://github.com/aritter/twitter_download`

Corpus	Positive	Negative	Objective / Neutral	Total
Twitter2013-train	3,662	1,466	4,600	9,728
Twitter2013-dev	575	340	739	1,654
Twitter2013-test	1,572	601	1,640	3,813
SMS2013-test	492	394	1,207	2,093
Twitter2014-test	982	202	669	1,853
Twitter2014-sarcasm	33	40	13	86
LiveJournal2014-test	427	304	411	1,142

Table 4 Dataset statistics for Subtask B.

Source	Message	Message-Level Polarity
Twitter	Why would you [still]- wear shorts when it's this cold?! I [love]+ how Britain see's a bit of sun and they're [like 'OOOH]+ LET'S STRIP!'	positive
SMS	[Sorry]- I think tonight [cannot]- and I [not feeling well]- after my rest.	negative
LiveJournal	[Cool]+ posts , dude ; very [colorful]+ , and [artsy]+ .	positive
Twitter Sarcasm	[Thanks]+ manager for putting me on the schedule for Sunday	negative

Table 5 Example annotations for each source of messages. The target terms (i.e., subjective phrases) are marked in [...], and are followed by their polarity (subtask A); the message-level polarity is shown in the last column (subtask B).

Corpus	Subtask A			Subtask B
	Lower	Avg.	Upper	Avg.
Twitter2013-train	75.1	89.7	97.9	77.6
Twitter2013-dev	66.6	85.3	97.1	86.4
Twitter2013-test	76.8	90.3	98.0	75.9
SMS2013-test	75.9	97.5	89.6	77.5
Livejournal2014-test	61.7	82.3	94.5	76.2
Twitter2014-test	75.3	88.9	97.5	74.7
2014-test	62.6	83.1	95.6	71.2

Table 6 Average (over all HITs) overlap of the gold annotations with the worst, average, and the worst Turker for each HIT, for subtasks A and B.

As a result, the task participants had access to different number of training tweets depending on when they did the downloading,[9] as over time some tweets were deleted. Another major reason for tweet unavailability was Twitter users changing the status of their accounts from public to private. Note that this account status change goes in both directions and changes can be made frequently; thus, some task participants could actually download more tweets by trying several times on different dates.

[9] However, this did not have major impact on the results; see Section 6.3 for detail.

4 Scoring

The participating systems were required to perform a three-way classification for both subtasks. A particular marked phrase (for subtask A) or an entire message (for subtask B) was to be classified as *positive*, *negative* or *objective/neutral*. We evaluated the systems by computing a score for predicting positive/negative phrases/messages. For instance, to compute positive precision, P_{pos}, we find the number of phrases/messages that a system correctly predicted to be positive, and we divide that number by the total number it predicted to be positive. To compute positive recall, R_{pos}, we find the number of phrases/messages correctly predicted to be positive and we divide that number by the total number of positives in the gold standard. We then calculate F_1-score for the positive class as follows $F_{pos} = \frac{2P_{pos}R_{pos}}{P_{pos}+R_{pos}}$. We carry out similar computations for the negative phrases/messages, F_{neg}. The overall score is then the average of the F_1-scores for the positive and negative classes: $F = (F_{pos} + F_{neg})/2$.

We provided the participants with a scorer that outputs the overall score F, as well as P, R, and F_1 scores for each class (positive, negative, neutral) and for each test set.

5 Participants and Results

In the first edition of the task (SemEval-2013), there were 28 submissions by 23 teams for subtask A, and 51 submissions by 38 teams for subtask B; a total of 44 teams took part in the task overall. In the second year (SemEval-2014), the task again attracted a high number of participants: there were 27 submissions by 21 teams for subtask A, and 50 submissions by 44 teams for subtask B, a total of 46 different teams.[10] Eighteen teams participated in both years.

Most of the submissions were constrained, with just a few unconstrained. In any case, the best systems were constrained both years. Some teams participated with both a constrained and an unconstrained system, but the unconstrained system was not always better than the constrained one. There was a single ranking, which included both constrained and unconstrained systems, where the latter were marked accordingly.

5.1 Systems

Algorithms: In both years, most systems were supervised and used a variety of handcrafted features derived from *n*-grams, stems, punctuation, part-of-speech (POS) tags, and Twitter-specific encodings such as emoticons, hashtags, and abbreviations. The most popular classifiers included Support Vector Machines (SVM), Maximum Entropy (MaxEnt), and Naïve Bayes.

[10] In the ongoing third year of the task (SemEval-2015), there were submission by 41 teams: 11 teams participated in subtask A, 40 in subtask B [65].

Notably, only one of the top-performing systems in 2013, *teragram* [61] (SAS Institute, USA), was entirely rule-based, and fully relied on hand-written rules. We should also mention the emerging but quite promising approach of applying deep learning, as exemplified by the top-performing SemEval-2014 teams of *coooolll* [74] (Harbin Institute of Technology and Microsoft Research China) and *ThinkPositive* [68] (IBM Research Brazil).[11]

Preprocessing: In addition to standard NLP steps such as tokenization, stemming, lemmatization, stop-word removal and POS tagging, most teams applied some kind of Twitter-specific processing such as substitution/removal of URLs, substitution of emoticons, spelling correction, word normalization, abbreviation lookup, and punctuation removal. Several teams reported using Twitter-tuned NLP tools such as POS and named entity taggers [21,62].

External Lexical Resources: Many systems relied heavily on existing sentiment lexicons. Sentiment lexicons are lists of words (and sometimes phrases) with prior associations to positive, negative, and sometimes neutral sentiment. Some lexicons provide a real-valued or a discrete sentiment score for a term to indicate its intensity. Most of the lexicons that were created by manual annotation tend to be domain-independent and include a few thousand terms, but larger lexicons can be built automatically or semi-automatically. The most popular lexicons used by participants in both years included the manually created MPQA Subjectivity Lexicon [81], Bing Liu's Lexicon [26], as well as the automatically created SentiWordNet [2]. The winning team at SemEval-2013, *NRC-Canada* [46], reported huge gains from their automatically created high-coverage tweet-specific sentiment lexicons (Hashtag Sentiment Lexicon and Sentiment140 lexicon).[12] They also used the NRC Emotion Lexicon [47,48] and the Bing Liu Lexicon [26]. The NRC lexicons were released to the community, and were used by many teams in the subsequent editions of the SemEval Twitter sentiment task.

In addition to using sentiment lexicons, many top-performing systems used word representations built from large external collections of tweets or other corpora. Such representations serve to reduce the sparseness of the word space. Two general approaches for building word representations are word clustering and word embeddings. The Brown clustering algorithm [8] groups syntactically or semantically close words in a hierarchy of clusters. The CMU Twitter NLP tool provides word clusters produced with the Brown clustering algorithm on 56 million English-language tweets. Recently, several deep learning algorithms have been proposed to build continuous dense word representations, called word embeddings [12,43]. Similar to word clusters, syntactically or semantically close words should have similar embedding vectors. The pre-trained word embeddings are publicly available,[13] but they were generated from news articles. Therefore, some teams chose to train their own word embeddings on tweets using the available software word2vec [43].

[11] Neural nets and deep learning were also used by top-performing teams in 2015, e.g., by UNITN [69] (University of Trento and Qatar Computing Research Institute).

[12] `http://www.purl.com/net/lexicons`

[13] `https://code.google.com/p/word2vec/`

The *coooolll* team [74] (Harbin Institute of Technology and Microsoft Research China) went one step further and produced sentiment-specific word embeddings. They extended the neural network C&W model [12] to incorporate the sentiment information on sentences and modified the loss function to be a linear combination of syntactic loss and sentiment loss. Similarly, at SemEval-2015, the UNITN team [69] used an unsupervised neural language model to initialize word embeddings that they further tuned by a deep learning model using a separate corpus and distant supervision; they then continued training in a supervised way on the SemEval data.

Further details on individual systems can be found in the proceedings of SemEval-2013 [41], SemEval-2014 [50], and SemEval-2015 [51].

5.2 Baselines

There are several baselines that one might consider for this task, and below we will explore some of the most interesting ones.

Majority Class. This baseline always predicts the most frequent class as observed on the training dataset. As our official evaluation metric is an average of the F-score for the positive and for the negative classes, it makes sense to consider these two classes only. For our training dataset, this baseline predicts the positive class for both subtasks A and B as it is more frequent for both subtasks.

Target's Majority Class. This baseline is only applicable to subtask A. For that subtask, we can calculate the majority class for individual target terms. If a target term (a word or a phrase) from the test set occurs as target in the training dataset, this baseline predicts the most frequent class for that term. If the frequencies tie between two classes, the priority in the order of positive, negative, and neutral is used to break the tie. For example, if a term appears the same number of times as positive and as negative in the training dataset, we predict positive class for the term. If a target term does not occur in the training data, we predict the most frequent class from the entire training dataset, i.e., the positive class.

Lexicon-based. We add up the scores for lexicon words or phrases matched in the target term (for subtask A) or in the entire message (for subtask B), and we predict a positive class if the cumulative sum is greater than zero, a neutral class if it is zero, and a negative class if it is less than zero. If no matches are found, we predict neutral. We calculate this baseline using three different sentiment lexicons: MPQA Subjectivity Lexicon, Bing Liu's Lexicon, and SentiWordNet 3.0. We use a score of 1 for a positive entry and a score of -1 for a negative entry in the MPQA and Bing Liu's lexicons. As SentiWordNet has a real-valued positive score and a real-valued negative score assigned to a word sense, for it we average positive and negative scores over all senses of a word and we subtract the average negative score from the average positive score to get the final sentiment score for the target word or phrase.

Baseline	2013		2014		
	Tweet	**SMS**	**Tweet**	**Tweet sarcasm**	**Live-Journal**
Subtask A					
Majority Class	38.10	31.50	42.22	39.81	33.42
Target's Majority Class	71.62	68.60	72.52	60.86	55.33
Lexicon-based					
MPQA	55.57	54.43	50.69	48.22	59.38
Bing Liu's	58.88	49.89	50.07	45.72	59.21
SentiWordNet	64.16	68.69	60.38	50.41	73.44
SVM unigrams	83.56	81.50	80.57	77.33	78.78
SVM unigrams+bigrams	83.82	81.71	81.03	76.95	77.96
Subtask B					
Majority Class	29.19	19.03	34.64	27.73	27.21
Lexicon-based					
MPQA	46.21	47.17	46.09	33.68	55.49
Bing Liu's	53.59	53.32	49.96	31.67	61.09
SentiWordNet	45.43	43.55	44.85	46.66	56.49
SVM unigrams	56.95	54.21	58.58	47.71	59.47
SVM unigrams+bigrams	57.59	53.81	58.14	48.40	57.47

Table 7 The macro-averaged F-scores for different baselines.

SVM unigrams. This is a more sophisticated baseline, which trains a Support Vector Machine classifier on the training dataset, using unigrams as features. In the experiments, we used the LibSVM package [9] with linear kernel and a value of the C parameter that we optimized on the development dataset.

SVM unigrams+bigrams. This baseline is similar to the previous one with the exception that the feature set now includes unigrams and bigrams.

Table 7 shows the macro-averaged F-scores for different baselines. First, note that for almost all baselines the scores for subtask A are substantially higher than the corresponding scores for subtask B. Second, we can see that for subtask A the Target's Majority Class baseline and the SVM unigrams baseline achieve remarkable results: by simply predicting a target's majority class one can obtain F-scores in the low seventies, and by training an SVM model with only unigram features one can get F-scores in the low eighties. For comparison, for subtask B the SVM unigrams baseline only achieves F-scores in the fifties. We explore the differences between the subtasks and the corresponding datasets in more detail in Section 6.7.

Overall, for both subtasks, statistical machine learning yields stronger baseline results than simple lexicon-based methods. Therefore, it is not surprising that most participants relied on statistical learning from the training dataset and used sentiment lexicons to obtain additional features.

Finally, note that most baselines perform badly on sarcastic tweets, even though the Majority Class baseline score on this dataset does not significantly differ from the corresponding scores on the other test datasets.

#	System	2013: Progress		2014: Official		
		Tweet	SMS	Tweet	Tweet sarcasm	Live-Journal
1	NRC-Canada	90.14_{1}	88.03_{4}	86.63_{1}	77.13_{5}	85.49_{2}
2	SentiKLUE	90.11_{2}	85.16_{8}	84.83_{2}	79.32_{3}	85.61_{1}
3	CMUQ-Hybrid	88.94_{4}	87.98_{5}	84.40_{3}	76.99_{6}	84.21_{3}
4	CMU-Qatar	89.85_{3}	88.08_{3}	83.45_{4}	78.07_{4}	83.89_{5}
5	ECNU (*)	87.29_{6}	89.26_{2}	82.93_{5}	73.71_{8}	81.69_{7}
6	ECNU	87.28_{7}	89.31_{1}	82.67_{6}	73.71_{9}	81.67_{8}
7	Think_Positive (*)	88.06_{5}	87.65_{6}	82.05_{7}	76.74_{7}	80.90_{12}
8	Kea	84.83_{10}	84.14_{10}	81.22_{8}	65.94_{17}	81.16_{11}
9	Lt_3	86.28_{8}	85.26_{7}	81.02_{9}	70.76_{13}	80.44_{13}
10	senti.ue	84.05_{11}	78.72_{16}	80.54_{10}	82.75_{1}	81.90_{6}
11	LyS	85.69_{9}	81.44_{12}	79.92_{11}	71.67_{10}	83.95_{4}
12	UKPDIPF	80.45_{15}	79.05_{14}	79.67_{12}	65.63_{18}	81.42_{9}
13	UKPDIPF (*)	80.45_{16}	79.05_{15}	79.67_{13}	65.63_{19}	81.42_{10}
14	TJP	81.13_{14}	84.41_{9}	79.30_{14}	71.20_{12}	78.27_{15}
15	SAP-RI	80.32_{17}	80.26_{13}	77.26_{15}	70.64_{14}	77.68_{18}
16	senti.ue (*)	83.80_{12}	82.93_{11}	77.07_{16}	80.02_{2}	79.70_{14}
17	SAIL	78.47_{18}	74.46_{20}	76.89_{17}	65.56_{20}	70.62_{22}
18	columbia_nlp	81.50_{13}	74.55_{19}	76.54_{18}	61.76_{22}	78.19_{16}
19	IIT-Patna	76.54_{20}	75.99_{18}	76.43_{19}	71.43_{11}	77.99_{17}
20	Citius (*)	76.59_{19}	69.31_{21}	75.21_{20}	68.40_{15}	75.82_{20}
21	Citius	74.71_{21}	61.44_{25}	73.03_{21}	65.18_{21}	71.64_{21}
22	IITPatna	70.91_{23}	77.04_{17}	72.25_{22}	66.32_{16}	76.03_{19}
23	SU-sentilab	74.34_{22}	62.58_{24}	68.26_{23}	53.31_{25}	69.53_{23}
24	Univ. Warwick	62.25_{26}	60.12_{26}	67.28_{24}	58.08_{24}	64.89_{25}
25	Univ. Warwick (*)	64.91_{25}	63.01_{23}	67.17_{25}	60.59_{23}	67.46_{24}
26	DAEDALUS	67.42_{24}	63.92_{22}	60.98_{26}	45.27_{27}	61.01_{26}
27	DAEDALUS (*)	61.95_{27}	55.97_{27}	58.11_{27}	49.19_{26}	58.65_{27}

Table 8 Results for subtask A. The systems are sorted by their score on the Twitter2014 test dataset; the rankings on the individual datasets are indicated with a subscript. The (*) indicates an unconstrained submission.

5.3 Results

The results for the 2014 edition of the task are shown in Tables 8 and 9, and the corresponding team affiliations are shown in Table 11. The tables show results on the two progress test datasets (tweets and SMS messages), which are the official test datasets from the 2013 edition of the task, and on the three official 2014 test sets (tweets, tweets with sarcasm, and LiveJournal). There is an index for each result showing its relative rank within the respective column. The systems are ranked by their score on the Twitter-2014 test set, which is the official ranking for the task; all remaining rankings are secondary.

5.3.1 Subtask A

Table 8 shows the results for subtask A, which attracted 27 submissions from 21 teams at SemEval-2014. There were seven unconstrained submissions: five teams submitted both a constrained and an unconstrained run, and two teams submitted an unconstrained run only. The best systems were constrained.

#	System	2013: Progress		2014: Official		
		Tweet	SMS	Tweet	Tweet sarcasm	Live-Journal
1	TeamX	72.12_{1}	57.36_{26}	70.96_{1}	56.50_{3}	69.44_{15}
2	coooolll	70.40_{3}	67.68_{2}	70.14_{2}	46.66_{24}	72.90_{5}
3	RTRGO	69.10_{5}	67.51_{3}	69.95_{3}	47.09_{23}	72.20_{6}
4	NRC-Canada	70.75_{2}	70.28_{1}	69.85_{4}	58.16_{1}	74.84_{1}
5	TUGAS	65.64_{13}	62.77_{11}	69.00_{5}	52.87_{12}	69.79_{13}
6	CISUC_KIS	67.56_{8}	65.90_{6}	67.95_{6}	55.49_{5}	74.46_{2}
7	SAIL	66.80_{11}	56.98_{28}	67.77_{7}	57.26_{2}	69.34_{17}
8	SWISS-CHOCOLATE	64.81_{18}	66.43_{5}	67.54_{8}	49.46_{16}	73.25_{4}
9	Synalp-Empathic	63.65_{23}	62.54_{12}	67.43_{9}	51.06_{15}	71.75_{9}
10	Think_Positive (*)	68.15_{7}	63.20_{9}	67.04_{10}	47.85_{21}	66.96_{24}
11	SentiKLUE	69.06_{6}	67.40_{4}	67.02_{11}	43.36_{30}	73.99_{3}
12	JOINT_FORCES (*)	66.61_{12}	62.20_{13}	66.79_{12}	45.40_{26}	70.02_{12}
13	AMI_ERIC	70.09_{4}	60.29_{20}	66.55_{13}	48.19_{20}	65.32_{26}
14	AUEB	63.92_{21}	64.32_{8}	66.38_{14}	56.16_{4}	70.75_{11}
15	CMU-Qatar	65.11_{17}	62.95_{10}	65.53_{15}	40.52_{38}	65.63_{25}
16	Lt_3	65.56_{14}	64.78_{7}	65.47_{16}	47.76_{22}	68.56_{20}
17	columbia_nlp	64.60_{19}	59.84_{21}	65.42_{17}	40.02_{40}	68.79_{19}
18	LyS	66.92_{10}	60.45_{19}	64.92_{18}	42.40_{33}	69.79_{14}
19	NILC_USP	65.39_{15}	61.35_{16}	63.94_{19}	42.06_{34}	69.02_{18}
20	senti.ue	67.34_{9}	59.34_{23}	63.81_{20}	55.31_{6}	71.39_{10}
21	UKPDIPF	60.65_{29}	60.56_{17}	63.77_{21}	54.59_{7}	71.92_{7}
22	UKPDIPF (*)	60.65_{30}	60.56_{18}	63.77_{22}	54.59_{8}	71.92_{8}
23	SU-FMI	60.96_{28}	61.67_{15}	63.62_{23}	48.34_{19}	68.24_{21}
24	ECNU	62.31_{27}	59.75_{22}	63.17_{24}	51.43_{14}	69.44_{16}
25	ECNU (*)	63.72_{22}	56.73_{29}	63.04_{25}	49.33_{17}	64.08_{31}
26	Rapanakis	58.52_{32}	54.02_{35}	63.01_{26}	44.69_{27}	59.71_{37}
27	Citius (*)	63.25_{24}	58.28_{24}	62.94_{27}	46.13_{25}	64.54_{29}
28	CMUQ-Hybrid	63.22_{25}	61.75_{14}	62.71_{28}	40.95_{37}	65.14_{27}
29	Citius	62.53_{26}	57.69_{25}	61.92_{29}	41.00_{36}	62.40_{33}
30	KUNLPLab	58.12_{33}	55.89_{31}	61.72_{30}	44.60_{28}	63.77_{32}
31	senti.ue (*)	65.21_{16}	56.16_{30}	61.47_{31}	54.09_{9}	68.08_{22}
32	UPV-ELiRF	63.97_{20}	55.36_{33}	59.33_{32}	37.46_{42}	64.11_{30}
33	USP_Biocom	58.05_{34}	53.57_{36}	59.21_{33}	43.56_{29}	67.80_{23}
34	DAEDALUS (*)	58.94_{31}	54.96_{34}	57.64_{34}	35.26_{44}	60.99_{35}
35	IIT-Patna	52.58_{40}	51.96_{37}	57.25_{35}	41.33_{35}	60.39_{36}
36	DejaVu	57.43_{36}	55.57_{32}	57.02_{36}	42.46_{32}	64.69_{28}
37	GPLSI	57.49_{35}	46.63_{42}	56.06_{37}	53.90_{10}	57.32_{41}
38	BUAP	56.85_{37}	44.27_{44}	55.76_{38}	51.52_{13}	53.94_{44}
39	SAP-RI	50.18_{44}	49.00_{41}	55.47_{39}	48.64_{18}	57.86_{40}
40	UMCC_DLSI_Sem	51.96_{41}	50.01_{38}	55.40_{40}	42.76_{31}	53.12_{45}
41	IBM_EG	54.51_{38}	46.62_{43}	52.26_{41}	34.14_{46}	59.24_{38}
42	Alberta	53.85_{39}	49.05_{40}	52.06_{42}	40.40_{39}	52.38_{46}
43	lsis_lif	46.38_{46}	38.56_{47}	52.02_{43}	34.64_{45}	61.09_{34}
44	SU-sentilab	50.17_{45}	49.60_{39}	49.52_{44}	31.49_{47}	55.11_{42}
45	SINAI	50.59_{42}	57.34_{27}	49.50_{45}	31.15_{49}	58.33_{39}
46	IITPatna	50.32_{43}	40.56_{46}	48.22_{46}	36.73_{43}	54.68_{43}
47	Univ. Warwick	39.17_{48}	29.50_{49}	45.56_{47}	39.77_{41}	39.60_{49}
48	UMCC_DLSI_Graph	43.24_{47}	36.66_{48}	45.49_{48}	53.15_{11}	47.81_{47}
49	Univ. Warwick (*)	34.23_{50}	24.63_{50}	45.11_{49}	31.40_{48}	29.34_{50}
50	DAEDALUS	36.57_{49}	40.86_{45}	33.03_{50}	28.96_{50}	40.83_{48}

Table 9 Results for subtask B. The systems are sorted by their score on the Twitter2014 test dataset; the rankings on the individual datasets are indicated with a subscript. The (*) indicates an unconstrained submission.

Comparing Table 8 to Table 7, we can see that all participating systems outperformed the Majority Class baseline by a sizable margin. However, some systems could not beat the Target's Majority Class baseline, and most systems could not compete against the SVM-based baselines.

5.3.2 Subtask B

The results for subtask B are shown in Table 9. The subtask attracted 50 submissions from 44 teams at SemEval-2014. There were eight unconstrained submissions: six teams submitted both a constrained and an unconstrained run, and two teams submitted an unconstrained run only. As for subtask A, the best systems were constrained.

Comparing Table 9 to Table 7, we see that almost all participating systems outperformed the Majority Class baseline, but some ended up performing slightly lower on some of the datasets. Moreover, several systems could not beat the remaining stronger baselines; in particular, about a third of the systems could not compete against the SVM-based baselines.

6 Analysis

In this section, we analyze the results from several perspectives. In particular, we discuss the progress over the first two years of the SemEval task, the system independence of the training domain, the need for external lexical resources, the impact of different techniques for handling negation and context, and the differences between the two subtasks.

6.1 Progress over the First Two Years

As Table 11 shows, 18 out of the 46 teams in 2014 also participated in the 2013 edition of the task. Comparing the results on the progress Twitter test dataset in 2013 [49] and 2014 [66], we can see that *NRC-Canada*, the 2013 winner for subtask A, has now improved their F-score from 88.93 [46] to 90.14 [84], which is the 2014 best. The best score on the progress SMS test dataset in 2014 of 89.31 belongs to *ECNU* [82]; this is a big jump compared to their 2013 score of 76.69 [75], but it is lower compared to the 2013 best of 88.37 achieved by *GU-MLT-LT* [23].

For subtask B, on the Twitter progress test dataset, the 2013 winner, *NRC-Canada*, improves their 2013 result from 69.02 [46] to 70.75 [84], which is the second best in 2014; the winner in 2014, *TeamX*, achieves 72.12 [44]. On the SMS progress test set, the 2013 winner, *NRC-Canada*, improves its F-score from 68.46 to 70.28. Overall, we see consistent improvements on the progress test datasets for both subtasks: 0-1 and 2-3 points absolute for subtasks A and B, respectively.

For both subtasks, the best systems on the Twitter2014-test dataset are those that performed best on the progress Twitter2013-test dataset: *NRC-Canada* for subtask A, and *TeamX* (Fuji Xerox Co., Ltd.) for subtask B. However, the best results on Twitter2014-test are substantially lower than those for the Twitter2013-test for both subtask A (86.63 vs. 90.14) and subtask B (70.96 vs 72.12). This is so despite the Majority Class baselines for Twitter2014-test being higher than those for Twitter2013-test: 42.2 vs. 38.1 for subtask A, and 34.6 vs. 29.2 for subtask B. Most likely, having access to the Twitter2013-test at development time, teams have overfitted on it. It could be also the case that some of the sentiment lexicons that were built in 2013 have become somewhat outdated by 2014.

6.2 Performance on Out-of-Domain Data

All participating systems were trained on tweets only. No training data were provided for the other test domains, SMS and blogs, nor was there training data for sarcastic tweets. Some teams, such as *NRC-Canada*, performed well across all test datasets. Surprisingly, on the out-of-domain test datasets they were able to achieve results comparable to those they obtained on tweets, or even better. Other teams, such as *TeamX*, chose to tune a weighting scheme specifically for class imbalances in tweets and, as a result, were only strong on Twitter datasets.

The Twitter2014-sarcasm dataset turned out to be the most challenging test dataset for most of the participants in both subtasks. The differences in performance on general and sarcastic tweets was 5–10 points for subtask A and 10–20 points for subtask B for most of the systems.

6.3 Impact of Training Data Size

As we mentioned above, due to Twitter's terms of service, we could not deliver the annotated tweets to the participants directly, and they had to download them on their own, which caused problems as at different times different subsets of the tweets could be downloaded. Thus, task participants had access to different number of training tweets depending on when they did the downloading.

To give some statistics, in the 2014 edition of the task, the number of tweets that participants could download and use for subtask B varied between 5,215 tweets and 10,882. On average, the teams were able to collect close to 9,000 tweets; teams that did not participate in 2013, and thus had to download the data later, could download about 8,500 tweets.

The difference in training data size did not seem to have had a major impact. In fact, the top two teams in subtask B in 2014 (*coooolll* [74] and *TeamX* [44]) used less than 8,500 tweets for training.

6.4 Use of External Resources

The participating systems were allowed to make use of external resources. As described in Section 2, a submission that directly used additional labeled data as part of the training dataset was considered *unconstrained*. In both 2013 and 2014, there were cases of a team submitting a constrained and an unconstrained run and the constrained run performing better. It is unclear why unconstrained systems did not always outperform the corresponding constrained ones. It could be because participants did not use enough external data or because the data they used was too different from Twitter or from our annotation method.

Several teams chose to use external (weakly) labeled tweet data indirectly, by creating sentiment lexicons or sentiment word representations, e.g., sentiment word embeddings. This approach allowed the systems to qualify as *constrained*, but it also offered some further benefits. First, it allowed to incorporate large amounts of noisily labeled data quickly and efficiently. Second, the classification systems were robust to the introduced noise because the noisy data were incorporated not directly as training instances but indirectly as features. Third, the generated sentiment resources could be easily distributed to the research community and used in other applications and domains [32].

These newly built sentiment resources, which leveraged on large collections of tweets, yielded large performance gains and assured top ranks for the teams that made use of them. For example, *NRC-Canada* reported 2 and 6.5 points of absolute improvement for subtasks A and B, respectively, by using their tweet-specific sentiment lexicons. On top of that, the *coooolll* team achieved another 3–4 points absolute improvement on the tweet test datasets for Subtask B thanks to sentiment-specific word embeddings.

Most participants greatly benefited from the use of existing general-domain sentiment lexicons. Even though the contribution of these lexicons on top of the Twitter-specific resources was usually modest, namely 1–2 points absolute, on the Twitter test datasets, the general-domain lexicons were particularly useful on out-of-domain data, such as the SMS test dataset, where their use resulted in gains of up to 3.5 points absolute for some participants.

Similarly, general-domain word representations, such as word clusters and word embeddings, showed larger gains on the out-of-domain SMS test dataset (1–2 points absolute) than on Twitter test datasets (0.5–1 points absolute).

6.5 Negation Handling

Many teams incorporated special handling of negation into their systems. The most popular approach transformed any word that appeared in a negated context by adding a suffix *_NEG* to it, e.g., *good* would become *good_NEG* [13, 56]. A negated context was defined as a text span between a negation word, e.g., *no, not, shouldn't*, and a punctuation mark or the end of the message.

Alternatively, some systems flipped the polarity of sentiment words when they occurred in a negated context, e.g., the positive word *good* would become negative when negated. The *RTRGO* team [24] reported an improvement of 1.5 points absolute for Subtask B on Twitter data when using both approaches together.

In [83], the authors argued that negation affects different words differently, and that a simple reversing strategy cannot adequately capture this complex behavior. Therefore, they proposed an empirical method to determine the sentiment of words in the presence of negation by creating a separate sentiment lexicon for negated words [33]. Their system, *NRC-Canada*, achieved 1.5 points of absolute improvement for Subtask A and 2.5 points for Subtask B by using sentiment lexicons generated for affirmative and negated contexts separately.

6.6 Use of Context in Subtask A

As suggested by the name of subtask A, *Contextual Polarity Disambiguation*, a model built for this subtask is expected to explore the context around a target term. For example, the top-performing *NRC-Canada* system used unigrams and bigrams extracted within four words on either side of the target term. The system also extracted additional features from the entire message in the same way as it extracted features from the target terms themselves. The inclusion of these context features resulted in F-score improvements of 4.08 points absolute on Twitter2013-test and 2.41 points on SMS2013-test. The second-best system in 2013, *AVAYA* [4], used dependency parse features such as the paths between the head of the target term and the root of the entire message. Similarly, the third-best *BOUNCE* system [34] used features and words extracted from neighboring target phrases, achieving 6.4 points of absolute improvement on Twitter2013-dev. The fourth-best *LVIC-LIMSI* system [42] also used the words surrounding the target terms during development, but their effect on the overall performance was not reported. The *SentiKLUE* system [17], second-best in 2014, used context in the form of automatically predicted message-level polarity.

6.7 Why Subtask A Seems Easier than Subtask B

The performance of the sentiment analysis systems is significantly higher for subtask A than for subtask B. A similar difference can also be observed for many baselines, including the SVM-unigrams baseline. Furthermore, a simple Target's Majority Class baseline showed surprisingly strong results on subtask A. Thus, we analyzed the data in order to determine why these baselines performed so well for subtask A. We found that 85.1% of the target words in Twitter2013-test and 88.8% of those in Twitter2014-test occurred as target tokens in the training data. Moreover, the distribution of occurrences of a target word that has been observed with different polarities is skewed towards one polarity.

Classifier	Targets fully seen in training	Targets partially seen in training	Targets unseen in training
(a) all features	93.31	85.42	84.09
(b) all, but no lexicons	92.96 (-0.35)	81.26 (-4.16)*	69.55 (-14.54)*
(c) all, but no n-grams	89.30 (-4.01)*	81.61 (-3.81)*	80.62 (-3.47)*

Table 10 Subtask A: macro-averaged F-scores for the *NRC-Canada* system on different subsets of Twitter2013-test with one of the feature groups removed. The number in brackets shows the absolute difference compared to the scores in row (a). Scores marked with a * are statistically significantly different ($p < .05$) from the corresponding scores in row (a).

Finally, the average ratio of instances pertaining to the dominant polarity of a target term to the total number of instances of that target term is 0.80. (Note that this ratio is calculated for all target words that occurred more than once in the training and in the test datasets.) These observations explain, at least in part, the high overall result and the dominant role of unigrams for subtask A.

We have conducted an experiment to examine the impact of sentiment resources in subtask A in the situation where the test targets would not appear in the training set. For this, we split the Twitter2013-test set into three subsets. In the first subset, "targets fully seen in training", each instance has a target with the following property: there exist instances in the training data with exactly the same target; this subset comprises 55% of the test set. In the second subset, "targets partially seen in training", each instance has a target X with the following property: there exist instances in the training data whose target expression includes one or more, but not all, tokens in X; this subset comprises 31% of the test set. In the third subset, "targets unseen in training", each instance has a target X with the following property: there are no instances in the training data whose target includes any of the tokens in X; this subset comprises 14% of the test set. We then ran the top-performing *NRC-Canada* system on each of the three subsets (a) using all features, (b) using all but the lexicon features, and (c) using all but the n-gram features. The results are shown in Table 10. We can see that on instances with unseen targets the sentiment lexicons play the most prominent role, yielding a gain of 14.54 points absolute.

7 Related Work

Sentiment analysis has enjoyed a lot of research attention over the last fifteen years, especially in sub-areas such as detecting subjective and objective sentences; classifying sentences as positive, negative, or neutral; and more recently, detecting the target of the sentiment. Much of this work focused on customer reviews of products and services, but tweets, Facebook posts, and other social media data are now increasingly being explored. Recent surveys by Pang and Lee [55] and Liu and Zhang [39] give detailed summaries of research on sentiment analysis.

Initially, the problem was regarded as standard document classification into topics, e.g., [56] experimented with various classifiers such as maximum entropy, Naïve Bayes and SVM, using standard features such as unigram/bigrams, word counts/present, word position and part-of-speech tagging. Around the same time, other researchers realized the importance of external sentiment lexicons, e.g., [77] proposed an unsupervised approach to learn the sentiment orientation of words/phrases: positive vs. negative. Later work studied the linguistic aspects of expressing opinions, evaluations, and speculations [79], the role of context in determining the sentiment orientation [81], of deeper linguistic processing such as negation handling [55], of finer-grained sentiment distinctions [54], of positional information [60], etc. Moreover, it was recognized that in many cases, it is crucial to know not just the polarity of the sentiment, but also the topic towards which this sentiment is expressed [71].

Naturally, most research in sentiment analysis was done for English, and much less efforts were devoted to other languages [1,11,30,31,57,73].

Early sentiment analysis research focused on customer reviews of movies, and later of hotels, phones, laptops, etc. Later, with the emergence of social media, sentiment analysis in Twitter became a hot research topic. Yet, there was a lack of suitable datasets for training, evaluating, and comparing different systems. This situation changed with the emergence of the SemEval task on Sentiment Analysis in Twitter, which ran in 2013-2015 [49,66,65]. The task created standard datasets of several thousand tweets annotated for sentiment polarity.

In fact, there was an even earlier shared task on sentiment analysis of text: the SemEval-2007 Affective Text Task [72]. However, it was framed as an unsupervised task where newspaper headlines were to be labeled with eight affect categories—positive and negative sentiment, as well as six emotions (joy, sadness, fear, anger, surprise, and disgust). For each headline–affect category pair, human annotators assigned scores from 0 to 100 indicating how strongly the headline expressed the affect category. In contrast, in our task, we focus on tweets, SMS messages, and blog posts. Moreover, apart from our main subtask on message-level sentiment, we also include a subtask on determining phrase-level sentiment.

Since our 2013 shared task, several other shared tasks have been proposed that further explored various sub-problems in sentiment analysis. We describe them briefly below.

7.1 Aspect-Based Sentiment Analysis

The goal of the SemEval-2014 Task 4 on Aspect-Based Sentiment Analysis (ABSA) was to identify aspect terms and the sentiment towards those aspect terms from customer reviews, where the focus was on two domains: laptops and restaurants [59].[14]

[14] `http://alt.qcri.org/semeval2014/task4`

For example, a review may gush positively about the lasagna at a restaurant, but negatively about the long wait before the food has arrived. In the restaurant domain, the aspect terms were further aggregated into coarse categories such as *food*, *service*, *ambiance*, *price*, and *miscellaneous*. The goal was to identify these aspect categories and the sentiment expressed towards them.

The ABSA task attracted 32 teams, who contributed 165 submissions. There is substantial overlap in the approaches and resources used by the participants in our task and in the ABSA task. Moreover, one of the top performing systems in our competition, *NRC-Canada*, also participated in the ABSA task and achieved the best scores in three out of the six subtask-domain combinations, including two out of the three sentiment subtasks [85]. The use of automatically created in-domain sentiment resources proved to be valuable for this task as well. Other useful features were derived from dependency parse trees in order to establish the relationship between aspect terms and sentiment expressions.

There is an ongoing follow-up task, SemEval-2015 Task 12 [58], which consolidates the subtasks from 2014 into a principled unified framework, where opinion target expressions, aspects and sentiment polarities are linked to each other in tuples. This is arguably useful when generating structured aspect-based opinion summaries from user reviews in real-world applications (e.g., customer review sites). The task is further extended to multiple sentences, and a new domain is added: reviews of hotels. Overall, this follow-up task has attracted 93 submissions by 16 teams.

7.2 Sentiment Analysis of Figurative Language

Social media posts are often teeming with creative and figurative language, rich in irony, sarcasm, and metaphors. The SemEval-2015 Task 11 [20] on Sentiment Analysis of Figurative Language[15] is interested in understanding how this creativity impacts perceived affect. For this purpose, tweets rich in irony, sarcasm, and metaphor were annotated on an 11-point discrete scale from -5 (most negative) to +5 (most positive). The participating systems were asked to predict this human-annotated fine-grained sentiment score, and were evaluated not only on the full dataset, but also separately on irony, sarcasm, and metaphor. One of the goals of the task was to explore how conventional sentiment analysis techniques can be altered to deal with non-literal content.

While our task also had evaluation on sarcastic tweets, for us this was just a separate (arguably harder) test set: we did not focus specifically on sarcasm and we did not provide specific training data for it. In contrast, SemEval-2015 Task 11 was fully dedicated to analyzing figurative language on Twitter (which includes not only sarcasm, but also irony and metaphor); moreover, they used an 11-point scale, while we were interested in predicting three classes. The task has attracted 15 teams, who submitted 29 runs.

[15] `http://alt.qcri.org/semeval2015/task11`

7.3 Detecting Events and Polarity Towards Events

SemEval-2015 Task 9 *CLIPEval Implicit Polarity of Events* [67] focuses on the implicit sentiment polarity towards events.[16] There are two subtasks. The first one asks to determine the sentiment (positive, negative, or neutral) towards an event instance, while the second one requires to identify both event instantiations and their associated polarity values. The task is based on a dataset of events annotated as instantiations of pleasant and unpleasant events previously collected in psychological research [38, 40]. It has attracted two teams who submitted three runs.

7.4 Sentiment Analysis of Movie Reviews

A popular test bed for sentiment analysis systems has been the movie reviews dataset from `rottentomatoes.com` collected initially by Pang and Lee [54]. State-of-the-art results were obtained on this test set using a recursive deep neural network [70]: an F-score of 85.4 on detecting review-level polarity (positive or negative). Even though this method does not require any handcrafted features or external semantic knowledge, it relies on extensive phrase-level sentiment annotations during training, which are expensive to acquire for most real-world applications.

Very comparable results (an F-score of 85.5) were reported using more conventional machine learning techniques, and crucially, large-scale sentiment lexicons generated automatically from tweets [33].

Finally, there is an ongoing 2015 Kaggle competition *Classify the sentiment of sentences from the Rotten Tomatoes dataset*, which aims to bring together sentiment analysis systems for fine-grained sentiment analysis of movie reviews.[17]

8 SemEval-2015 and Beyond

8.1 The SemEval-2015 Edition of the Task

In addition to the two subtasks described above (contextual and message-level polarity), we have added three new subtasks[18] in 2015. The first two focus on the sentiment *towards a given topic* in a single tweet or in a set of tweets, respectively. The third new subtask asks to determine the strength of prior association of Twitter terms with positive sentiment; this acts as an intrinsic evaluation of automatic methods that build Twitter-specific sentiment lexicons with real-valued sentiment association scores.

[16] `http://alt.qcri.org/semeval2015/task9`
[17] `https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews`
[18] `http://alt.qcri.org/semeval2015/task10/`

Topic-Based Message Polarity Classification. Given a message and a topic, classify whether the message is of positive, negative, or neutral sentiment towards the given topic.

Detecting Trends Towards a Topic. Given a set of messages on a given topic from the same period of time, classify the overall sentiment towards the topic in these messages as (a) strongly positive, (b) weakly positive, (c) neutral, (d) weakly negative, or (e) strongly negative.

Determining the Strength of Twitter Sentiment Terms. Given a word or a phrase, propose a score between 0 (lowest) and 1 (highest) that is indicative of the strength of association of that word/phrase with positive sentiment. If a word/phrase is more positive than another one, it should be assigned a relatively higher score.

8.2 Outlook on SemEval-2016

There is a new edition of the task which will run as part of SemEval-2016. In this new edition,[19] we will focus on sentiment with respect to a topic, but on a five-point scale, which is used for human review ratings on popular websites such as Amazon, TripAdvisor, Yelp, etc. From a research perspective, moving to an ordered five-point scale means moving from binary classification to *ordinal regression*.

We further plan to continue the trend detection subtask, which represents a move from classification to *quantification*,[20] and is on par with what applications need. In real-world applications, the focus often is not on the sentiment of a particular tweet, but rather on the percentage of tweets that are positive/negative.

Finally, we plan a new subtask on trend detection, but using a five-point scale, which would get us even closer to what business (e.g., marketing studies), and researchers, (e.g., in political science or public policy), want nowadays. From a research perspective, this is a problem of *ordinal quantification* [15].

9 Conclusion

We have presented the development and evaluation of a SemEval task on Sentiment Analysis in Twitter. The task included the creation of a large contextual and message-level polarity corpus consisting of tweets, SMS messages, LiveJournal messages, and a special test set of sarcastic tweets. It ran in 2013, 2014, and 2015, attracting the highest number of participating teams in all three years, with new challenging subtasks added in 2015, and some coming in 2016.

[19] `http://alt.qcri.org/semeval2016/task4/`

[20] Note that a good classifier is not necessarily a good quantifier, and vice versa [18]. See [16] for pointers to literature on text quantification.

The task has fostered the creation of some freely-available resources such as NRC's Hashtag Sentiment lexicon and the Sentiment140 lexicon [46], which the NRC-Canada team initially developed for their participation in SemEval-2013 task 2, and which were key for their winning the competition. Further specialized resources were developed for 2014 and for 2015 as well.

We hope that the long-lasting role of this task and the accompanying datasets, which we release freely for general research use,[21] will be to serve as a test bed for comparing different approaches and for fostering the creation of new relevant resources. This would facilitate research, would lead to better understanding of how sentiment is conveyed in social media, and ultimately to the creation of better sentiment analysis systems.

In future work, we plan to extend the task with new data from additional domains. We further plan to work on getting the setup as close as possible to what real-world applications need; this could mean altering the task/subtask definition, the data filtering process, the data annotation procedure, and/or the evaluation setup. Last but not least, we are interested in comparing annotations obtained from crowdsourcing with annotations from experts [7].

Acknowledgements We would like to thank Theresa Wilson, who was coorganizer of SemEval-2013 Task 2 and has contributed tremendously to the data collection and to the overall organization of the task. We would also like to thank Kathleen McKeown for her insight in creating the Amazon Mechanical Turk annotation task.

For the 2013 Amazon Mechanical Turk annotations, we received funding by the JHU Human Language Technology Center of Excellence and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the U.S. Army Research Lab. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

The 2014 Amazon Mechanical Turk annotations were funded by Kathleen McKeown and Smaranda Muresan.

The 2015 Amazon Mechanical Turk annotations were partially funded by SIGLEX.

References

1. Abdul-Mageed, M., Diab, M.T., Korayem, M.: Subjectivity and sentiment analysis of Modern Standard Arabic. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, ACL-HLT '11, pp. 587–591. Portland, Oregon (2011)
2. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC '10. Valletta, Malta (2010)
3. Barbosa, L., Feng, J.: Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pp. 36–44. Beijing, China (2010)
4. Becker, L., Erhart, G., Skiba, D., Matula, V.: AVAYA: Sentiment analysis on Twitter with self-training and polarity lexicon expansion. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval' 13, pp. 333–340. Atlanta, Georgia (2013)

[21] Available at `https://www.cs.york.ac.uk/semeval-2013/task2/`, `http://alt.qcri.org/semeval2014/task9/`, and `http://alt.qcri.org/semeval2015/task10/`

5. Bifet, A., Holmes, G., Pfahringer, B., Gavaldà, R.: Detecting sentiment change in Twitter streaming data. Journal of Machine Learning Research, Proceedings Track **17**, 5–11 (2011)
6. Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M., Maynard, D., Aswani, N.: TwitIE: An open-source information extraction pipeline for microblog text. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP '13, pp. 83–90. Hissar, Bulgaria (2013)
7. Borgholt, L., Simonsen, P., Hovy, D.: The rating game: Sentiment rating reproducibility from text. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP '15, pp. 2527–2532. Lisbon, Portugal (2015)
8. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Comput. Linguist. **18**(4), 467–479 (1992)
9. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011)
10. Chen, T., Kan, M.Y.: Creating a live, public short message service corpus: the NUS SMS corpus. Language Resources and Evaluation **47**(2), 299–335 (2013)
11. Chetviorkin, I., Loukachevitch, N.: Evaluating sentiment analysis systems in Russian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pp. 12–17. Sofia, Bulgaria (2013)
12. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**, 2493–2537 (2011)
13. Das, S.R., Chen, M.Y.: Yahoo! for Amazon: Sentiment extraction from small talk on the web. Manage. Sci. **53**(9), 1375–1388 (2007)
14. Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcasm in Twitter and Amazon. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10, pp. 107–116. Uppsala, Sweden (2010)
15. Esuli, A., Sebastiani, F.: Sentiment quantification. IEEE Intelligent Systems **25**, 72–75 (2010)
16. Esuli, A., Sebastiani, F.: Optimizing text quantifiers for multivariate loss functions. ACM Trans. Knowl. Discov. Data **9**(4), 27:1–27:27 (2015)
17. Evert, S., Proisl, T., Greiner, P., Kabashi, B.: SentiKLUE: Updating a polarity classifier in 48 hours. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 551–555. Dublin, Ireland (2014)
18. Forman, G.: Quantifying counts and costs via classification. Data Min. Knowl. Discov. **17**(2), 164–206 (2008)
19. Fort, K., Adda, G., Cohen, K.B.: Amazon Mechanical Turk: Gold mine or coal mine? Comput. Linguist. **37**(2), 413–420 (2011)
20. Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pp. 470–478. Denver, Colorado (2015)
21. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for Twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT '11, pp. 42–47. Portland, Oregon (2011)
22. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford **1**, 12 (2009)
23. Günther, T., Furrer, L.: GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 328–332. Atlanta, Georgia (2013)
24. Günther, T., Vancoppenolle, J., Johansson, R.: RTRGO: Enhancing the GU-MLT-LT system for sentiment analysis of short messages. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 497–502. Dublin, Ireland (2014)

25. Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., Hovy, E.: Learning whom to trust with MACE. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '13, pp. 1120–1130. Atlanta, Georgia (2013)
26. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pp. 168–177. Seattle, Washington (2004)
27. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. CoRR **abs/0812.1045** (2008)
28. Jansen, B., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. **60**(11), 2169–2188 (2009)
29. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56–65 (2007)
30. Jovanoski, D., Pachovski, V., Nakov, P.: Sentiment analysis in Twitter for Macedonian. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '15, pp. 249–257. Hissar, Bulgaria (2015)
31. Kapukaranov, B., Nakov, P.: Fine-grained sentiment analysis for movie reviews in Bulgarian. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '15, pp. 266–274. Hissar, Bulgaria (2015)
32. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.M.: NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval '14, pp. 437–442. Dublin, Ireland (2014)
33. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research **50**, 723–762 (2014)
34. Kökciyan, N., Çelebi, A., Özgür, A., Üsküdarl, S.: BOUNCE: Sentiment classification in Twitter using rich feature sets. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 554–561. Atlanta, Georgia (2013)
35. Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., Smith, A.N.: A dependency parser for tweets. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP '14, pp. 1001–1012. Doha, Qatar (2014)
36. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the OMG! In: Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM '11, pp. 538–541. Barcelona, Catalonia, Spain (2011)
37. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, pp. 591–600. Raleigh, North Carolina (2010)
38. Lewinsohn, J., Amenson, C.: Some relations between pleasant and unpleasant events and depression. Journal of Abnormal Psychology **87**(6), 644–654 (1978)
39. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: C.C. Aggarwal, C. Zhai (eds.) Mining Text Data, pp. 415–463. Springer US (2012)
40. MacPhillamy, D., Lewinsohn, P.M.: The pleasant event schedule: Studies on reliability, validity, and scale intercorrelation. Journal of Counseling and Clinical Psychology **50**(3), 363–380 (1982)
41. Manandhar, S., Yuret, D. (eds.): Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation. SemEval '13. Association for Computational Linguistics, Atlanta, Georgia (2013)
42. Marchand, M., Ginsca, A., Besançon, R., Mesnard, O.: LVIC-LIMSI: Using syntactic features and multi-polarity words for sentiment analysis in Twitter. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 418–424. Atlanta, Georgia (2013)
43. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of a Workshop at ICLR (2013)

44. Miura, Y., Sakaki, S., Hattori, K., Ohkuma, T.: TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 628–632. Dublin, Ireland (2014)
45. Mohammad, S.: #Emotional tweets. In: Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, *SEM '12, pp. 246–255. Montreal, Canada (2012)
46. Mohammad, S., Kiritchenko, S., Zhu, X.: NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 321–327. Atlanta, Georgia (2013)
47. Mohammad, S.M., Turney, P.D.: Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET '10, pp. 26–34. Los Angeles, California (2010)
48. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence **29**(3), 436–465 (2013)
49. Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T.: SemEval-2013 task 2: Sentiment analysis in Twitter. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 312–320. Atlanta, Georgia (2013)
50. Nakov, P., Zesch, T. (eds.): Proceedings of the 8th International Workshop on Semantic Evaluation. SemEval '14. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (2014)
51. Nakov, P., Zesch, T., Cer, D., Jurgens, D. (eds.): Proceedings of the 9th International Workshop on Semantic Evaluation. SemEval '15. Association for Computational Linguistics, Denver, Colorado (2015)
52. O'Connor, B., Balasubramanyan, R., Routledge, B., Smith, N.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM '10, pp. 122–129. Washington, DC (2010)
53. Pak, A., Paroubek, P.: Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, pp. 436–439. Uppsala, Sweden (2010)
54. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '05, pp. 115–124. Ann Arbor, Michigan (2005)
55. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1–2), 1–135 (2008)
56. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '02, pp. 79–86. Philadelphia, Pennsylvania (2002)
57. Perez-Rosas, V., Banea, C., Mihalcea, R.: Learning sentiment lexicons in Spanish. In: Proceedings of the Eight International Conference on Language Resources and Evaluation, LREC '12. Istanbul, Turkey (2012)
58. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: Aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pp. 486–495. Denver, Colorado (2015)
59. Pontiki, M., Papageorgiou, H., Galanis, D., Androutsopoulos, I., Pavlopoulos, J., Manandhar, S.: SemEval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 27–35. Dublin, Ireland (2014)
60. Raychev, V., Nakov, P.: Language-independent sentiment analysis using subjectivity and positional information. In: Proceedings of the International Conference on Recent

Advances in Natural Language Processing, RANLP '09, pp. 360–364. Borovets, Bulgaria (2009)
61. Reckman, H., Baird, C., Crawford, J., Crowell, R., Micciulla, L., Sethi, S., Veress, F.: teragram: Rule-based detection of sentiment phrases using SAS sentiment analysis. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 513–519. Atlanta, Georgia (2013)
62. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: An experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 1524–1534. Edinburgh, Scotland, UK (2011)
63. Ritter, A., Etzioni, O., Clark, S., et al.: Open domain event extraction from Twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, pp. 1104–1112. Beijing, China (2012)
64. Rosenthal, S., McKeown, K.: Detecting opinionated claims in online discussions. In: Proceedings of the 2012 IEEE Sixth International Conference on Semantic Computing, ICSC '12, pp. 30–37. Washington, DC (2012)
65. Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V.: SemEval-2015 task 10: Sentiment analysis in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pp. 450–462. Denver, Colorado (2015)
66. Rosenthal, S., Ritter, A., Nakov, P., Stoyanov, V.: SemEval-2014 Task 9: Sentiment analysis in Twitter. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 73–80. Dublin, Ireland (2014)
67. Russo, I., Caselli, T., Strapparava, C.: SemEval-2015 task 9: CLIPEval implicit polarity of events. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pp. 442–449. Denver, Colorado (2015)
68. dos Santos, C.: Think Positive: Towards Twitter sentiment analysis from scratch. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 647–651. Dublin, Ireland (2014)
69. Severyn, A., Moschitti, A.: UNITN: Training deep convolutional neural network for Twitter sentiment classification. In: Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pp. 464–469. Denver, Colorado (2015)
70. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '13, pp. 1631–1642. Seattle, Washington (2013)
71. Stoyanov, V., Cardie, C.: Topic identification for fine-grained opinion analysis. In: Proceedings of the 22nd International Conference on Computational Linguistics, COLING '08, pp. 817–824. Manchester, United Kingdom (2008)
72. Strapparava, C., Mihalcea, R.: SemEval-2007 task 14: Affective text. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval '07, pp. 70–74. Prague, Czech Republic (2007)
73. Tan, S., Zhang, J.: An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications **34**(4), 2622 – 2629 (2008)
74. Tang, D., Wei, F., Qin, B., Liu, T., Zhou, M.: Coooolll: A deep learning system for Twitter sentiment classification. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 208–212. Dublin, Ireland (2014)
75. Tiantian, Z., Fangxi, Z., Lan, M.: ECNUCS: A surface information based system description of sentiment analysis in Twitter in the SemEval-2013 (task 2). In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, SemEval '13, pp. 408–413. Atlanta, Georgia (2013)
76. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I.: Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM '10, pp. 178–185. Washington, DC (2010)
77. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '02, pp. 417–424. Philadelphia, Pennsylvania (2002)

78. Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., Cristóbal, J.C.G.: TASS - Workshop on Sentiment Analysis at SEPLN. Procesamiento del Lenguaje Natural **50**, 37–44 (2013)
79. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M.: Learning subjective language. Comput. Linguist. **30**(3), 277–308 (2004)
80. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation **39**(2-3), 165–210 (2005)
81. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT-EMNLP '05, pp. 347–354. Vancouver, British Columbia, Canada (2005)
82. Zhao, J., Lan, M., Zhu, T.: ECNU: Expression- and message-level sentiment orientation classification in Twitter using multiple effective features. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 259–264. Dublin, Ireland (2014)
83. Zhu, X., Guo, H., Mohammad, S.M., Kiritchenko, S.: An empirical study on the effect of negation words on sentiment. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL '14, pp. 304–313. Baltimore, Maryland (2014)
84. Zhu, X., Kiritchenko, S., Mohammad, S.: NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets. In: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pp. 443–447. Dublin, Ireland (2014)
85. Zhu, X., Kiritchenko, S., Mohammad, S.M.: NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the International Workshop on Semantic Evaluation, SemEval '14, pp. 437–442. Dublin, Ireland (2014)

Team	Affiliation	2013?
Alberta	University of Alberta	
AMI_ERIC	AMI Software R&D and Université de Lyon (ERIC LYON 2)	yes
AUEB	Athens University of Economics and Business	yes
BUAP	Benemérita Universidad Autónoma de Puebla	
CISUC_KIS	University of Coimbra	
Citius	University of Santiago de Compostela	
CMU-Qatar	Carnegie Mellon University, Qatar	
CMUQ-Hybrid	Carnegie Mellon University, Qatar	
columbia_nlp	Columbia University	yes
coooll	Harbin Institute of Technology	
DAEDALUS	Daedalus	
DejaVu	Indian Institute of Technology, Kanpur	
ECNU	East China Normal University	yes
GPLSI	University of Alicante	
IBM_EG	IBM Egypt	
IITPatna	Indian Institute of Technology, Patna	
IIT-Patna	Indian Institute of Technology, Patna	
JOINT_FORCES	Zurich University of Applied Sciences	
Kea	York University, Toronto	yes
KUNLPLab	Koç University	
lsis_lif	Aix-Marseille University	yes
Lt_3	Ghent University	
LyS	Universidade da Coruña	
NILC_USP	University of São Paulo	yes
NRC-Canada	National Research Council Canada	yes
Rapanakis	Stamatis Rapanakis	
RTRGO	Retresco GmbH and University of Gothenburg	yes
SAIL	Signal Analysis and Interpretation Laboratory	yes
SAP-RI	SAP Research and Innovation	
senti.ue	Universidade de Évora	yes
SentiKLUE	Friedrich-Alexander-Universität Erlangen-Nürnberg	yes
SINAI	University of Jaén	yes
SU-FMI	Sofia University	
SU-sentilab	Sabanci University	yes
SWISS-CHOCOLATE	ETH Zurich	
Synalp-Empathic	University of Lorraine	
TeamX	Fuji Xerox Co., Ltd.	
Think_Positive	IBM Research, Brazil	
TJP	University of Northumbria at Newcastle Upon Tyne	yes
TUGAS	Instituto de Engenharia de Sistemas e Computadores, Investigação e Desenvolvimento em Lisboa	yes
UKPDIPF	Ubiquitous Knowledge Processing Lab	
UMCC_DLSI_Graph	Universidad de Matanzas and Univarsidad de Alicante	yes
UMCC_DLSI_Sem	Universidad de Matanzas and Univarsidad de Alicante	yes
Univ. Warwick	University of Warwick	
UPV-ELiRF	Universitat Politècnica de València	
USP_Biocom	University of São Paulo and Federal University of São Carlos	

Table 11 The teams that participated in SemEval-2014 task 9, their affiliations, and an indication whether each team participated in SemEval-2013 Task 2.