SemEval-2015 Task 10: Sentiment Analysis in Twitter

Sara Rosenthal
Columbia University
sara@cs.columbia.edu

Preslav Nakov
Qatar Computing Research Institute
pnakov@qf.org.qa

Svetlana Kiritchenko
National Research Council Canada
Svetlana.Kiritchenko@nrc-cnrc.gc.ca

Saif M Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Alan Ritter
The Ohio State University
aritter@cs.washington.edu

Veselin Stoyanov
Facebook
vesko.st@gmail.com

Abstract

In this paper, we describe the 2015 iteration of the SemEval shared task on Sentiment Analysis in Twitter. This was the most popular sentiment analysis shared task to date with more than 40 teams participating in each of the last three years. This year's shared task competition consisted of five sentiment prediction subtasks. Two were reruns from previous years: (A) sentiment expressed by a phrase in the context of a tweet, and (B) overall sentiment of a tweet. We further included three new subtasks asking to predict (C) the sentiment towards a topic in a single tweet, (D) the overall sentiment towards a topic in a set of tweets, and (E) the degree of prior polarity of a phrase.

1 Introduction

Social media such as Weblogs, microblogs, and discussion forums are used daily to express personal thoughts, which allows researchers to gain valuable insight into the opinions of a very large number of individuals, i.e., at a scale that was simply not possible a few years ago. As a result, nowadays, sentiment analysis is commonly used to study the public opinion towards persons, objects, and events. In particular, opinion mining and opinion detection are applied to product reviews (Hu and Liu, 2004), for agreement detection (Hillard et al., 2003), and even for sarcasm identification (González-Ibáñez et al., 2011; Liebrecht et al., 2013).

Early work on detecting sentiment focused on newswire text (Wiebe et al., 2005; Baccianella et al., 2010; Pang et al., 2002; Hu and Liu, 2004). As later research turned towards social media, people realized this presented a number of new challenges.

Misspellings, poor grammatical structure, emoticons, acronyms, and slang were common in these new media, and were explored by a number of researchers (Barbosa and Feng, 2010; Bifet et al., 2011; Davidov et al., 2010; Jansen et al., 2009; Kouloumpis et al., 2011; O'Connor et al., 2010; Pak and Paroubek, 2010). Later, specialized shared tasks emerged, e.g., at SemEval (Nakov et al., 2013; Rosenthal et al., 2014), which compared teams against each other in a controlled environment using the same training and testing datasets. These shared tasks had the side effect to foster the emergence of a number of new resources, which eventually spread well beyond SemEval, e.g., NRC's Hashtag Sentiment lexicon and the Sentiment140 lexicon (Mohammad et al., 2013).[1]

Below, we discuss the public evaluation done as part of SemEval-2015 Task 10. In its third year, the SemEval task on Sentiment Analysis in Twitter has once again attracted a large number of participants: 41 teams across five subtasks, with most teams participating in more than one subtask.

This year the task included reruns of two legacy subtasks, which asked to detect the sentiment expressed in a tweet or by a particular phrase in a tweet. The task further added three new subtasks. The first two focused on the sentiment towards a given topic in a single tweet or in a set of tweets, respectively. The third new subtask focused on determining 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.

[1] http://www.purl.com/net/lexicons

In the remainder of this paper, we first introduce the problem of sentiment polarity classification and our subtasks. We then describe the process of creating the training, development, and testing datasets. We list and briefly describe the participating systems, the results, and the lessons learned. Finally, we compare the task to other related efforts and we point to possible directions for future research.

2 Task Description

Below, we describe the five subtasks of SemEval-2015 Task 10 on Sentiment Analysis in Twitter.

- **Subtask A. Contextual Polarity Disambiguation:** Given an instance of a word/phrase in the context of a message, determine whether it expresses a positive, a negative or a neutral sentiment in that context.

- **Subtask B. Message Polarity Classification:** Given a message, determine whether it expresses a positive, a negative, or a neutral/objective sentiment. If both positive and negative sentiment are expressed, the stronger one should be chosen.

- **Subtask C. Topic-Based Message Polarity Classification:** Given a message and a topic, decide whether the message expresses a positive, a negative, or a neutral sentiment towards the topic. If both positive and negative sentiment are expressed, the stronger one should be chosen.

- **Subtask D. Detecting Trend 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.

- **Subtask E. Determining Strength of Association of Twitter Terms with Positive Sentiment (Degree of Prior Polarity):** Given a word/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.

3 Datasets

In this section, we describe the process of collecting and annotating our datasets of short social media text messages. We focus our discussion on the 2015 datasets; more detail about the 2013 and the 2014 datasets can be found in (Nakov et al., 2013) and (Rosenthal et al., 2014).

3.1 Data Collection

3.1.1 Subtasks A–D

First, we gathered tweets that express sentiment about popular topics. For this purpose, we extracted named entities from millions of tweets, using a Twitter-tuned NER system (Ritter et al., 2011). Our initial training set was collected over a one-year period spanning from January 2012 to January 2013. Each subsequent Twitter test set was collected a few months prior to the corresponding evaluation. We used the public streaming Twitter API to download the tweets.

We then identified popular topics as those named entities that are frequently mentioned in association with a specific date (Ritter et al., 2012). 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.

The collected tweets 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.[2]

For subtasks C and D, we did some manual pruning based on the topics. First, we excluded topics that were incomprehensible, ambiguous (e.g., *Barcelona*, which is a name of a sports team and also of a place), or were too general (e.g., *Paris*, which is a name of a big city). Second, we discarded tweets that were just mentioning the topic, but were not really about the topic. Finally, we discarded topics with too few tweets, namely less than 10.

[2]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.

Instructions: Subjective words are ones which convey an opinion or sentiment. Given a Twitter message, identify whether it is objective, positive, negative, or neutral. Then, identify each subjective word or phrase in the context of the sentence and mark the position of its start and end in the text boxes below. The number above each word indicates its position. The word/phrase will be generated in the adjacent textbox so that you can confirm that you chose the correct range. Choose the polarity of the word or phrase by selecting one of the radio buttons: positive, negative, or neutral. If a sentence is not subjective please select the checkbox indicating that "There are no subjective words/phrases". If a tweet is sarcastic, please select the checkbox indicating that "The tweet is sarcastic". Please read the examples and invalid responses before beginning if this is your first time answering this hit.

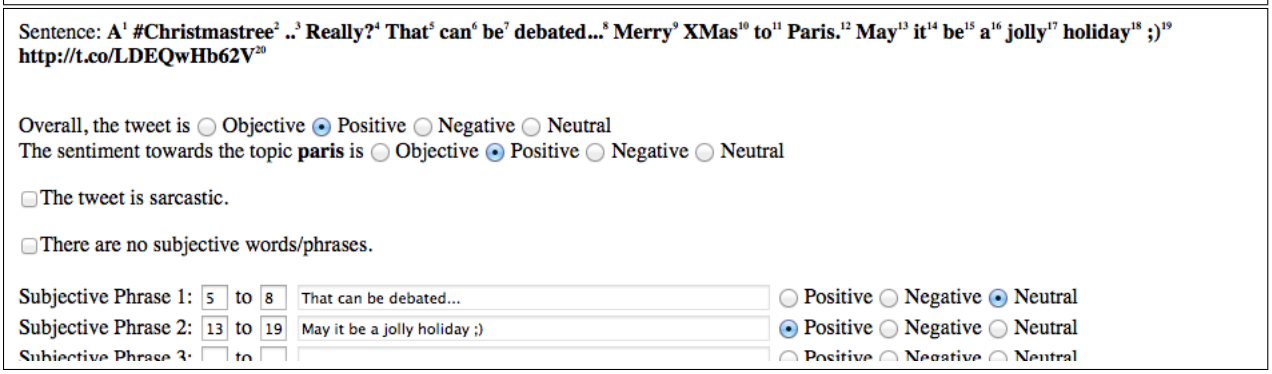

Figure 1: The instructions we gave to the workers on Mechanical Turk, followed by a screenshot.

3.1.2 Subtask E

We selected high-frequency target terms from the Sentiment140 and the Hashtag Sentiment tweet corpora (Kiritchenko et al., 2014). In order to reduce the skewness towards the neutral class, we selected terms from different ranges of automatically determined sentiment values as provided by the corresponding Sentiment140 and Hashtag Sentiment lexicons. The term set comprised regular English words, hashtagged words (e.g., *#loveumom*), misspelled or creatively spelled words (e.g., *parlament* or *happeeee*), abbreviations, shortenings, and slang. Some terms were negated expressions such as *no fun*. (It is known that negation impacts the sentiment of its scope in complex ways (Zhu et al., 2014).) We annotated these terms for degree of sentiment manually. Further details about the data collection and the annotation process can be found in Section 3.2.2 as well as in (Kiritchenko et al., 2014).

The trial dataset consisted of 200 instances, and no training dataset was provided. Note, however, that the trial data was large enough to be used as a development set, or even as a training set. Moreover, the participants were free to use any additional manually or automatically generated resources when building their systems for subtask E. The testset included 1,315 instances.

3.2 Annotation

Below we describe the data annotation process.

3.2.1 Subtasks A–D

We used Amazon's Mechanical Turk for the annotations of subtasks A–D. Each tweet message was annotated by five Mechanical Turk workers, also known as Turkers. The annotations for subtasks A–D were done concurrently, in a single task. A 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). He/she also had to indicate the overall polarity of the tweet message in general (subtask B) as well as the overall polarity of the message towards the given target topic (subtasks C and D). The instructions we gave to the Turkers, along with an example, are shown in Figure 1. We further made available to the Turkers several additional examples, which we show in Table 1.

Providing all the required annotations for a given tweet message constituted a Human Intelligence Task, or a HIT. In order to qualify to work on our HITs, a Turker had to have an approval rate greater than 95% and should have completed at least 50 approved HITs.

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 and annotations we provided to the Turkers. All subjective phrases are italicized and color-coded: positive phrases are in green, negative ones are in red, and neutral ones 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. Words and phrases that were marked as subjective are in bold italic. The first five rows are annotations provided by Turkers, and the final row shows their intersection. The last column shows the token-level accuracy for each annotation compared to the intersection.

We further discarded the following types of message annotations:

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

Recall that each tweet message was annotated by five different Turkers. We consolidated these annotations for subtask A using intersection as shown in the last row of Table 2. A word had to appear in 3/5 of the annotations in order to be considered subjective. It further had to be labeled with a particular polarity (positive, negative, or neutral) by three of the five Turkers in order to receive that polarity label. As the example shows, this effectively shortens the spans of the annotated phrases, often to single words, as it is hard to agree on long phrases.

Corpus	Pos.	Neg.	Obj. / Neu.	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
Twitter2015-test	1899	1008	190	3097

Table 3: Dataset statistics for subtask A.

We also experimented with two alternative methods for combining annotations: (*i*) by computing the union of the annotations for the sentence, and (*ii*) by taking the annotations by the Turker who has annotated the highest number of HITs. However, our manual analysis has shown that both alternatives performed worse than using the intersection.

Corpus	Pos.	Neg.	Obj. / Neu.	Total
Twitter2013-train	3,662	1,466	4,600	9,728
Twitter2013-dev	575	340	739	1,654
Twitter2013-test	1,572	601	1,640	3,813
SMS2013-test	492	394	1,207	2,093
Twitter2014-test	982	202	669	1,853
Twitter2014-sarcasm	33	40	13	86
LiveJournal2014-test	427	304	411	1,142
Twitter2015-test	1040	365	987	2392

Table 4: Dataset statistics for subtask B.

Corpus	Topics	Pos.	Neg.	Obj. / Neu.	Total
Train	44	142	56	288	530
Test	137	870	260	1256	2386

Table 5: Twitter-2015 statistics for subtasks C & D.

For subtasks B and C, we consolidated the tweet-level annotations using majority voting, requiring that the winning label be proposed by at least three of the five Turkers; we discarded all tweets for which 3/5 majority could not be achieved. As in previous years, we combined the objective and the neutral labels, which Turkers tended to mix up.

We used these consolidated annotations as gold labels for subtasks A, B, C & D. The statistics for all datasets for these subtasks are shown in Tables 3, 4, and 5, respectively. Each dataset is marked with the year of the SemEval edition it was produced for. An annotated example from each source (Twitter, SMS, LiveJournal) is shown in Table 6; examples for sentiment towards a topic can be seen in Table 7.

3.2.2 Subtask E

Subtask E asks systems to propose a numerical score for the positiveness of a given word or phrase. Many studies have shown that people are actually quite bad at assigning such absolute scores: inter-annotator agreement is low, and annotators struggle even to remain self-consistent. In contrast, it is much easier to make relative judgments, e.g., to say whether one word is more positive than another. Moreover, it is possible to derive an absolute score from pairwise judgments, but this requires a much larger number of annotations. Fortunately, there are schemes that allow to infer more pairwise annotations from less judgments.

One such annotation scheme is MaxDiff (Louviere, 1991), which is widely used in market surveys (Almquist and Lee, 2009); it was also used in a previous SemEval task (Jurgens et al., 2012).

In MaxDiff, the annotator is presented with four terms and asked which term is most positive and which is least positive. By answering just these two questions, five out of six pairwise rankings become known. Consider a set in which a judge evaluates A, B, C, and D. If she says that A and D are the most and the least positive, we can infer the following: $A > B, A > C, A > D, B > D, C > D$. The responses to the MaxDiff questions can then be easily translated into a ranking for all the terms and also into a real-valued score for each term. We crowdsourced the MaxDiff questions on CrowdFlower, recruiting ten annotators per MaxDiff example. Further details can be found in Section 6.1.2. of (Kiritchenko et al., 2014).

3.3 Lower & Upper Bounds

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 agreement between annotators. Unfortunately, as we derive annotations by agreement, we cannot calculate standard statistics such as Kappa. 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 8. 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. 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 it from a system.

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 6: Example annotations for each source of messages. The 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).

Topic	Message	Message-Level Polarity	Topic-Level Polarity
leeds united	Saturday without Leeds United is like Sunday dinner it doesn't feel normal at all (Ryan)	negative	positive
demi lovato	Who are you tomorrow? Will you make me smile or just bring me sorrow? #HottieOfTheWeek Demi Lovato	neutral	positive

Table 7: Example of annotations in Twitter showing differences between topic- and message-level polarity.

Corpus	Subtask A			Subtask B
	Low	Avg	Up	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
Sarcasm2014-test	62.6	83.1	95.6	71.2
Twitter2015-test	73.2	87.6	96.8	75.7

Table 8: 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.

3.4 Tweets Delivery

Due to restrictions in the 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 that extracts the corresponding tweets via the Twitter API.[3]

As a result, different teams had access to different number of training tweets depending on when they did the downloading. However, our analysis has shown that this did not have a major impact and many high-scoring teams had less training data compared to some lower-scoring ones.

[3]https://dev.twitter.com

4 Scoring

4.1 Subtasks A-C: Phrase-Level, Message-Level, and Topic-Level Polarity

The participating systems were required to perform a three-way classification, i.e., to assign one of the folowing three labels: *positive*, *negative* or *objective/neutral*. We evaluated the systems in terms of a macro-averaged F_1 score for predicting positive and negative phrases/messages.

We first computed positive precision, P_{pos} as follows: we found the number of phrases/messages that a system correctly predicted to be positive, and we divided that number by the total number of examples it predicted to be positive. To compute positive recall, R_{pos}, we found the number of phrases/messages correctly predicted to be positive and we divided that number by the total number of positives in the gold standard. We then calculated an F_1 score for the positive class as follows $F_{pos} = \frac{2P_{pos}R_{pos}}{P_{pos}+R_{pos}}$. We carried out similar computations for the negative phrases/messages, F_{neg}. The overall score was then computed as the average of the F_1scores for the positive and for the 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.

4.2 Subtask D: Overall Polarity Towards a Topic

This subtask asks to predict the overall sentiment of a set of tweets towards a given topic. In other words, to predict the ratio r_i of positive (pos_i) tweets to the number of positive and negative sentiment tweets in the set of tweets about the i-th topic:

$$r_i = Pos_i/(Pos_i + Neg_i)$$

Note, that neutral tweets do not participate in the above formula; they have only an indirect impact on the calculation, similarly to subtasks A–C.

We use the following two evaluation measures for subtask D:

- **AvgDiff** (official score): Calculates the absolute difference betweeen the predicted r'_i and the gold r_i for each i, and then averages this difference over all topics.

- **AvgLevelDiff** (unofficial score): This calculation is the same as AvgDiff, but with r'_i and r_i first remapped to five coarse numerical categories: 5 (strongly positive), 4 (weakly positive), 3 (mixed), 2 (weakly negative), and 1 (strongly negative). We define this remapping based on intervals as follows:
  - 5: $0.8 < x \leq 1.0$
  - 4: $0.6 < x \leq 0.8$
  - 3: $0.4 < x \leq 0.6$
  - 2: $0.2 < x \leq 0.4$
  - 1: $0.0 \leq x \leq 0.2$

4.3 Subtask E: Degree of Prior Polarity

The scores proposed by the participating systems were evaluated by first ranking the terms according to the proposed sentiment score and then comparing this ranked list to a ranked list obtained from aggregating the human ranking annotations. We used Kendall's rank correlation (Kendall's τ) as the official evaluation metric to compare the ranked lists (Kendall, 1938). We also calculated scores for Spearman's rank correlation (Lehmann and D'Abrera, 2006), as an unofficial score.

Team ID	Affiliation
CIS-positiv	University of Munich
CLaC-SentiPipe	CLaC Labs, Concordia University
DIEGOLab	Arizona State University
ECNU	East China Normal University
elirf	Universitat Politècnica de València
Frisbee	Frisbee
Gradiant-Analytics	Gradiant
GTI	AtlantTIC Center, University of Vigo
IHS-RD	IHS inc
iitpsemeval	Indian Institute of Technology, Patna
IIIT-H	IIIT, Hyderabad
INESC-ID	IST, INESC-ID
IOA	Institute of Acoustics, Chinese Academy of Sciences
KLUEless	FAU Erlangen-Nürnberg
lsislif	Aix-Marseille University
NLP	NLP
RGUSentimentMiners123	Robert Gordon University
RoseMerry	The University of Melbourne
Sentibase	IIIT, Hyderabad
SeNTU	Nanyang Technological University, Singapore
SHELLFBK	Fondazione Bruno Kessler
sigma2320	Peking University
Splusplus	Beihang University
SWASH	Swarthmore College
SWATAC	Swarthmore College
SWATCMW	Swarthmore College
SWATCS65	Swarthmore College
Swiss-Chocolate	Zurich University of Applied Sciences
TwitterHawk	University of Massachusetts, Lowell
UDLAP2014	Universidad de las Amèricas Puebla, Mexico
UIR-PKU	University of International Relations
UMDuluth-CS8761	University of Minnesota, Duluth
UNIBA	University of Bari Aldo Moro
unitn	University of Trento
UPF-taln	Universitat Pompeu Fabra
WarwickDCS	University of Warwick
Webis	Bauhaus-Universität Weimar
whu-iss	International Software School, Wuhan University
Whu-Nlp	Computer School, Wuhan University
wxiaoac	Hong Kong University of Science and Technology
ZWJYYC	Peking University

Table 9: The participating teams and their affiliations.

5 Participants and Results

The task attracted 41 teams: 11 teams participated in subtask A, 40 in subtask B, 7 in subtask C, 6 in subtask D, and 10 in subtask E. The IDs and affiliations of the participating teams are shown in Table 9.

5.1 Subtask A: Phrase-Level Polarity

The results (macro-averaged F_1 score) for subtask A are shown in Table 10. The official results on the new Twitter2015-test dataset are shown in the last column, while the first five columns show F_1 on the 2013 and on the 2014 progress test datasets:[4] Twitter2013-test, SMS2013-test, Twitter2014-test, Twitter2014-sarcasm, and LiveJournal2014-test. There is an index for each result showing the relative rank of that result within the respective column. The participating systems are ranked by their score on the Twitter2015-test dataset, which is the official ranking for subtask A; all remaining rankings are secondary.

[4]Note that the 2013 and the 2014 test datasets were made available for development, but it was explicitly forbidden to use them for training.

#	System	2013: Progress		2014: Progress			2015: Official
		Tweet	SMS	Tweet	Tweet sarcasm	Live-Journal	Tweet
1	unitn	90.10_1	88.60_2	87.12_1	73.65_5	84.46_2	$\mathbf{84.79}_1$
2	KLUEless	88.56_2	88.62_1	84.99_3	75.59_4	83.94_4	$\mathbf{84.51}_2$
3	IOA	83.90_7	84.18_7	85.37_2	71.58_6	85.61_1	$\mathbf{82.76}_3$
4	WarwickDCS	84.08_6	84.40_5	83.89_5	78.03_2	83.18_5	$\mathbf{82.46}_4$
5	TwitterHawk	82.87_8	83.64_8	84.05_4	75.62_3	83.97_3	$\mathbf{82.32}_5$
6	iitpsemeval	85.81_3	85.86_3	82.73_6	65.71_9	81.76_7	$\mathbf{81.31}_6$
7	ECNU	85.28_4	84.70_4	82.09_7	70.96_7	82.49_6	$\mathbf{81.08}_7$
8	Whu-Nlp	79.76_9	81.78_9	81.69_8	63.14_{11}	80.87_9	$\mathbf{78.84}_8$
9	GTI	84.64_5	84.37_6	79.48_9	81.53_1	81.61_8	$\mathbf{77.27}_9$
10	whu-iss	74.02_{10}	70.26_{11}	72.20_{10}	69.33_8	73.57_{10}	$\mathbf{71.35}_{10}$
11	UMDuluth-CS8761	72.71_{11}	71.80_{10}	69.84_{11}	64.53_{10}	71.53_{11}	$\mathbf{66.21}_{11}$
	baseline	38.1	31.5	42.2	39.8	33.4	**38.0**

Table 10: **Results for subtask A: Phrase-Level Polarity.** The systems are ordered by their score on the Twitter2015 test dataset; the rankings on the individual datasets are indicated with a subscript.

There were less participants this year, probably due to having a new similar subtask: C. Notably, many of the participating teams were newcomers.

We can see that all systems beat the majority class baseline by 25-40 F_1 points absolute on all datasets. The winning team unitn (using deep convolutional neural networks) achieved an F_1 of 84.79 on Twitter2015-test, followed closely by KLUEless (using logistic regression) with F_1=84.51.

Looking at the progress datasets, we can see that unitn was also first on both progress Tweet datasets, and second on SMS and on LiveJournal. KLUEless won SMS and was second on Twitter2013-test. The best result on LiveJournal was achieved by IOA, who were also second on Twitter2014-test and third on the official Twitter2015-test. None of these teams was ranked in top-3 on Twitter2014-sarcasm, where the best team was GTI, followed by WarwickDCS.

Compared to 2014, there is an improvement on Twitter2014-test from 86.63 in 2014 (NRC-Canada) to 87.12 in 2015 (unitn). The best result on Twitter2013-test of 90.10 (unitn) this year is very close to the best in 2014 (90.14 by NRC-Canada). Similarly, the best result on LiveJournal stays exactly the same, i.e., F_1=85.61 (SentiKLUE in 2014 and IOA in 2015). However, there is slight degradation for SMS2013-test from 89.31 (ECNU) in 2014 to 88.62 (KLUEless) in 2015. The results also degraded for Twitter2014-sarcasm from 82.75 (senti.ue) to 81.53 (GTI).

5.2 Subtask B: Message-Level Polarity

The results for subtask B are shown in Table 11. Again, we show results on the five progress test datasets from 2013 and 2014, in addition to those for the official Twitter2015-test datasets.

Subtask B attracted 40 teams, both newcomers and returning, similarly to 2013 and 2014. All managed to beat the baseline with the exception of one system for Twitter2015-test, and one for Twitter2014-test. There is a cluster of four teams at the top: Webis (ensemble combining four Twitter sentiment classification approaches that participated in previous editions) with an F_1 of 64.84, unitn with 64.59, lsislif (logistic regression with special weighting for positives and negatives) with 64.27, and INESC-ID (word embeddings) with 64.17.

The last column in the table shows the results for the 2015 sarcastic tweets. Note that, unlike in 2014, this time they were not collected separately and did not have a special #sarcasm tag; instead, they are a subset of 75 tweets from Twitter2015-test that were flagged as sarcastic by the human annotators. The top system is IOA with an F_1 of 65.77, followed by INESC-ID with 64.91, and NLP with 63.62.

Looking at the progress datasets, we can see that the second ranked unitn is also second on SMS and on Twitter2014-test, and third on Twitter2013-test. INESC-ID in turn is third on Twitter2014-test and also third on Twitter2014-sarcasm. Webis and lsislif were less strong on the progress datasets.

#	System	2013: Progress		2014: Progress			2015: Official	
		Tweet	SMS	Tweet	Tweet sarcasm	Live-Journal	Tweet	Tweet sarcasm
1	Webis	68.49_{10}	63.92_{14}	70.86_{7}	49.33_{12}	71.64_{14}	$\mathbf{64.84}_{1}$	53.59_{22}
2	unitn	72.79_{2}	68.37_{2}	73.60_{2}	55.44_{5}	72.48_{12}	$\mathbf{64.59}_{2}$	55.01_{19}
3	lsislif	71.34_{4}	63.42_{17}	71.54_{5}	46.57_{22}	73.01_{10}	$\mathbf{64.27}_{3}$	46.00_{33}
4	INESC-ID*	71.97_{3}	63.78_{15}	72.52_{3}	56.23_{3}	69.78_{22}	$\mathbf{64.17}_{4}$	64.91_{2}
5	Splusplus	72.80_{1}	67.16_{5}	74.42_{1}	42.86_{31}	75.34_{1}	$\mathbf{63.73}_{5}$	60.99_{7}
6	wxiaoac	66.43_{16}	64.04_{13}	68.96_{11}	54.38_{7}	73.36_{9}	$\mathbf{63.00}_{6}$	52.22_{26}
7	IOA	71.32_{5}	68.14_{3}	71.86_{4}	51.48_{9}	74.52_{2}	$\mathbf{62.62}_{7}$	65.77_{1}
8	Swiss-Chocolate	68.80_{9}	65.56_{6}	68.74_{12}	48.22_{16}	73.95_{4}	$\mathbf{62.61}_{8}$	54.66_{20}
9	CLaC-SentiPipe	70.42_{7}	63.05_{18}	70.16_{10}	51.43_{10}	73.59_{6}	$\mathbf{62.00}_{9}$	58.55_{9}
10	TwitterHawk	68.44_{11}	62.12_{20}	70.64_{9}	56.02_{4}	70.17_{19}	$\mathbf{61.99}_{10}$	61.24_{6}
11	SWATCS65	68.21_{12}	65.49_{8}	67.23_{14}	37.23_{39}	73.37_{8}	$\mathbf{61.89}_{11}$	52.64_{24}
12	UNIBA	61.66_{29}	65.50_{7}	65.11_{25}	37.30_{38}	70.05_{20}	$\mathbf{61.55}_{12}$	48.16_{32}
13	KLUEless	70.64_{6}	67.66_{4}	70.89_{6}	45.36_{26}	73.50_{7}	$\mathbf{61.20}_{13}$	56.19_{17}
14	NLP	66.96_{14}	61.05_{25}	67.45_{13}	39.87_{34}	66.12_{31}	$\mathbf{60.93}_{14}$	63.62_{3}
15	ZWJYYC	69.56_{8}	64.72_{11}	70.77_{8}	46.34_{23}	71.60_{15}	$\mathbf{60.77}_{15}$	52.40_{25}
16	Gradiant-Analytics	65.29_{22}	61.97_{21}	66.87_{17}	59.11_{1}	72.63_{11}	$\mathbf{60.62}_{16}$	56.45_{16}
17	IIIT-H	65.68_{20}	62.25_{19}	67.04_{16}	57.50_{2}	69.91_{21}	$\mathbf{59.83}_{17}$	62.75_{5}
18	ECNU	65.25_{23}	68.49_{1}	66.37_{20}	45.87_{25}	74.40_{3}	$\mathbf{59.72}_{18}$	52.67_{23}
19	CIS-positiv	64.82_{24}	65.14_{10}	66.05_{21}	49.23_{14}	71.47_{16}	$\mathbf{59.57}_{19}$	57.74_{11}
20	SWASH	63.07_{27}	56.49_{34}	62.93_{31}	48.42_{15}	69.43_{24}	$\mathbf{59.26}_{20}$	54.30_{21}
21	GTI	64.03_{25}	63.50_{16}	65.65_{22}	55.38_{6}	70.50_{17}	$\mathbf{58.95}_{21}$	57.02_{13}
22	iitpsemeval	60.78_{31}	60.56_{26}	65.09_{26}	47.32_{19}	73.70_{5}	$\mathbf{58.80}_{22}$	58.18_{10}
23	elirf	57.05_{32}	60.20_{28}	61.17_{35}	45.98_{24}	68.33_{28}	$\mathbf{58.58}_{23}$	43.91_{34}
24	SWATAC	65.86_{19}	61.30_{24}	66.64_{19}	39.45_{35}	68.67_{27}	$\mathbf{58.43}_{24}$	50.66_{27}
25	UIR-PKU*	67.41_{13}	64.67_{12}	67.18_{15}	52.58_{8}	70.44_{18}	$\mathbf{57.65}_{25}$	59.43_{8}
26	SWATCMW	65.67_{21}	65.43_{9}	65.62_{23}	37.48_{36}	69.52_{23}	$\mathbf{57.60}_{26}$	56.69_{14}
27	WarwickDCS	66.57_{15}	61.92_{22}	65.47_{24}	45.03_{28}	68.98_{25}	$\mathbf{57.32}_{27}$	56.58_{15}
28	SeNTU	63.50_{26}	60.53_{27}	66.85_{18}	45.18_{27}	68.70_{26}	$\mathbf{57.06}_{28}$	49.53_{29}
29	DIEGOLab	62.49_{28}	58.60_{30}	63.99_{28}	47.62_{18}	63.74_{34}	$\mathbf{56.72}_{29}$	55.56_{18}
30	Sentibase	61.56_{30}	59.26_{29}	63.29_{30}	47.07_{20}	67.55_{29}	$\mathbf{56.67}_{30}$	62.96_{4}
31	Whu-Nlp	65.97_{18}	61.31_{23}	63.93_{29}	46.93_{21}	71.83_{13}	$\mathbf{56.39}_{31}$	22.25_{40}
32	UPF-taln	66.15_{17}	57.84_{31}	65.05_{27}	50.93_{11}	64.50_{32}	$\mathbf{55.59}_{32}$	41.63_{35}
33	RGUSentimentMiners123	56.41_{34}	57.14_{32}	59.44_{36}	44.72_{29}	64.39_{33}	$\mathbf{53.73}_{33}$	48.21_{31}
34	IHS-RD*	55.06_{35}	57.08_{33}	61.39_{32}	37.32_{37}	66.99_{30}	$\mathbf{52.65}_{34}$	36.02_{37}
35	RoseMerry	52.33_{37}	53.00_{36}	61.27_{34}	49.25_{13}	62.54_{35}	$\mathbf{51.18}_{35}$	49.62_{28}
36	Frisbee	49.37_{38}	46.59_{38}	53.92_{38}	42.07_{32}	57.94_{38}	$\mathbf{49.19}_{36}$	48.26_{30}
37	UMDuluth-CS8761	54.17_{36}	50.64_{37}	55.82_{37}	43.74_{30}	60.23_{37}	$\mathbf{47.77}_{37}$	34.40_{38}
38	UDLAP2014	41.93_{39}	39.35_{39}	45.93_{39}	41.04_{33}	50.11_{39}	$\mathbf{42.10}_{38}$	40.59_{36}
39	SHELLFBK	32.14_{40}	26.14_{40}	32.20_{40}	35.58_{40}	34.06_{40}	$\mathbf{32.45}_{39}$	25.73_{39}
40	whu-iss	56.51_{33}	54.28_{35}	61.31_{33}	47.78_{17}	61.98_{36}	$\mathbf{24.80}_{40}$	57.73_{12}
	baseline	29.2	19.0	34.6	27.7	27.2	**30.3**	30.2

Table 11: **Results for subtask B: Message-Level Polarity.** The systems are ordered by their score on the Twitter2015 test dataset; the rankings on the individual datasets are indicated with a subscript. Systems with late submissions for the *progress* test datasets (but with timely submissions for the official 2015 test dataset) are marked with a *.

Compared to 2014, there is improvement on Twitter2013-test from 72.12 (TeamX) to 72.80 (Splusplus), on Twitter2014-test from 70.96 (TeamX) to 74.42 (Spluplus), on Twitter2014-sarcasm from 58.16 (NRC-Canada) to 59.11 (Gradiant-Analytics), and on LiveJournal from 74.84 (NRC-Canada) to 75.34 (Splusplus), but not on SMS: 70.28 (NRC-Canada) vs. 68.49 (ECNU).

#	System	Tweet	Tweet sarcasm
1	TwitterHawk	**50.51**$_1$	31.30$_2$
2	KLUEless	**45.48**$_2$	39.26$_1$
3	Whu-Nlp	**40.70**$_3$	23.37$_5$
4	whu-iss	**25.62**$_4$	28.90$_4$
5	ECNU	**25.38**$_5$	16.20$_6$
6	WarwickDCS	**22.79**$_6$	13.57$_7$
7	UMDuluth-CS8761	**18.99**$_7$	29.91$_3$
	baseline	**26.7**	26.4

Table 12: **Results for Subtask C: Topic-Level Polarity.** The systems are ordered by the official 2015 score.

#	Team	avgDiff	avgLevelDiff
1	KLUEless	**0.202**	0.810
2	Whu-Nlp	**0.210**	0.869
3	TwitterHawk	**0.214**	0.978
4	whu-iss	**0.256**	1.007
5	ECNU	**0.300**	1.190
6	UMDuluth-CS8761	**0.309**	1.314
	baseline	**0.277**	0.985

Table 13: **Results for Subtask D: Trend Towards a Topic.** The systems are sorted by the official 2015 score.

5.3 Subtask C: Topic-Level Polarity

The results for subtask C are shown in Table 12. This proved to be a hard subtask, and only three of the seven teams that participated in it managed to improve over a majority vote baseline. These three teams, TwitterHawk (using subtask B data to help with subtask C) with F_1=50.51, KLUEless (which ignored the topics as if it was subtask B) with F_1=45.48, and Whu-Nlp with F_1=40.70, achieved scores that outperform the rest by a sizable margin: 15-25 points absolute more than the fourth team.

Note that, despite the apparent similarity, subtask C is much harder than subtask B: the top-3 teams achieved an F_1 of 64-65 for subtask B vs. an F_1 of 41-51 for subtask C. This cannot be blamed on the class distribution, as the difference in performance of the majority class baseline is much smaller: 30.3 for B vs. 26.7 for C.

Finally, the last column in the table reports the results for the 75 sarcastic 2015 tweets. The winner here is KLUEless with an F_1 of 39.26, followed by TwitterHawk with F_1=31.30, and then by UMDuluth-CS8761 with F_1=29.91.

5.4 Subtask D: Trend Towards a Topic

The results for subtask D are shown in Table 13. This subtask is closely related to subtask C (in fact, one obvious way to solve D is to solve C and then to calculate the proportion), and thus it has attracted the same teams, except for one. Again, only three of the participating teams managed to improve over the baseline; not suprisingly, those were the same three teams that were in top-3 for subtask C. However, the ranking is different from that in subtask C, e.g., TwitterHawk has dropped to third position, while KLUEless and Why-Nlp have each climbed one position up to positions 1 and 2, respectively.

Finally, note that avgDiff and avgLevelDiff yielded the same rankings.

5.5 Subtask E: Degree of Prior Polarity

Ten teams participated in subtask E. Many chose an unsupervised approach and leveraged newly-created and pre-existing sentiment lexicons such as the Hashtag Sentiment Lexicon, the Sentiment140 Lexicon (Kiritchenko et al., 2014), the MPQA Subjectivity Lexicon (Wilson et al., 2005), and SentiWordNet (Baccianella et al., 2010), among others. Several participants further automatically created their own sentiment lexicons from large collections of tweets. Three teams, including the winner INESC-ID, adopted a supervised approach and used word embeddings (supplemented with lexicon features) to train a regression model.

The results are presented in Table 14. The last row shows the performance of a lexicon-based baseline. For this baseline, we chose the two most frequently used existing, publicly available, and automatically generated sentiment lexicons: Hashtag Sentiment Lexicon and Sentiment140 Lexicon (Kiritchenko et al., 2014).[5] These lexicons have real-valued sentiment scores for most of the terms in the test set. For negated phrases, we use the scores of the corresponding negated entries in the lexicons. For each term, we take its score from the Sentiment140 Lexicon if present; otherwise, we take the term's score from the Hashtag Sentiment Lexicon. For terms not found in any lexicon, we use the score of 0, which indicates a neutral term in these lexicons. The top three teams were able to improve over the baseline.

[5] http://www.purl.com/net/lexicons

Team	Kendall's τ coefficient	Spearman's ρ coefficient
INESC-ID	**0.6251**	0.8172
lsislif	**0.6211**	0.8202
ECNU	**0.5907**	0.7861
CLaC-SentiPipe	**0.5836**	0.7771
KLUEless	**0.5771**	0.7662
UMDuluth-CS8761-10	**0.5733**	0.7618
IHS-RD-Belarus	**0.5143**	0.7121
sigma2320	**0.5132**	0.7086
iitpsemeval	**0.4131**	0.5859
RGUSentminers123	**0.2537**	0.3728
Baseline	**0.5842**	0.7843

Table 14: **Results for Subtask E: Degree of Prior Polarity.** The systems are ordered by their Kendall's τ score, which was the official score.

6 Discussion

As in the previous two years, almost all systems used supervised learning. Popular machine learning approaches included SVM, maximum entropy, CRFs, and linear regression. In several of the subtasks, the top system used deep neural networks and word embeddings, and some systems benefited from special weighting of the positive and negative examples.

Once again, the most important features were those derived from sentiment lexicons. Other important features included bag-of-words features, hashtags, handling of negation, word shape and punctuation features, elongated words, etc. Moreover, tweet pre-processing and normalization were an important part of the processing pipeline.

Note that this year we did not make a distinction between constrained and unconstrained systems, and participants were free to use any additional data, resources and tools they wished to.

Overall, the task has attracted a total of 41 teams, which is comparable to previous editions: there were 46 teams in 2014, and 44 in 2013. As in previous years, subtask B was most popular, attracting almost all teams (40 out of 41). However, subtask A attracted just a quarter of the participants (11 out of 41), compared to about half in previous years, most likely due to the introduction of two new, very related subtasks C and D (with 6 and 7 participants, respectively). There was also a fifth subtask (E, with 10 participants), which further contributed to the participant split.

We should further note that our task was part of a larger Sentiment Track, together with three other closely-related tasks, which were also interested in sentiment analysis: Task 9 on CLIPEval Implicit Polarity of Events, Task 11 on Sentiment Analysis of Figurative Language in Twitter, and Task 12 on Aspect Based Sentiment Analysis. Another related task was Task 1 on Paraphrase and Semantic Similarity in Twitter, from the Text Similarity and Question Answering track, which also focused on tweets.

7 Conclusion

We have described the five subtasks organized as part of SemEval-2015 Task 10 on Sentiment Analysis in Twitter: detecting sentiment of terms in context (subtask A), classifiying the sentiment of an entire tweet, SMS message or blog post (subtask B), predicting polarity towards a topic (subtask C), quantifying polarity towards a topic (subtask D), and proposing real-valued prior sentiment scores for Twitter terms (subtask E). Over 40 teams participated in these subtasks, using various techniques.

We plan a new edition of the task as part of SemEval-2016, where we will focus on sentiment with respect to a topic, but this time 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*, and is on par with what applications need. They are not interested in the sentiment of a particular tweet but rather in 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*.

Acknowledgements

The authors would like to thank SIGLEX for supporting subtasks A–D, and the National Research Council Canada for funding subtask E.

References

Eric Almquist and Jason Lee. 2009. What do customers really want? *Harvard Business Review*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC '10, pages 2200–2204, Valletta, Malta.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Beijing, China.

Albert Bifet, Geoffrey Holmes, Bernhard Pfahringer, and Ricard Gavaldà. 2011. Detecting sentiment change in Twitter streaming data. *Journal of Machine Learning Research, Proceedings Track*, 17:5–11.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Uppsala, Sweden.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Short Papers*, ACL-HLT '11, pages 581–586, Portland, Oregon, USA.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Volume 2*, NAACL '03, pages 34–36, Edmonton, Canada.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA.

Bernard Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 356–364, Montréal, Canada.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, pages 81–93.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research (JAIR)*, 50:723–762.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM '11, pages 538–541, Barcelona, Catalonia, Spain.

Erich Leo Lehmann and Howard JM D'Abrera. 2006. *Nonparametrics: statistical methods based on ranks*. Springer New York.

Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, USA.

Jordan J. Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, University of Alberta.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 321–327, Atlanta, Georgia, USA.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 312–320, Atlanta, Georgia, USA.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan Routledge, and Noah Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10, pages 122–129, Washington, DC, USA.

Alexander Pak and Patrick Paroubek. 2010. Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 436–439, Uppsala, Sweden.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Philadephia, Pennsylvania, USA.

Alan Ritter, Sam Clark, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1524–1534, Edinburgh, Scotland, UK.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, Beijing, China.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 73–80, Dublin, Ireland.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. 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, pages 347–354, Vancouver, British Columbia, Canada.

Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '14, pages 304–313, Baltimore, Maryland, USA.