**Task 4: Sentiment Analysis in Twitter**

**SUMMARY**

**This is a rerun of**[**SemEval-2015 Task 10**](http://alt.qcri.org/semeval2015/task10/)**with two important changes:**

**- focus on new machine learning problems: quantification and ordinal classification  
- 2-point and 5-point scale (vs. 3-point we used in the past)**

NOTE: Due to its popularity, we also keep the general 3-way sentiment polarity classification subtask (subtask A), but we retire the expression-level subtask (what was SemEval-2015 task 10, subtask A).

**I. Introduction**

In the past decade, new forms of communication, such as microblogging and text messaging, have emerged and become ubiquitous.  While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them.

Working with these informal text genres presents challenges for natural language processing beyond those typically encountered when working with more traditional text genres, such as newswire data.  Tweets and texts are short: a sentence or a headline rather than a document.  The language used is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for “re-tweet” and #hashtags, which are a type of tagging for Twitter messages.  How to handle such challenges so as to automatically mine and understand the opinions and sentiments that people are communicating has only very recently been the subject of research (Jansen et al., 2009; Barbosa and Feng, 2010; Bifet and Frank, 2010; Davidov et al., 2010; O’Connor et al., 2010; Pak and Paroubek, 2010; Tumasjen et al., 2010; Kouloumpis et al., 2011).

We believe that a freely available, annotated corpus that can be used as a common testbed is needed in order to promote research that will lead to a better understanding of how sentiment is conveyed in tweets and texts.  Our primary goal in this task is to create such a resource: a corpus of tweets marked with their message-level polarity, in general and towards a specific topic.  The few corpora with detailed opinion and sentiment annotation that have been made freely available, e.g., the MPQA corpus (Wiebe et al., 2005) of newswire data, have proved to be valuable resources for learning about the language of sentiment.  While a few twitter sentiment datasets have been created, they are either small and proprietary, such as the i-sieve corpus (Kouloumpis et al., 2011), or they rely on noisy labels obtained from emoticons or hashtags.

**II. What is New This Year**

We propose three new subtasks, each of them a variant of the basic binary sentiment classification task for Twitter, i.e., classify a tweet known to be about a certain topic as expressing a Positive or a Negative view about that topic. By topic here we mean anything people on Twitter usually express opinions about; for example, a product (e.g., iPhone6), a political candidate (e.g., Hillary Clinton), a policy (e.g., Obamacare), an event (e.g., the Pope's visit to Palestine), etc. We consider the topic as given, i.e., it is NOT the task of the participants to find out whether the tweet is about the topic or not, or whether the sentiment expressed is about the topic.

The new subtasks stem from two new directions, taken individually and in combination:

**(1) Replacing classification with quantification.** Essentially, this starts from the consideration that, when it comes to Twitter, from an application point of view nobody is interested in whether A SPECIFIC PERSON has a positive or a negative view of the topic; rather, people are interested in HOW MANY people have a positive or a negative view of the topic (or, more precisely, in estimating the PERCENTAGE of tweets that are positive and the PERCENTAGE of tweets that are negative in a given set of tweets about the topic). This is always true in political science, computational social studies, market research, online reputation management, etc.; in short, it is true in each of the major fields which are interested in sentiment classification of tweets.

Estimating these percentages (more generally, estimating the distribution of the classes in a set of unlabelled items) by leveraging training data is called, in data mining and several other fields, QUANTIFICATION. In the literature, it has been argued that classification is not the same as quantification, since (a) a good classifier is not necessarily a good quantifier, and vice versa; see, e.g., (Forman, 2008); (b) quantification requires evaluation measures different from classification, since it needs to evaluate results at the aggregate rather than at the individual level. Quantification-specific learning approaches have been proposed over the years; Sections 2 and 5 of (Esuli and Sebastiani, 2015) contain several pointers to such literature.

**(2) Replacing the standard two-point scale (Positive / Negative) or three-point scale (Positive + Neutral + Negative) with a five-point scale (VeryPositive + Positive + OK + Negative + VeryNegative).**This stems from the consideration that a five-point scale is now ubiquitous in the corporate world where human ratings are involved; e.g., Amazon, TripAdvisor, Yelp, and many others, all use a five-point scale for their reviews.

Moving from a two-point scale to an ordered five-point scale means, in scientific terms, moving from binary classification to ordinal classification (a.k.a. ordinal regression). Again, learning approaches that take into consideration the fact that a total order is defined on the set of classes have been studied for many years.

**III. Subtasks**

* **(rerun) Subtask A: Message Polarity Classification:**Given a tweet, predict whether the tweet is of positive, negative, or neutral sentiment. *(This is SemEval-2015 task 10, subtask B, which we want to keep due to its popularity -- it has attracted 40 teams; however, we are retiring what was SemEval-2015 task 10, subtask A)*
* **(partially new) Subtask B: Tweet classification according to a two-point scale:**Given a tweet known to be about a given topic, classify whether the tweet conveys a positive or a negative sentiment towards the topic. *(This is a simplification of Subtask C as from SemEval-2015 task 10, which also required to filter out tweets that were not about the topic, and which (like Subtask A does now) also involved the Neutral class.)*
* **(new) Subtask C: Tweet classification according to a five-point scale:**Given a tweet known to be about a given topic, estimate the sentiment conveyed by the tweet towards the topic on a five-point scale.
* **(new) Subtask D: Tweet quantification according to a two-point scale:**Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the Positive and Negative classes.
* **(new) Subtask E: Tweet quantification according to a five-point scale:**Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the five classes of a five-point scale.

**IV. Evaluation**

  There will be different evaluation measures for the different subtasks. A detailed description can be found [here](http://alt.qcri.org/semeval2016/task4/data/uploads/eval.pdf).

**V. Organizers**

* Preslav Nakov, Qatar Computing Research Institute, HBKU
* Alan Ritter, Ohio State University
* Sara Rosenthal, Columbia University
* Fabrizio Sebastiani, Qatar Computing Research Institute, HBKU
* Veselin Stoyanov, Facebook