SemEval-2017 Task 4

Sentiment Analysis in Twitter

Sentiment Analysis in Twitter

Summary

This will be a rerun of SemEval-2016 Task 4 with several changes:

- new subtasks: another language and user information
- new evaluation measures
- new training datasets
- new test datasets

I. Introduction

The recent rise of social media has greatly democratized content creation. Facebook, Twitter, Skype, Whatsapp and LiveJournal are now commonly used to share thoughts and opinions about anything in the surrounding world. This proliferation of social media content has created new opportunities to study public opinion, with Twitter being especially popular for research due to its scale, representativeness, variety of topics discussed, as well as ease of public access to content.

Unfortunately, research in that direction was hindered by the unavailability of suitable datasets and lexicons for system training, development and testing. While some Twitter-specific resources were developed, initially they were either small and proprietary, such as the i-sieve corpus (Kouloumpis et al., 2011), were created only for specific languages (e.g., Villena-Roman et al., 2013), or relied on noisy labels obtained automatically (Mohammad, 2012; Pang et al., 2002).

This situation changed with the shared task on Sentiment Analysis on Twitter, part of the International Workshop on Semantic Evaluation (SemEval), a semantic evaluation forum previously known as SensEval. The task ran in 2013, 2014, 2015 and 2016, attracting over 40+ participating teams in all four editions. While the focus was on general tweets, the task also featured out-of-domain testing on SMS messages, LiveJournal messages, as well as on sarcastic tweets. SemEval-2013 Task 2 (Nakov et al., 2013) and SemEval-2014 Task 9 (Rosenthal et al., 2014) had an expression-level and a message-level polarity subtasks. SemEval-2015 Task 10 (Rosenthal et al., 2015; Nakov et al., 2016b) further added subtasks on topic-based message polarity classification, on detecting trends towards a topic, and on determining the out-of-context (a priori) strength of association of terms with positive sentiment. SemEval-2016 Task 4 (Nakov et al., 2016a) dropped the phrase-level subtask and the strength of association subtask, and focused on sentiment with respect to a topic. It further introduced a 5-point scale, which is used for human review ratings on popular websites such as Amazon, TripAdvisor, Yelp, etc.; from a research perspective, this meant moving from classification to ordinal regression. Moreover, it also focused on quantification, i.e., determining what proportion of a set of tweets on a given topic are positive/negative about it. It also featured a 5-point scale ordinal quantification subtask (Gao and Sebastiani, 2015).

Other related (mostly non-Twitter) SemEval tasks have explored aspect-based sentiment analysis (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016), sentiment analysis of figurative language on Twitter (Ghosh et al., 2015), implicit event polarity (Russo et al., 2015), stance in tweets (Mohammad et al., 2016), out-of-context sentiment intensity of phrases (Kiritchenko et al., 2016), and emotion detection (Strapparava and Mihalcea, 2007). Some of these tasks featured languages other than English, such as Arabic (Pontiki et al., 2016; Mohammad et al.,2016); they did not involve sentiment analysis of tweets or topics in tweets.

There has been emerging work in sentiment analysis for Arabic, especially as more resources and tools for Arabic NLP become readily available. Earlier studies focused on sentiment analysis in Arabic newswire (AbdulMageed & Diab., 2011, Elarnaoty et al., 2012), but recently there has been work on social media, which has largely included Twitter (Mourad & Darwish. 2013; AbdulMageed & Diab., 2014; Refaee & Reiser., 2014; Mohammad et al., 2015). There have been increasing efforts towards building Arabic sentiment lexicons (AbdulMageed & Diab., 2011; Badaro & al., 2015; Eskander & Rambow., 2015), and Arabic dialectal lexicons aimed towards sentiment analysis in Twitter (Refaee & Reiser., 2014; Mohammad et al., 2015). Some work studied the effect of cross-lingual MT-based methods for Arabic sentiment analysis (Mohammad et al., 2015; Salameh et al., 2015; Refaee & Reiser., 2015), identification of sentiment holders (Elarnaoty et al., 2012) and sentiment targets or topics (Al Smadi et al., 2015; Farra et al,. 2015). We believe the development of a standard Arabic Twitter dataset for sentiment, and particularly with respect to topics, will be helpful for encouraging further

Contact Info

- Sara Rosenthal, IBM Research
- Noura Farra, Columbia University
- Preslav Nakov, Qatar Computing Research Institute, HBKU

email: semevaltweet@googlegroups.com

Other Info

Announcements

- **Results, and gold labels are** released
- Arabic and English TEST INPUT v1.0 for phase 2 (subtasks B, D) released
- Arabic and English TEST INPUT v3.0 for phase 1 (subtasks A, C, E) released
- Arabic and English training data released
- CodaLab development sets on Data and Tools page

research in this regard.
We expect the quest for more interesting formulations of the general sentiment analysis task to continue. We see SemEval as the engine of this innovation, as it not only does head-to-head comparisons, but also creates databases and tools that enable follow-up research for many years afterwards.
The proposed new edition of the task reiterates the new tasks from 2016, and offers two new directions: (i) adding a new language, Arabic, and (ii) adding information about the Twitter users. Moreover, we will add new evaluation measures, new training datasets and new test datasets.

II. Subtasks

Subtask A. (rerun): Message Polarity Classification: Given a message, classify whether the message is of positive, negative, or neutral sentiment.
Subtasks B-C. (rerun): Topic-Based Message Polarity Classification:
Given a message and a topic, classify the message on
B) two-point scale: positive or negative sentiment towards that topic
C) five-point scale: sentiment conveyed by that tweet towards the topic on a five-point scale.
Subtasks D-E. (rerun): Tweet quantification:
Given a set of tweets about a given topic, estimate the distribution of the tweets across
D) two-point scale: the "Positive" and "Negative" classes
E) five-point scale: the five classes of a five-point scale.
I. (new) Polarity classification in another language: In addition to performing subtasks A-E in English, we will also perform multilingual experiments by providing a run of these tasks using a test set containing tweets in Arabic.
II. (new) User Information: We aim to harness the user profile information provided in Twitter such as demographics (e.g., age, location) as well as the social network. We would like to analyze its impact on improving sentiment analysis.
III. Data

From SemEval-2016 Task 4, we already have datasets with Twitter messages on a range of topics, including a mixture of entities (e.g., Gadafi, Steve Jobs), products (e.g., kindle, android phone), and events (e.g., Japan earthquake, NHL playoffs).

These preexisting datasets will be made available for training, tuning and dev-testing the systems for the task reruns; they would be also allowed to be used for the new subtasks, if participants think they can make use of them in addition to new datasets that we are going to prepare.

We will do annotations using CrowdFlower or Mechanical Turk (most likely the former). We already have experience with both in previous years. We have secured funding to pay for the annotations: about 2000 new test tweets per English subtask + 8000-10000 new tweets for the Arabic data.

In addition to providing data, this year we will provide scripts that can be used to download user profile information, such as age, and location, as well as friend lists. This information can be used for the new addition to all tasks that is described in the summary as "II. (new) User Information".

IV. Evaluation

The metric for evaluating the participating systems will be as follows:
For Subtask A we will replace the measure used in 2016 (F1 averaged across the positives and the negatives) with macroaveraged recall (recall averaged across the three classes), since the latter has better theoretical properties than the former (Sebastiani, 2015), and since this provides better consistency with Subtask B
For Subtask B we will maintain the same measure used in 2016, i.e., macroaveraged recall;
For Subtask C, we will use macroaveraged mean absolute error as the main measure (consistently with what we did in 2016), but we will also add a secondary measure, namely, an extension of macroaveraged recall for ordinal regression (Sebastiani, 2016); again, this will provide better consistency with Subtasks A and B
For Subtask D, we will keep Kullback-Leibler Divergence (the measure used in 2016) as the main measure, but we will also add a secondary measure, namely, Pearson Divergence.
For Subtask E, the only measure we will consider, consistently with 2016, will be the Earth Mover's Distance, since it is the only known measure for ordinal quantification.

Each participating team will initially have access to the training data only. Later, the unlabelled test data will be released. After SemEval-2017, the labels for the test data will be released as well. We will ask the participants to submit their predictions, and the organizers will calculate the results for each participant. We will make no distinction between constrained and unconstrained systems, but the participants will be asked to report what additional resources they have used for each submitted run.

V. Baselines, scorers, format checkers

As in previous years, we will provide scorers, format checkers and simple baselines. Some of these are readily available, e.g., for the task reruns.

VII. Organizers

Noura Farra (noura@cs.columbia.edu) Columbia University
Preslav Nakov (pnakov@qf.org.qa) Qatar Computing Research Institute
Sara Rosenthal (srosenthal@us.ibm.com) IBM Watson Health Research

VIII. References

Muhammad Abdul-Mageed, Mona T. Diab, and Mohammed Korayem. 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., Association for Computational Linguistics, 2011.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. SAMAR: Subjectivity and sentiment analysis for Arabic social media. In Computer Speech & Language 28.1, pp. 20-37, 2014.

Mohammad Al-Smadi, Omar Qawasmeh, Bashar Talafha, and Muhannad Quwaide. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on. IEEE, 2015.

Gilbert Badaro, Badaro, Gilbert, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. A large scale Arabic sentiment lexicon for Arabic opinion mining. In the 1st Workshop on Arabic Natural Language Processing, 2014.

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. Evaluation Measures for Ordinal Regression. In Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA 2009), Pisa, IT, 2009, pp. 283-287.

Barbosa, L. and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of Coling.

Bifet, A. and Frank, E. 2010. Sentiment knowledge discovery in twitter streaming data. Proceedings of 14th International Conference on Discovery Science.

Davidov, D., Tsur, O., and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. Proceedings of Coling.

Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. A machine learning approach for opinion holder extraction in Arabic language. International Journal of Artificial Intelligence & Applications, 2012.

Ramy Eskander, and Owen Rambow. SLSA: A Sentiment Lexicon for Standard Arabic. In EMNLP 2015.

Andrea Esuli and Fabrizio Sebastiani. Sentiment Quantification. IEEE Intelligent Systems, 25(4): 72-75, 2010.

Andrea Esuli and Fabrizio Sebastiani. Optimizing Text Quantifiers for Multivariate Loss Functions. ACM Transactions on Knowledge Discovery from Data. 2015, in press.

Noura Farra, Kathleen McKeown, and Nizar Habash. Annotating Targets of Opinions in Arabic using Crowdsourcing. In Proceedings of the 2nd workshop on Arabic Natural Language Processing, Association for Computational Linguistics, 2015.

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

Wei Gao and Fabrizio Sebastiani. 2015. Tweet sentiment: From classification to quantification. In Proceedings of the 7th International Conference on Advances in Social Network Analysis and Mining, ASONAM '15, pages 97–104, Paris, FR.

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

Jansen, B.J., Zhang, M., Sobel, K., and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11):2169-2188.

Svetlana Kiritchenko, Saif M Mohammad, and Mohammad Salameh. 2016. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California.

Kouloumpis, E., Wilson, T., and Moore, J. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of ICWSM.

Saif Mohammad. 2012. #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, pages 246–255, Montreal, Canada.

Mohammad, Saif M., Mohammad Salameh, and Svetlana Kiritchenko. How Translation Alters Sentiment. Journal of Artificial Intelligence Research 54 (2015): 1-20.

Mohammad Salameh, Saif M. Mohammad, and Svetlana Kiritchenko. Sentiment after translation: A case-study on arabic social media posts. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015.

Ahmed Mourad and Kareem Darwish. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media

analysis. 2013.
Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S. Stoyanov, V. and Wilson, T. Semeval-2013 Task 2: Sentiment Analysis in Twitter To appear in Proceedings of the 7th International Workshop on Semantic Evaluation. Association for Computational Linguistics. June 2013, Atlanta, Georgia
Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016a. SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California.
Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M. Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2016b. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. Language Resources and Evaluation, 50(1):35–65.
O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. Proceedings of ICWSM.
Pak, A. and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.
Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the Con- ference on Empirical Methods in Natural Language Processing, EMNLP '02, pages 79–86, Philadelphia, Pennsylvania.
Maria Pontiki, Harris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, pages 27–35, Dublin, Ireland.
Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pages 486–495, Denver, Colorado.
Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California.
Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California.
Eshrag Refaee and Verena Rieser. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme. 2014.
Eshrag Refaee and Verena Rieser. Benchmarking Machine Translated Sentiment Analysis for Arabic Tweets. NAACL-HLT 2015 Student Research Workshop (SRW). 2015.
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.
Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pages 450–462, Denver, Colorado.
Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. SemEval-2015 task 9: CLIPEval implicit polarity of events. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval '15, pages 442–449, Denver, Colorado.
Fabrizio Sebastiani. An Axiomatically Derived Measure for the Evaluation of Classification Algorithms. Proceedings of the 5th ACM International Conference on the Theory of Information Retrieval (ICTIR 2015), Northampton, US, 2015, pp. 11-20.
Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In Proceedings of the International Workshop on Semantic Evaluation, SemEval '07, pages 70–74, Prague, Czech Republic.
Tumasjan, A., Sprenger, T.O., Sandner, P., and Welpe, I. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. Proceedings of ICWSM.
Julio Villena-Roman, Sara Lana-Serrano, Eugenio Mart inez-Camara, and Jose Carlos Gonzalez Cristóbal. 2013. TASS-Workshop on Sentiment Analysis at SEPLN. Natural, 50:37–44.
Janyce Wiebe, Theresa Wilson and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.