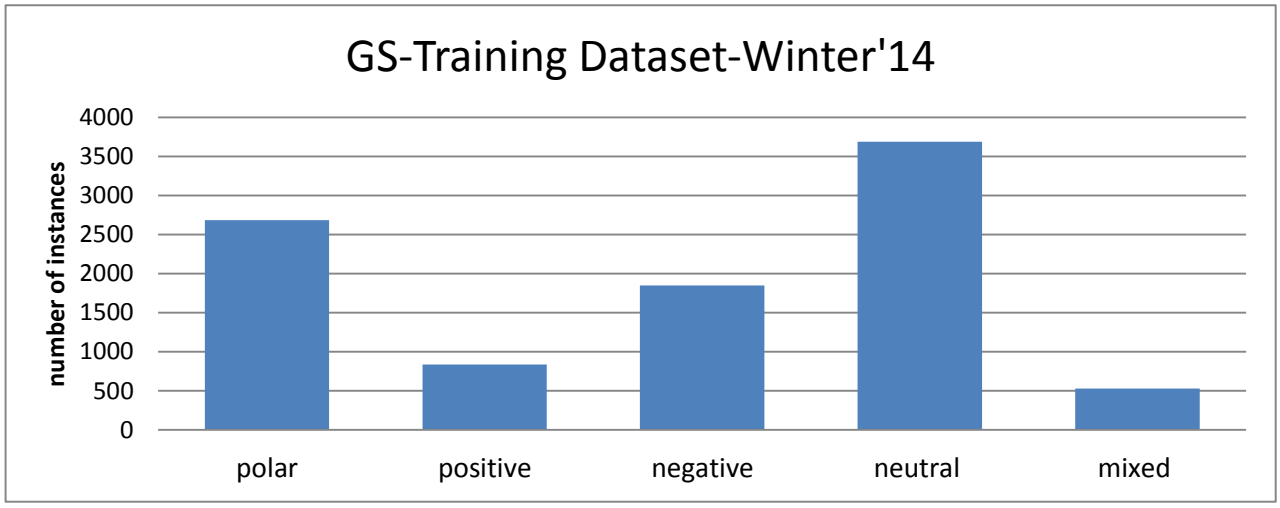
Refaee & Rieser (R&R): an Arabic Gold-Standard Twitter Sentiment Corpus 2014

- Data collection: 20/01/14 to 21/02/14
- The dataset was randomised and a subset of 10,000 instances was selected
- The following annotation scheme was given to two native speaker annotators to assign a single label denoting the emotional state of a given text

Label	Definition	Example
positive	▪ Clear positive indicator	كم انت عظيم يا بشار الاسد How great you are, Bashar Al-Asad
Negative	▪ Clear negative indicator	حنا للأسف نستخدم ايفون Unfortunately, we use the iPhone
Neutral	• Simple factual statement/ news ▪ Open questions with no emotions indicated	وفاة جديدة بإنش7إن9 بالصين A new reported death case with H7N9 in China بكم سعر الايفون 5 حاليا؟ How much is the iPhone these days?
Mixed	▪ Mixed positive and negative indicators	نحن نعشق الديمقراطية و نكره فوضى الاخوان المسلمين التي تريد تدمير حرياتنا we love democracy, but hate the mess that Muslim Brotherhood is making to destroy our freedom
Uncertain	▪ Undeterminable indicators/neither positive or negative/ lack subjective cues	المساواة في قمع الحريات الشخصية عدل Equality in suppressing personal freedoms is justice أحياناً فهمنا للأمور بطريقة خطا يكون هو الصح :) sometimes, the wrong understanding of the things is the right thing :)
Skipped	▪ Repeated/redundant/advertising tweet	-

- The distribution of sentiment labels is as follows:



label	number of instances
Polar (positive + negative)	2,681
positive	833*
negative	1,848*
neutral	3,685*
mixed	528*
Skipped	1,876
Uncertain	1,248
Total to be released (positive+negative+neutral+Mixed)	6,894*

- Weighted Kappa= 0.816

Word frequencies	96,493
Word tokens	26,724

- Average words/tokens per tweet= 20
- Chi-squared selection of the most informative attributes

Word	English	Chi2
عداء	Hostility	29.7918
باريس	Paris	24.9781
سعيد	Happy	18.4852
انستجرام	Instagram	18.0461
انبياء	Prophets	14.1358

زياده	Increase	14.1358
طرق	Ways	13.83
يوتيوب	YouTube	11.2036
سام	Poisonous	9.6753
حميد	benignant	9.4724

- Features sets:
  - morphological features (extracted using a state-of-art morphological analyser MADA version 3.2)
  - semantic features (binary features examine the presence of sentiment-bearing words)
  - stylistic features (binary features examine the presence of emoticons)
  - Affective cues (binary features examine the presence of a set of social signals: consent, dazzle, laughs, regret, sigh)
  - Tweet category (nominal feature identifies a theme/category of each given tweet)
  - Language style (MSA/DA, isSarcastic, length)
  - Twitter-specific features (has-Hashtag, has-URL, isFavorite, isRetweet)

Release

Version 1.0 March 2014: First public release via LREC repository.

Version 1.1 April 2014: 1) Revised and corrected sentiment labels 2) Additional feature-sets

Release format: ARFF and CSV

Due to the restrictions by Twitter's API terms of service (https://dev.twitter.com/terms/api-terms/diff updated on July 2013), we may not share datasets of Tweet text. However, we may share a dataset of Twitter object IDs, and user-IDs that can be used to poll Twitter content using the statuses/show API methods.

All data from Twitter is covered by Twitter's terms of service (https://dev.twitter.com/terms/api-terms/diff). The sentiment labels and the associated features are provided free of charge and may be used for research purposes. Citing the following paper is appreciated if you use this corpus.

Citation

Refaee, E. and Rieser, V. (2014, May). An Arabic twitter corpus for subjectivity and sentiment analysis. In proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14). Reykjavik-Iceland, 26-31 May 2014.

Contact

Eshrag Refaee
Eaar1@hw.ac.uk
eshragrefaee@gmail.com
Twitter: @eshragR
Interaction Lab
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh