Introduction

- First release of an Arabic Twitter corpus for subjectivity and sentiment analysis (SSA).
- SSA aims to determine the attitude of the tweet’s author with respect to a topic.
- Lots of interest in analysing Arabic tweets (Abdul-Maged et al., 2012, Al-Sabagh and Girju, 2012, & Murad and Dawish, 2013) datasets not freely available yet.

Challenges of Arabic SSA in Social Networks:

- Users on social networks tend to use Dialectal Arabic (DA) — more than other languages — instead of the standard MSA. (Zaidan and Callison-Burch, 2013).
- Tweets may convey sarcasm, mixed and/or unclear polarity content as in the following example:

Football match in English, spelled using the Arabic alphabet.

Great job Ali! (referring to a famous football club)

- Tweets can be composed of local dialects, slangs, abbreviations, non-standard spellings, and non-English specific terminology (i.e. re:polarity text, and show significant variation from MSA)
- DA poses additional challenges for NLP researchers i.e. they lack standardisation, are written in free text, and show significant variation from MSA (Zaidan and Callison-Burch, 2013).
- Multi/bi-lingual users tend to use a mixture of languages, for instance:

Corpuses Collection

- Twitter Search API for harvesting a stream of tweets.
- Dataset collected from 20 Jan to 21 Feb 2014.
- Search queries:

Corpus Release

Version 1.0 March 2014: First public release via LREC repository.
Version 1.1 April 2014: Revised and corrected sentiment labels 2) Additional feature-sets
Release format: 1) Attribute-Relation File Format ARFF 2) Comma-Separated File CSV
Release Policy:
- Release of a dataset of Twitter object IDs, and user-IDs
- The sentiment labels and the associated features are provided free of charge via LREC repository and may be used for research purposes.
- Corpus also available at: http://bit.ly/11gKxk

Corpus Citation:


Supervised machine learning approach


Distant supervision emotion-based learning


Current Work

References


Contact

Ehsrag Refae
Eah@lthw.ac.uk
ehsragrefae@gmail.com

Twitter: @ehsragrefae

Interaction Lab
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh
United Kingdom

Feature-sets

Type Feature-sets

- Diachronic, aspect, gender, mood, person, past tense, present tense, state, tense, tone, verb, word-morphology analysis

Sentiments

- Positive, negative, neutral

- 28 positive and neutral items

- 5 negative items

- 79 negative and neutral items

- Value for each sentiment

- Pre-processing

- Normalising query-terms, user-names, hashtags, URLs, digits and non-Arabic characters.

Corpus Annotation

- 10,000 instances manually annotated.
- Annotation schematic:

- Coreference
- Clear positive indicator
- Clear negative indicator
- Mixed positive and negative indicators
- Simple factoid statement / news
- Opinion with no emotions indicated
- Mixed positive and negative indicators
- Underminable sentiment indicators
- Repealed/redundant advertising tweet

- Distribution of the sentiment labels:
  - gold standard training dataset

- Size of gold-standard data-set

- Average word tokens per tweet = 20.
- Most informative tokens (according to Chi-squared):