Analyzing Twitter messages with Big data resources and lexical resources

Rosa Meo, Emilio Sulis, Department of Computer Science, University of Turin

In the following we describe the dataset and the lexica we will use in the laboratory.

1. DATA-SET

In the corpus developed by Suttles and Ide [Suttles and Ide 2013], emotional tweets (written in English) are collected by manually labelling with the eight Plutchik's emotions an initial set of 56 hashtags. Then they used these hashtags to collect and label tweets. Their approach applies distant supervision as in [Mintz et al. 2009]. The original hashtags were selected among the most frequent ones in a 38.9 million tweet dataset. According to these emotional tokens, a huge data-set of 5.9 millions of micro-blog messages had been extracted. Then, tweets containing one or more emotional tokens from both classes of an opposing binary pair were discarded.

Messages were tokenized and normalized: each mention was replaced with the keyword `USERNAME` and each web address with the keyword `URL`. The words with more then two consecutive letters (elongated words) were replaced with only two. Finally, messages with quotes were discarded, as they may contain someone's else opinion or they are forwarding someone's else content (retweet).

Exploiting this large data-set, we extracted a sample of messages containing more than ten elements, whether words, emoticons, emoji and so on. In a pre-processing phase, we excluded very short messages with less than 10 tokens as they have poor textual information. Moreover, we manually checked the corpus to remove some spam messages [1].

To sum up, our corpus includes 48,000 messages, and more precisely 6k for each emotion.

2. DESCRIPTION OF THE LEXICA

We consider the occurrences of terms and concepts in several lexica, defining two categories of features related to polarity and affective resources. In this section we introduce our selection of ten dictionaries among the ones commonly used in this kind of studies. For instance, we opted for AFINN as it is specifically created for Twitter and SentiSense for its many emotional categories. In addition, we included several emotional resources.

The *polarity features* are related to lexica which assign a positive or negative polarity to each term. We consider here five lexical resources: `AFINN`, `Hu-Liu`, `General Inquirer`, `LIWC`, and `EffectWordNet`. The last four include two lists of positive and negative terms, while `AFINN` associates a single score, as we briefly describe here.

(i) AFINN: The dictionary includes 2,477 English manually labelled words with a sentiment score in a range from -5 up to $+5$. The list was collected by Finn Årup Nielsen [Nielsen 2011], including slang acronyms or obscene words used on the Internet[2]. A negative score represents a negative affect while a positive score a positive one. The words with a negative score are 1,598, while the positive ones are 878.

[1] For instance, we removed tweets created by meteo or traffic information services that do not contain explicit sentiment information by users

[2] https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt

(ii) HL: The Hu–Liu's lexicon has been largely used for opinion mining [Hu and Liu 2004]. The 6,789 terms[3] are both negative (4,783) and positive (2,006).

(iii) GI: The Harward General Inquirer includes 182 dictionary categories and subcategories[4]. We consider here one lists of 1,915 positive words and another one of 2,291 negative words.

(iv) LIWC: The Linguistic Inquiry and Word Counts [Pennebaker et al. 2001; Pennebaker et al. 2007] is a dictionary including $4,500$ words distributed in 80 linguistic and psychological categories[5]. In particular, two lists of words contain 405 positive and 500 negative emotion terms.

(v) EWN: The Effect WordNet lexicon has been recently developed by Choi [Choi and Wiebe 2014] exploiting the corresponding synsets in WordNet. It includes two lists of 3,298 positive and 2,427 negative terms[6].

The *affective resources* are mainly lists of terms labelled with a single emotion, as `EmoLex`, *EmoSN* and *SS*. In addition, we explored two dictionaries where terms are annotated in several psychological dimensions from the resources `ANEW` and `DAL`. In the following, we describe the five resources concerning the categories of emotions and the dimensional representation.

(i) EmoLex:[7] it was developed by Saif Mohammad [Mohammad and Turney 2013]. The dictionary contains 14,182 words labelled with the eight Plutchik's primary emotions: Sadness, Joy, Disgust, Anger, Fear, Surprise, Trust, and Anticipation.

(ii) EmoSN: EmoSenticNet includes 13,189 entries for the six Ekman's emotions of Joy, Sadness, Anger, Fear, Surprise and Disgust. The resource was developed by assigning WordNet Affect emotion labels to SenticNet concepts [Poria et al. 2013; Poria et al. 2014]. The last one is a list of common-sense knowledge concepts with a polarity score [Cambria et al. 2014] referring to the multidisciplinary approach of Sentic Computing [Cambria and Hussain 2015].

(iii) SS: SentiSense is a concept-based affective lexicon with a wide set of categories developed by Carrillo de Albornoz [Carrillo de Albornoz et al. 2012], including 5,496 words and 2,190 synsets from WordNet, labeled with an emotion from a set of 14 categories[8].

(iv) ANEW: The dictionary Affective Norms for English Words includes terms rated from 1 to 9 for each of the three dimensions of Valence, Arousal and Dominance.

(v) DAL: The Dictionary of Affective Language developed by Whissell [Whissell 2009] contains words belonging to the dimensions of Pleasantness, Activation and Imagery. The 8,742 terms are rated in a three-point scale[9].

These lexica can be grouped on the basis of two dichotomies. The first one distinguishes between *Polarity-lexicon* dictionaries, composed by positive and negative words and *Emotion-lexicon* dictionaries, composed by terms with the same emotional content. The second dichotomy distinguishes between *Categorical* dictionaries, with entries grouped into a category and *Annotated values* dictionaries, with list of entries annotated with a single score.

For example, `EmoLex` includes a list of terms for each emotion, such as Joy, Sadness, Anger and so on. A resource such as `AFINN` includes lists of annotated terms with values

[3] http://www.cs.uic.edu/~liub/FBS/

[4] http://www.wjh.harvard.edu/~inquirer/homecat.htm

[5] http://www.liwc.net, http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_operatormanual.pdf

[6] http://mpqa.cs.pitt.edu/

[7] EmoLex is also called NRC word-emotion association lexicon, cf. http://www.saifmohammad.com/WebPages/lexicons.html

[8] nlp.uned.es/~jcalbornoz/SentiSense.html

[9] ftp://perceptmx.com/wdalman.pdf

Table I. Adopted lexica organized by subject (emotion-lexicon or sentiment polarity-lexicon) and typology (single value or category)

Description	Emotion-lexicon	Polarity-lexicon
Categorical	EmoLex, EmoSN, SS	EWN, GI, HuLiu, LIWC
Annotated values	ANEW, DAL	AFINN

which express the polarity of the terms as a whole. For instance, "funny": 0.4, "damn": -0.4 and so on. Instead, in DAL the term *butterfly* is associated to three values: $+2.6$, $+1.6364$ and $+3.0$ that represent respectively the value of *Pleasantness*, *Activation* and *Imagery*. Table I summarizes the different dictionaries used in this work.

REFERENCES

Erik Cambria and Amir Hussain. 2015. *Sentic computing: a common-sense-based framework for concept-level sentiment analysis*. Vol. 1. Springer.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Carla E. Brodley and Peter Stone (Eds.). AAAI Press, 1515–1521.

Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervas. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1181–1191.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*. 168–177.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2 (ACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1003–1011. http://dl.acm.org/citation.cfm?id=1690219.1690287

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

Finn Årup Nielsen. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, Vol. 718. CEUR-WS.org, 93–98.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. The Development and Psychometric Properties of LIWC2007. (2007).

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems* 69 (2014), 108–123.

Soujanya Poria, Alexander Gelbukh, Amir Hussain, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining. *IEEE Intelligent Systems* 28, 2 (March 2013), 31–38. DOI:http://dx.doi.org/10.1109/MIS.2013.4

Jared Suttles and Nancy Ide. 2013. Distant Supervision for Emotion Classification with Discrete Binary Values. In *Computational Linguistics and Intelligent Text Processing*, Alexander Gelbukh (Ed.). Lecture Notes in Computer Science, Vol. 7817. Springer Berlin Heidelberg, 121–136. DOI:http://dx.doi.org/10.1007/978-3-642-37256-8_11

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural languages. *Psychological Reports* 2, 105 (2009), 509–521.